Spatio-temporal analysis of extreme wind velocities for infrastructure design

Dissertation submitted in partial fulfillment of the requirements

for the Degree of Master of Science in Geospatial Technologies

Jan 2020

Alexys Herleym Rodríguez Avellaneda

alexyshr@gmail.com

https://github.com/alexyshr

Supervised by:

Prof. Dr. Edzer Pebesma

Institute for Geoinformatics

University of Münster - Germany

Co-supervised by:

Prof. Dr. Juan C. Reyes

Department of Civil and Environmental Engineering

Universidad de los Andes - Colombia

Co-supervised by:

Prof. Dr. Sara Ribero

Information Management School

Universidade Nova de Lisboa - Portugal

Declaration of Academic Integrity

I hereby confirm that this thesis on *Spatio-temporal analysis of extreme wind velocities for infrastructure design* is solely my own work and that I have used no sources or aids other than the ones stated.

All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

March 17, 2020

_____

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

March 17, 2020

_____

Acknowledgements

Special thanks to Prof. Dr. **Edzer Pebesma**, first, for all the contributions to the open source community, considering that main work in this thesis was done using his R packages, especially *sf*, *stars* and *gstat*, and second, for all high level knowledge transmitted through the subjects *Spatial Data Science with R* and *Analysis of Spatio-Temporal Data*, which were the motivation and basis to carry out the investigation.

Special thanks to Prof. Dr. **Juan C Reyes** for his contribution in selecting the research topic, and great contributions in information, methodology and support.

I would like to thank to:

Prof. Dr. **Edzer Pebesma**, Prof. Dr. **Juan C. Reyes**, and Prof. Dr. **Sara Ribero**, for supervising my work and spending their valuable time for discussions and feedback, it was really a huge advantage to have that support always available, and a pleasure to work beside you. Dr. **Adam Pintar**, for sharing its related POT-PP R Code, and for devoting much of his time to reviewing and commenting on my progress. Dr. **Joaquín Huerta Guijarro**, because he always was available to help and he was very friendly and receptive. Dr. **Christoph Brox**, because he was beside me in the difficult moments of the incident and the surgery. **European Union** -'Erasmus Mundus Grant', because their funding allow me to fulfill this dream to go further with my academic and professionals dreams. Engineer **Juan David Sandoval** for its valuable contributions. My mother **Ligia** made possible all my achievements, because she was always there with love, support, and valuable advice, I am grateful with all my heart. My daughter **Nicolle Chaely** for its love, support, and always pleasant company. Family members as **Elsa Manrique**, **Barbara Avellaneda**, and **Kevin Martinez**, because they were an important source of motivation and support. To all the beautiful people that shared with me different activities at **San Antonius Church of Münster**, with special mention of father **Alejandro Serrano Palacios** for support and friendship, and **choir friends**.

Preface

Models of extreme values are used for designing against the effects of *extreme events* like earthquakes, winds, rainfall, floods of different types of physical processes, see Beirlant, Goegebeur, Teugels, & Segers (2004), avoiding widespread destruction and loss of lives, see Haigh & Wahl (2019). This research presents a applied case of univariate extreme value analysis, explained in detail in Smith (2004), applied to wind velocities for infrastructure design, consequently, the main interest are probable future more extreme wind events, that structures need to be able to resist.

This work in its theoretical and methodological component was directed by ASCE7-16 Engineers (2017), considering that output products will be used to update the chapter B.6, wind forces, of the Colombian structure design norm, see Ministerio de Vivienda (2010), maintained by the Colombian Association of Seismic Engineering - AIS by its Spanish acronym. ASCE7-16, defines four risk categories, which implies the use of different wind loads (represented in wind extreme values for different mean recurrence intervals) for structures that belong to each category, 3000 years of MRI for risk IV, 1700 years for risk III, and 700 years for risk II and I.

This research has a particularly new situation regarding to the input data, and it is that not only time series of field measurements from meteorological stations are used (IDEAM data source), but also post-processed information coming from the Integrated Surface Database - ISD (USA database based on IDEAM data source), see Smith, Lott, & Vose (2011), and forecast reanalysis data from ERA5, see European Centre For Medium-Range Weather Forecasts (2017). This condition demanded a comparison of the different data sources, in order to verify the feasibility in the use of ERA5 and ISD, with a previous process of standardization of wind velocities (only for IDEAM and ISD), to reach the needed requirement of 3-s wind gust speed, 10 meters anemometer height, and terrain open space condition.

At each station the used method Peaks Over Threshold - Poisson Process, required to identify all the non-thunderstorm events in the non-hurricane dataset, through a process of declustering, choose a suitable threshold level to leave for the analysis only the most extreme values available, and then, fit to the data a intensity function, using maximum likelihood to find optimal parameters with the best goodness of fit. With the fitted model, it was possible to calculate return levels for required mean return intervals. Next, a process of spatial interpolation was done using Kriging, what allowed to have three continuous maps for the whole study area.

Table of Contents

List of Tables

List of Figures

Abstract

For the input non-hurricane, non tornadic data in each available station of the study area, this research calculate extreme winds or return levels for three different mean recurrence intervals - MRI, 700, 1700, and 3000 years, with a change of being equaled or exceeded only one time in the corresponding MRI period. Then, continuous maps of wind extreme velocities are interpolated to cover the study area, which are combined with existing wind extreme hurricane studies, to be used as input loads for infrastructure design.

The development of this research focused in non-hurricane data, covers three main areas, *downscaling support*, *temporal analysis*, and *spatial analysis*, and includes in the end an integration process with *existing results of hurricane studies*, which all together, allow to generate extreme winds maps with different mean recurrence intervals (MRIs), for the design of structures of different risk categories, namely, less risky/important structures for short MRIs (700 and 1700 years), and highly important structures for the longest MRI of 3000 years.

Due to the specific characteristics of the study area, where there is a lack of historical wind measurements (IDEAM data source), it became necessary to look for *alternative data sources*, ISD (model, based on IDEAM), and ERA5 (forecast data), which resulted in the downscaling issue, and that was confronted from a graphic comparison of all sources by matching stations, in the search of adequate support for the use of complementary data. The result of the comparison showed little similarity between the different sources. Prior to the comparison process, ISD and IDEAM data sources were standardized to represent 3-second wind gust, 10 meters anemometer height, and terrain open space roughness.

The method of temporal analysis used to calculate the return levels at each station, from the historical wind time series, is the Peaks Over Threshold - POT, using a non-homogeneous, bi-dimensional Poisson Process - PP, recommended by Engineers (2017), and developed and implemented in Pintar, Simiu, Lombardo, & Levitan (2015), considering from a maximum likelihood adjustment, the model with the best goodness of fit. Main components of this matters are de-clustering, thresholding, intensity function fitting, hazard curve, and return levels calculation.

Non-hurricane maps where created for data sources ISD and ERA5, using Kriging as spatial interpolation method, and after the integration with hurricane studies, the results for ERA5 showed the most reliable final maps, despite limitations in input data. Due to the limitation in the classification of storm and non-storm data, ISD final map showed very very high wind values, which are quite unlikely.

List of Acronyms

AIS	Seismic Engineering Association
ASCE	American Society of Civil Engineers
ASCE7-16	ASCE/SEI Design Loads Standard
cdf	Cumulative Distribution Function
EDA	Exploratory Data Analysis
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF climate reanalysis dataset
EVD	Extreme Value Distribution (GEVD, GEV)
GEVD	Generalized Extreme Value Distribution (EVD, GEV)
GEV	Generalized Extreme Value Distribution (GEVD, EVD)
GPD	Generalized Pareto Distribution
hf	Hazard Function
IDEAM	Institute of Hydrology, Meteorology and Environmental Studies
IDW	Inverse Distance Weighted
ISD	Integrated Surface Database
MRI	Mean Return Interval or Return Period
NSR	Seismic Resistant Norm
NOAA	National Oceanic and Atmospheric Administration
NetCDF	Network Common Data Form
NCEI	NOAA's National Centers for Environmental Information
P_e	Annual Exceedance Probability
pdf	Probability Distribution Function
P_n	Compound Exceedance Probability
POT	Peaks Over Threshold
ppf	Percent Point Function (Quantile)
PP	Poisson Process
POT-GPD	Peaks Over Threshold - Generalized Pareto Distribution
POT-PP	Peaks Over Threshold - Poisson Process
RL	Return Level
RMSE	Root Mean Squared Error
SEI	Structural Engineering Institute
SQL	Structured Query Language
WGS84	World Geodetic System 1984

Chapter 1

Introduction

This research aims to create non-hurricane non-tornadic maps of extreme wind speeds, for *three specific recurrence intervals* (700, 1700, and 3000 years) covering the study area (Colombian territory). These maps will be combined with existing hurricane wind speed studies, to be used as input loads due to wind, for infrastructure design.

For each station with wind speeds time series in the input data, following Pintar et al. (2015), extreme wind speeds corresponding to each recurrence interval are calculated using a *Peaks Over Threshold Poisson Process* extreme value model, onwards *POT-PP*, then wind velocities with the same recurrence interval are *spatially interpolated* to generate continuous maps for the whole study area.

A wind speed linked to a *mean recurrence interval - MRI* of *N-years* (N-years return value or return period) is interpreted as the highest probable wind speed along the period of N-years, see Engineers (2017). The annual probability of equal or exceed that wind speed is 1/N. The annual exceedance probability for all velocity values in 700-years output map will be 1/700, for the 1700-years map will be 1/1700, and 1/3000 for the 3000-years final map.

1.1 Context and Background

To design a specific structure, the horizontal forces, wind and earthquake, play an starring role. For the study area, Colombia, initially the wind force was considered as a fixed velocity $100\frac{Km}{h}$, later, a continuous map with a return period of 50 years was included in the official design standard, then, an additional map with return period of 700 year was included, see Ministerio de Vivienda (2010).

In the context of this study, extreme wind analysis is concerned with statistical methods applied to very high values of wind as random variable in a stochastic process, to allow statistical inference from historical data, namely, assess from the ordered sample of wind velocities, the probability of wind events that are more extreme than the ones previously observed and included in the mentioned input sample. Classical reference in this matter is Coles (2001), where a detailed study is done about classical extreme value theory and models

and threshold models.

There are four main approaches to deal with extreme value analysis, Smith (2004): a) sample maxima associated to a Generalized Extreme Value Distribution - GEV (traditional method), b) exceedances over threshold associated to a Generalized Pareto Distribution, onwards *POT-GPD*, c) the Poisson-GPD, an homogeneous Poisson process for the number of exceedances, and a GPD for the excess values, and d) the exceedances over threshold associated to a non-homogeneous, non-stationary, bi-dimensional Poisson process, a Point process approach also known as POT-PP. Main details will be discussed here for each method, but as the last one is recommended in Asce2017, a more indeed explanation will be provided in POT-Poisson Process. There is a whole section with the details about the background of this research, see Theoretical Framework

1.1.1 Sample Maxima

To work with random variables of sample maximum values, the used probability distribution function *pdf* is the GEV

$$H(y) = \exp\left\{-\left(1+\xi\frac{y-\mu}{\psi}\right)_+^{-\frac{1}{\xi}}\right\},$$

($y+ = max(y,0)$) where μ is the location parameter, $\psi > 0$ is a scale parameter, and ξ is a shape parameter. GEV can be seen as the integration in the same *psf* of the Gumbel distribution (limit $\xi \to 0$), Fréchet distribution ($\xi > 0$), and Weibull distribution ($\xi < 0$).

1.1.2 Exceedances Over Threshold

If the researcher needs to work only with extreme values above an specific threshold, Pickands (1971) showed that the GEV has a GPD approximation where shape ξ parameter in previous equation is the same parameter for next equation for GPD,

$$G(y,\sigma,\xi) = 1-\left(1+\xi\frac{y}{\sigma}\right)_+^{-\frac{1}{\xi}},$$

1.1.3 Poisson-GPD

If a rescale of the variable indexes above the threshold is performed, then the exceedances over threshold approach can be seen as a point process, namely, an homogeneous Poisson Process where:

1. The number of exceedances above the threshold has a Poisson distribution with mean λ
2. The excess values follow a GPD with $N \leq 1$

Its cumulative distribution function *cdf* is

$$F(y) = \exp\left\{-\lambda\left(1+\xi\frac{y-\mu}{\sigma}\right)_{+}^{-\frac{1}{\xi}}\right\},$$

1.2 Problem Statement and Motvation

Wind forces are important for infrastructure design, see Comarazamy (2005). For a civil engineer, designer of different types of structures, main forces to consider when designing a structure, for instance a bridge or a building, are a) dead load due to the weight of the structure, and b) live load due to earthquake and wind. For the study area, a developing country, the structure design standard has defined in great detail, all aspects related to seismic forces, and dead forces, but lack of detail in wind forces, actually, current map is 20 years outdated, and it is not appropriate for all types of structures, because it only includes two return periods. Additionally, it is well known that in recent years there have been accelerated changes in the climate of the planet, including issues related to winds, aspect that is reflected in frequent failures of structures due to wind forces, see Council (1994), as is stated in Rezapour & Baldock (2014), wind forces are able to completely destroy different types of infrastructures, reason why last five decades the way to assess wind loads in structural design has had remarkable changes, see Roberts (2012).

A complete study of extreme wind forces, need to address separately and using different scientific approaches, hurricane and non-hurricane data, to allow a final research product as the integration of the results in both fronts, see Engineers (2017). In the study area, hurricane winds are only present inland in the Caribbean Sea, therefore, only affects directly to 'San Andres y Providencia' island - one (1) of thirty-three (33) states. In 1102 of 1103 municipalities (more than 99%), the issue of non-hurricane winds is the only one relevant, and in addition, this lacks recent studies and research, however, all municipalities located near to the northern onshore border, may be impacted by side effects of hurricanes.

As a note of clarification on the motivation to carry out the research, the author of this thesis is a civil engineer, from Colombia (the study area), and has developed previous research work with 'Universidad de los Andes', related to geoinformatics, and analysis and evaluation of natural risks. Due to the proximity of the University with the Colombian Association of Seismic Engineering - AIS, the opportunity to contribute to the update of the standard has arisen.

1.3 Knowledge Gap

Nowadays, methodologies to deal with the inference of extreme wind maps are quite mature and advanced, and many of them already implemented and ready for use, reason why the main contribution in this research is not related to the theoretical foundations of the methods themselves, but to application of the method in a particular case in developing countries, where the lack of data plays a decisive role in achieving the results, see ADB (2014). Thereby,

the gaps in which this research aims to contribute, are related to the use of alternative data sources, and how to meet the downscaling challenge, considering that the main drawback of the research is the lack of field measurement data coming from weather stations.

1.4 Research Aim and Objectives

Main aim of this research is the estimation of wind extreme velocities to be used as input loads for the design of different types of structures, considering its risk categories, and covering any place in the whole study area.

Specific objectives are:

1. Complement the lack of field measured wind data, with other sources of information, then, analyze and compare different time series, to select and use the best data source (or combination of sources) for research, based on objective criteria, for instance similitude, completeness, coverage, etcetera, to achieve in this way a formal support for the decision made in this regard, in case of downscaling issue.
2. Select and apply a suitable probabilistic method to infer wind maps for infrastructure design, that allows to fulfill wind load requirements defined for the respective authority in the study area.
3. Estimate needed extreme wind values for the stations in the selected input data source, considering non-hurricane approaches.
4. Allow the comparison of wind extreme values estimations, using different methods, in order to verify and calibrate output results.
5. Generate needed continuous non-hurricane wind maps, using the most suitable spatial interpolation technique, considering the specific characteristics of the input data source, advantages, and disadvantages of the selected methods.
6. Combine output maps from non-hurricane analysis, with existing hurricane studies to allow the inclusion of final maps in the design standard.

1.5 Research Question

Main question of this research is directed to calculate future extreme velocities (return levels) for infrastructure design, then the research question could be

What extreme velocities (return levels) need to be used as load design forces for structures of different use category, in the study area?

If we remember that, for the case study area (Colombia), there are predefined requirements or mean return intervals - MRI to design structures depending of it use category, and that this MRI values are 700, 1700, and 3000 years, the research question could be more specific.

What extreme velocities (return levels) will be equaled or exceeded with a probability equal to $\frac{1}{MRI}$ in a given year?

What extreme velocities (return levels) will be equaled or exceeded only one time in the period defined for this specific MRIs: 700, 1700, and 3000 years?

If we consider not only the annual exceedance probability $\frac{1}{MRI}$, but also the exposure time (compound probability), understood as the time the structure will be in use, then the question will be

What extreme velocities (return levels) will have a occurrence compound probability of 67%, when the exposure time of the structure will be equal to the main return intervals 700, 1700, and 3000 years?

1.6 Outline

Main sections of thesis document are 1) Introduction, 2) Data, 3) Theoretical Framework, 4) Methodology, 5) Results and Discussion, 5) Conclusions, and 6) five Annexes, from A to E.

After introduction, in second section **Data**, main information about data sources IDEAM, ISD, and ERA5 are described, including at the end, details for data download and organization, topic that is complemented for ERA5, with the content of the Annex C - ERA5 Data Download and Integration.

Theoretical framework section is dedicated to introduce statistical concepts that are basis for the investigation, both in **probability distributions** (density function - pdf, distribution function - cdf, percent point function - ppf, and hazard function - hf) and in **extreme analysis** (annual exceedance probability - P_e, recurrence interval - MRI, and compound exceedance probability - P_n). Later, it is described in more detail, topics related to **extreme value analysis** (peaks over threshold with generalized Pareto - POT-GPD, and peaks over threshold with Poisson process - POT-PP), and at the end, a summary report is done about **wind load requirements** for the study, which is foundation for addressing the research.

In methodology are described main processes needed to meet the objectives and answer the research question, which, broadly speaking, are data standardization, downscaling support, POT-PP, spatial interpolation and integration with hurricane data.

Results and discussion section, shows first all results for data standardization and comparison to support the downscaling issue, second, all POT-PP results are set out for one ISD station, then all output maps for ISD and ERA5 data sources are exhibited, including discussions without bias, about the good and the bad of those finals results. These discussions are complemented by the following conclusions section.

To finalize the document, a series of appendices were created, to facilitate the reproducibility of the research. Appendix A, contains *research R code*, but it is necessary to keep in mind that the code provided by Dr. Adam Pintar, to do the de-clustering and thresholding in POT-PP, is not there, because its publication and distribution is not authorized. Appendix B contains all *results in digital format*. Appendix C compliments the information needed to *download*

and organize data from the sources, mainly with details related to ERA5. As the document for the thesis was done using package ´thesisdown´, which is based in ´bookdown´, the most important *document R code* to create the thesis document, mainly graphics, is shown in Appendix D, then at the end, in Appendix E, an *user manual* is presented, in order to provide instruction to apply the same methodology in a different case study, and provided using R code.

Chapter 2

Data

Input data is made up of three different sources a) IDEAM - Institute of Hydrology, Meteorology and Environmental Studies of Colombia `http://www.ideam.gov.co`, b) ISD - Integrated Surface Database `https://www.ncdc.noaa.gov/isd`, and c) ERA5 climate reanalysis `https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5`.

Table 2.1: Datasets description

Institution	Dataset	Details
IDEAM	Historical records at weather stations	IDEAM is responsible for the instalation, maintenance and management of all kind of weather stations located everywhere along the country
NOAA	ISD	ISD (Integrated Surface Database. NOAA's National Centers for Environmental Information - NCEI) Lite: A subset from the full ISD dataset containing eight common surface parameters in a fixed-width format free of duplicate values, sub-hourly data, and complicated flags.
ECMWF	ERA5	ERA5 is a reanalysis dataset with hourly estimates of atmospheric variables with horizontal resolution of 0.25^{o} (33 kilómeters), this is equally spaced cells every 0.25 degrees

Table 2.2: Datasets variables

Dataset	Variables	Description
IDEAM	vv_aut_2	Instantaneous wind velocity each two (2) minutes
	vv_aut_10	Instantaneous wind velocity each ten (10) minutes
ISD	v5	Maximun hourly five seconds (5-s) wind gust velocity
ERA5	fg10	10 metre wind gust since previous post-processing
	fsr	Forecast Surface Roughness

Table 2.3: Variables units and time

Variable	Units	Time	Stations
vvmx_aut_60	meters per second	Variable from 2001 until today. Irregular time series.	203
Wind speed	meters per second	Variable from 1941 until today. Note: There is too much variability in time (start, end, and time range) for each station. Irregutal time series.	101
fg10	meters per second	1979-Today	3381
fsr	meters per second	1979-Today	3381

Ideal data source to create extreme wind speeds maps should be field observed data from IDEAM, but there are not enough number of stations around the study area to represent all the local wind variability in a huge country with multiple variety of climates and and changing thermal floors, but there are other important motivation to include different sources trying to improve output results:

- As just mentioned, low quantity of IDEAM stations
- There are uncertainties related to the way IDEAM anemometers are registering data, then comparison with other data sources are needed to be able to do appropriate data standardization, needed as a prerequisite to the analysis.
- There is no time continuity in the registration of IDEAM data. Historical time series are different and variable in each station.

Importance of ISD database for this study is based on the fact that post-processed ISD database has wind extreme values, and it was used to create extreme wind maps for United States. ISD allows comparison with IDEAM records to take better decisions in order to do needed data standardization. Despite that ERA5 data are not observed data, but forecast, its main advantage is data availability to assess the local climatic variance every 0.25 square decimal degrees.

2.1 IDEAM

Historical observed wind speeds from 203 stations in Colombia are managed by the official environmental authority IDEAM. Table 2.4 shows a sample of five IDEAM stations. Figure 2.1 shows a map of IDEAM stations.

Table 2.4: IDEAM Stations Sample

Name[Code]	Latitud	Longitud
EMAS - AUT [26155230]	5.09	-75.51
SAN BENITO - AUT [25025380]	9.16	-75.04
AEROPUERTO ALFONSO LOPEZ - [28025502]	10.44	-73.25
TIBAITATA - AUT [21206990]	4.69	-74.21
ELDORADO CATAM - AUT [21205791]	4.71	-74.15

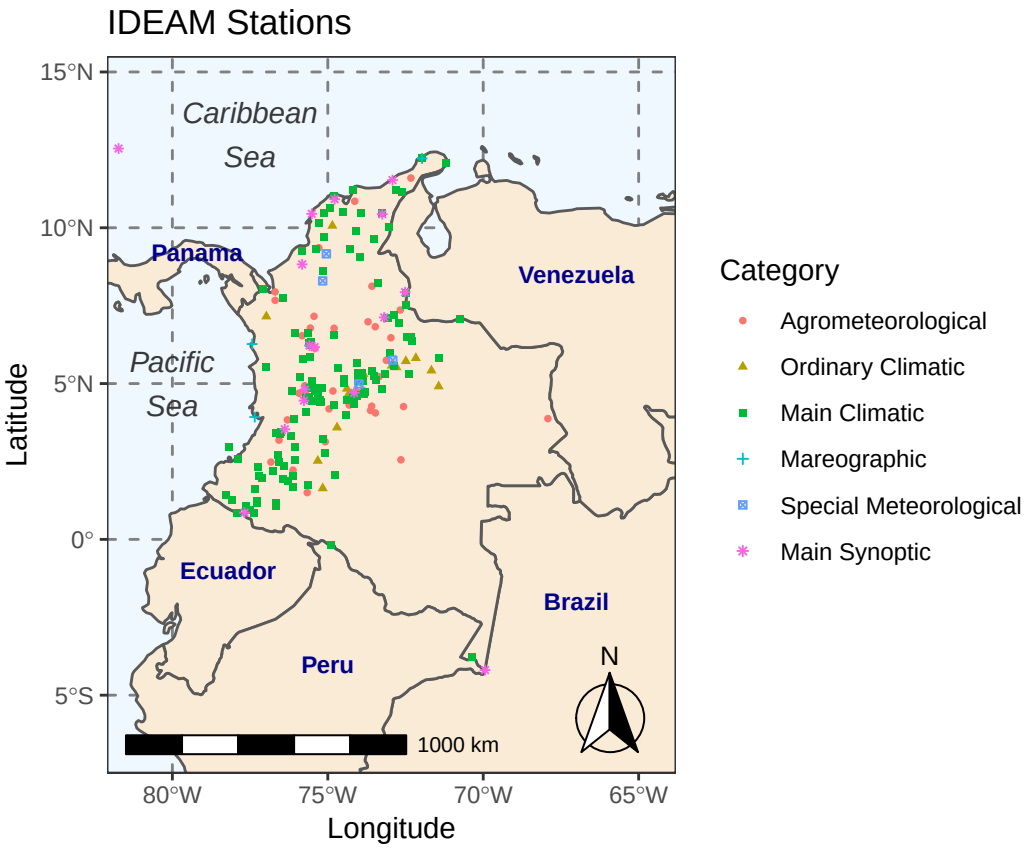

Figure 2.1: IDEAM Stations. Colombia

Following, the time series, autocorrelation function, and partial autocorrelation function, for IDEAM station "21205791" will be displayed.

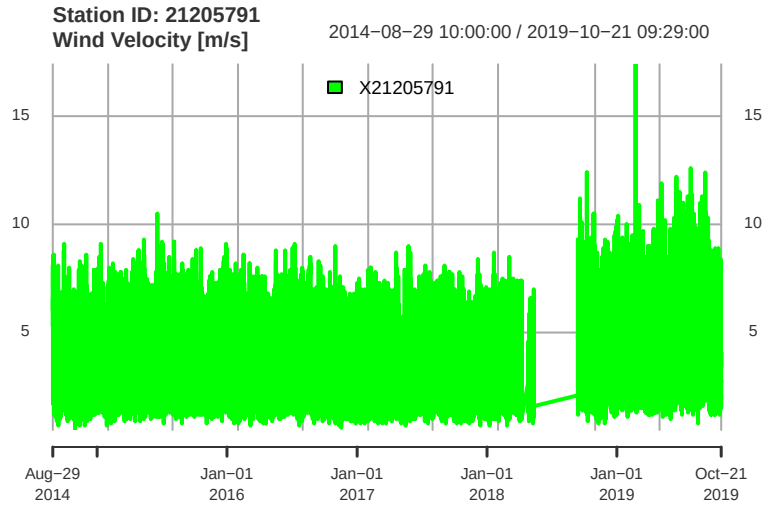

Figure 2.2: IDEAM Station ELDORADO CATAM - AUT - Time Series

2.2 ISD

ISD is a database with environmental variables, among them extreme wind speeds. ISD has data for the whole planet, and is based on observed data at meteorological stations in each country, which means that for Colombia is based on IDEAM data. Main advantage is data availability at neighbor countries and specialized post-processing made by NOAA's National Centers for Environmental Information - NCEI in United States, which facilitates its use. Table 2.5 shows a sample of five ISD stations. Figure 2.3 shows a map of ISD stations.

Table 2.5: ISD Stations Sample

Code	Name	Latitud	Longitud
804400	BARINAS	8.62	-70.22
800810	ALTO CURICHE	7.05	-76.35
801000	BAHIA SOLANO / JOSE MUTIS	6.18	-77.40
802590	ALFONSO BONILLA ARAGON INTL	3.54	-76.38
803150	BENITO SALAS	2.95	-75.29

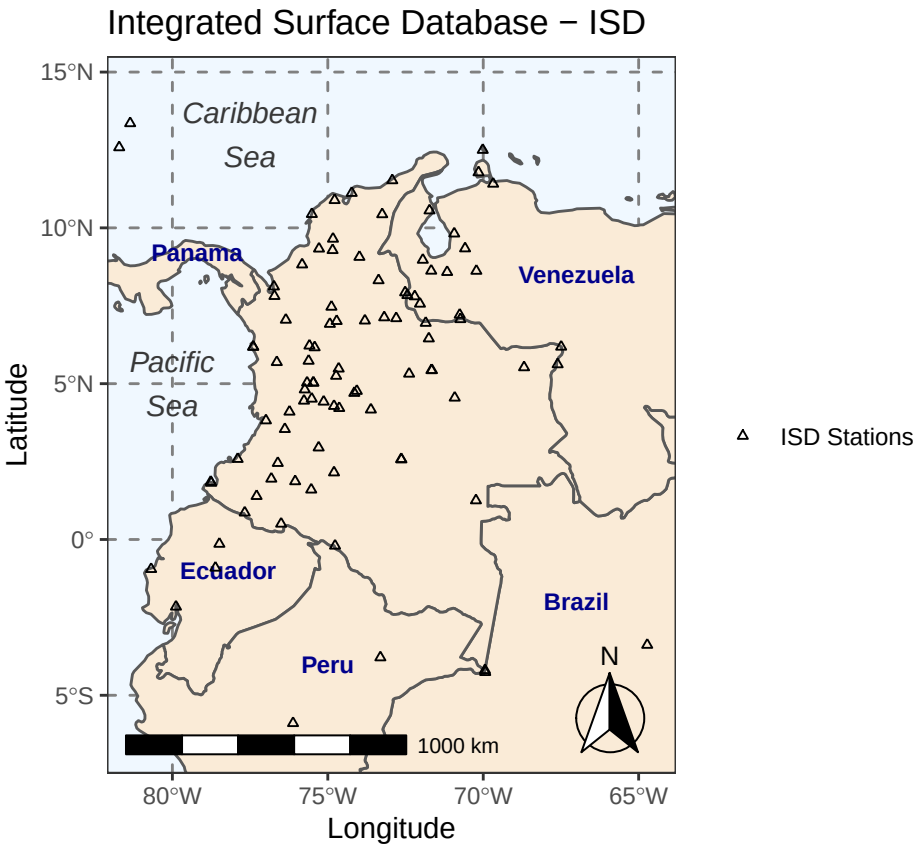

Figure 2.3: ISD Stations. Colombia and surroundings

Following, the time series, autocorrelation function, and partial autocorrelation function, for ISD station "802590" will be displayed.

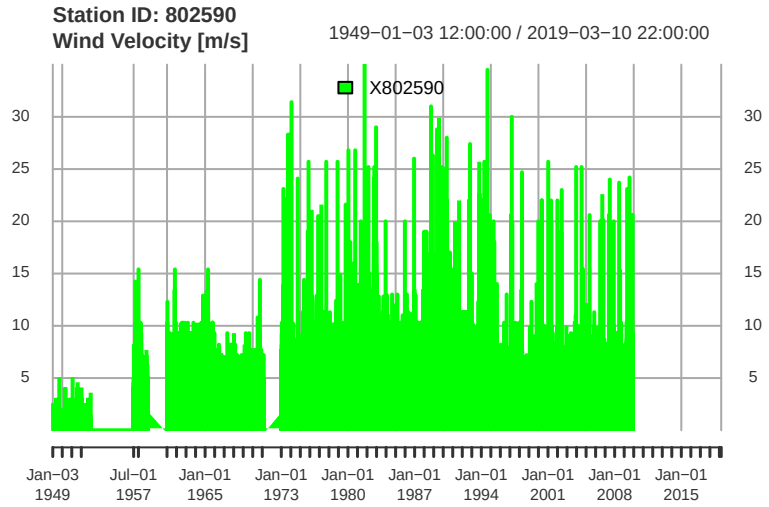

Figure 2.4: ISD Station ALFONSO BONILLA ARAGON INTL - Time Series

2.3 ERA5

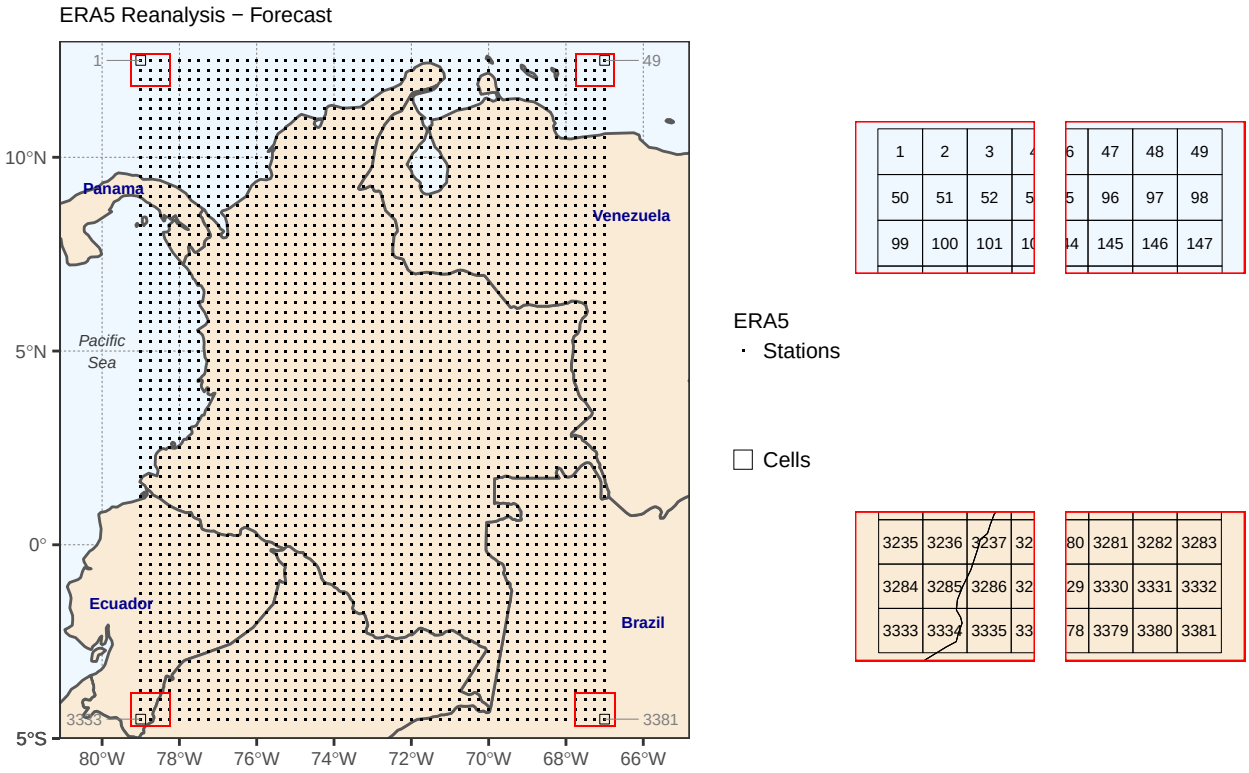

Figure 2.5: ERA5 Cells and Stations (cells centers). 49 cols by 69 rows. Cell size 0.25 decimal degrees (aprox 28 km in Colombia). Station IDs from 1 (lon=-79, lat=12.5) to 3381 (lon=-67, lat=-4.5)

ERA5 is forecast reanalysis data processed by the *European Centre for Medium-Range Weather Forecasts* - ECMWF with wind speeds time series in square cells *matrix of pixels* of 0.25 decimal degrees covering the whole planet. For the study area was extracted a raster of 69 rows by 49 columns in format netCDF. Figure 2.5 shows a map of ERA5 stations (cells centers).

2.4 Data Download and Data Organization

All data sources had different mechanisms for downloading. For IDEAM, the official procedure is through a written request using the e-mail *atencionalciudadano@ideam.gov.co*, then they will provide a link to get the information. For ISD, all files are available in the ftp site `ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/`, organized in folders by years, then a gzip file is available, with the station name in the format *ID-99999-YYYY.gz*, where ID is the USAF-ISD station identifier. For ISD, there are many files by station, one file for each year with data available. For ERA5 it is possible to make a request using a Python script, but since there is a size limit for downloading, it is necessary to split the request, then, use console commands to create an unified netCDF file. Files with all IDEAM and ISD stations are available in the Annex A - Results - Digital Files. For the Python code, and the commands to join netCDF files of ERA5 data source, see the Annex C - ERA5 Data Download and Integration.

Chapter 3

Theoretical Framework

3.1 Probability Concepts

Poisson process is an stochastic method that relies in the concepts of probability distributions. The main functions related to probability for extreme value analysis will be described below.

3.1.1 Probability Density Function - *pdf*

Pdf defines the probability that a continuous variable falls between two points, this is, in *pdf* the probability is related to the area below the curve (integral) between two points, as for continuous probability distributions the probability at a single point is zero. The term density is directly related to the probability of a portion of the curve, if the density function has high values the probability will be greater in comparison with the same portion of curve for low values.

$$\int_a^b f(x)dx = Pr[a \leq X \leq b]$$

Equation (3.1) is the Gumbel *pdf*.

$$\mathrm{f(x)} = \frac{1}{\beta} \exp\left\{-\frac{\mathrm{x}-\mu}{\beta}\right\} \exp\left\{-\exp\left\{-\left(\frac{\mathrm{x}-\mu}{\beta}\right)\right\}\right\}, \quad -\infty < \mathrm{x} < \infty \tag{3.1}$$

where $\exp\{.\} \mapsto \mathrm{e}^{\{.\}}$, β is the scale parameter, and μ is the location parameter. Location (μ) has the effect to shift the *pdf* to left or right along 'x' axis, thus, if location value is changed the effect is a movement of *pdf* to the left (small value for location), or to the right (big value for location). Scale has the effect to stretch ($\beta > 1$) of compress ($0 < \beta < 1$) the *pdf*, if scale parameter is close to zero the *pdf* approaches a spike.

Figure 3.1 shows *pdf* with location (μ) = 100 and scale (β) = 40, using Equation (3.1).

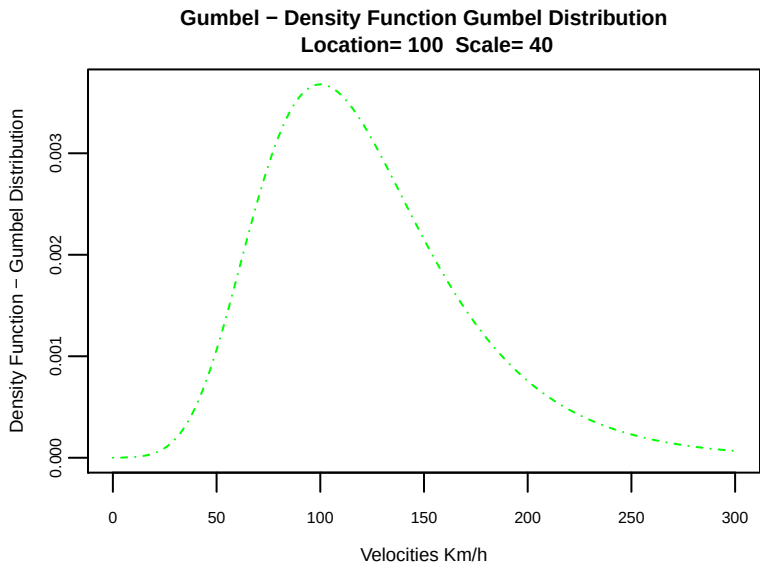

Figure 3.1: Gumbel pdf

Figure 3.2 shows *pdf* with location (μ) = 100 and scale (β) = 40, using function `dgumbel` of the package `RcmdrMisc`

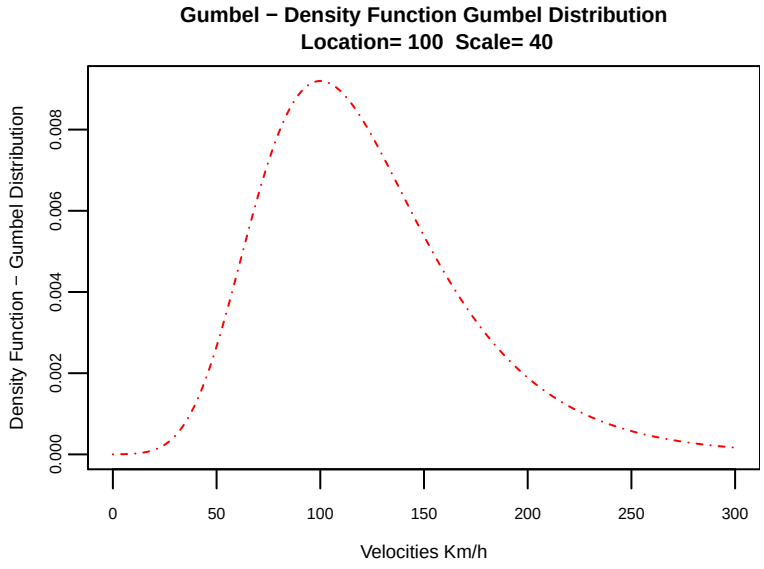

Figure 3.2: Gumbel pdf - dgumbel function

3.1.2 Cumulative Distribution Function - *cdf*

Cdf is the probability of taking a value less than or equal to x. That is

$$F(x) = Pr[X < x] = \alpha$$

For a continuous variable, *cdf* can be expressed as the integral of its *pdf*.

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Equation (3.2) is the Gumbel *cdf*.

$$\mathrm{F(x)} = \exp\left\{-\exp\left[-\left(\frac{\mathrm{x}-\mu}{\beta}\right)\right]\right\}, \quad -\infty < \mathrm{x} < \infty \tag{3.2}$$

Figure 3.3 shows Gumbel *cdf* with location (μ) = 100 and scale (β) = 40, using Equation (3.2). As previously done with *pdf*, similar result can be achieved using function `pgumbel` of package `RcmdrMisc`.

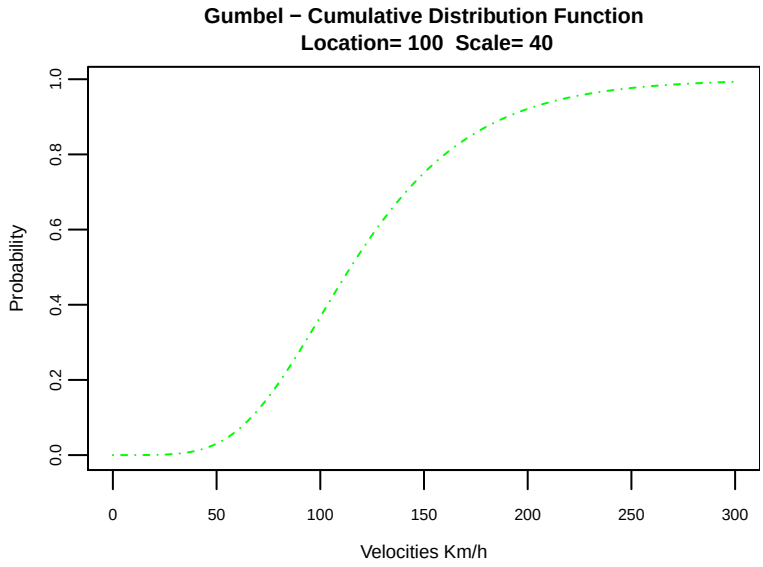

Figure 3.3: Gumbel cdf

3.1.3 Percent Point Function - *ppf*

Ppf is the inverse of *cdf*, also called the *quantile* function. This is, from a specific probability get the corresponding value x of the variable.

$$x = G(\alpha) = G(F(x))$$

Equation (3.3) is the Gumbel *ppf*.

$$\mathrm{G}(\alpha) = \mu - \beta\ln(-\ln(\alpha)) \quad 0 < \alpha < 1 \tag{3.3}$$

Figure 3.4 shows Gumbel *ppf*, using Equation (3.3). Similar result can be achieved using function `qgumbel` of package `RcmdrMisc`.

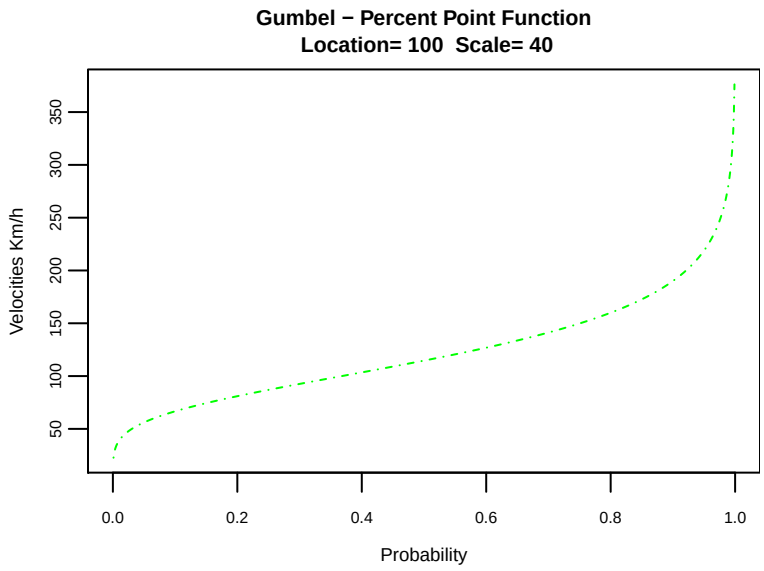

Figure 3.4: Gumbel ppf

3.1.4 Hazard Function - *hf*

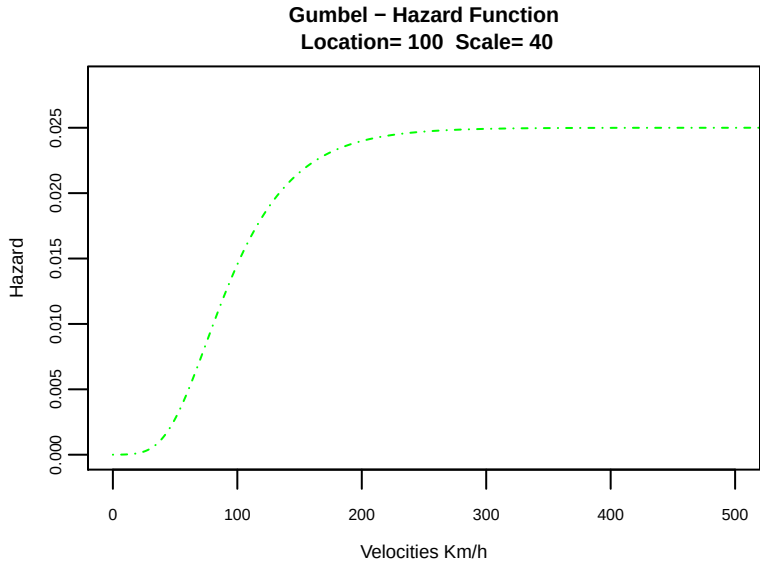

Figure 3.5: Gumbel hf

Using $S(x) = 1 - F(x)$ as survival function -*sf*, the probability that a variable takes a value greater than x $S(x) = Pr[X > x] = 1 - F(x)$, the *hf* is the ratio between *pdf* and *sf*.

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$$

Equation (3.4) is the Gumbel *ppf*.

$$\mathrm{h(x)} = \frac{1}{\beta} \frac{\exp(-(\mathrm{x}-\mu)/\beta)}{\exp(\exp(-(\mathrm{x}-\mu)/\beta)) - 1} \tag{3.4}$$

Figure 3.5 shows Gumbel *hf*, using Equation (3.4).

3.2 Statistical Concepts For Extreme Analysis

In order to approach the extreme value analysis, some statistical concepts are needed to understand the theoretical framework behind this knowledge area. In this section will be introduced the concepts annual exceedance probability, mean recurrence interval - MRI, exposure time, and compound probability for any given exposure time and MRI.

As an hypothetical example, a simulated database of extreme wind speed will be used. This database is supposed to have 10.000 years of simulated wind speeds.

3.2.1 Annual Exceedance Probability - P_e

Using the previously described database, a question arises to calculate the probability to exceed the highest probable loss due to the simulated winds. It is possible to conclude that there is only one event grater or equal (in this case equal) to the highest probable causing loss in 10.000 years, and it is the *highest wind*. If we sort the database by wind magnitude in descending order (small winds last), the question is solved calculating the annual exceedance probability Pe with next formula

Figure 3.6: Sorted Winds by Magnitude - wind simulation database

$$P_e = \frac{Event\,index\,after\,descending\,sorting}{Years\,of\,simulations} = \frac{1}{10.000} = 0.001 = 0.01\%$$

because the highest wind will be the first in the sorted list. Same exercise can be done with all winds to construct the annual exceedance probability curve, that in this case will represent the probability to equal or exceed different probable losses due to wind.

3.2.2 Return Period - Mean Recurrence Interval - MRI

Continuing with the previous section, if the inverse of the exceedance probability is taken, the return period (in years) is obtained. The return period or Mean Recurrence Interval -

MRI is associated with an specific return level (wind extreme velocity). MRI is the numbers of years (N) needed to obtain 63% of change that the corresponding return level will occur at least one time in that period. The return level is expected to be exceeded on average once every N-years. The annual exceedance probability of the return level corresponding to N-years of MRI, is $P_e = \frac{1}{MRI} = \frac{1}{N}$.

For an specific wind extreme event A, the probability that the event will occur in a period equal to MRI years is 63%. If we analyze for the same period a strongest wind extreme event B, its occurrence probability will be lest than 67%. If the purpose of this research is to design infrastructure considering wind loads, the structure will be more resistant to wind if we design with stronger winds, this is high MRIs, and low annual exceedance probability. Common approach for infrastructure design, considering any type of load (earthquake, wind, etc) is to choose high MRI according to the importance/use/risk/type of the structure. For highly important structures, like hospitals or coliseums, where the risk of collapse must be diminished, the MRI used to design is higher in comparison to common structures (for instance a normal house), which implies less risks for its use and importance.

$$\mathrm{P_e} = \begin{cases} 1 - \exp\left(-\frac{1}{MRI}\right), & for\, MRI < 10\, years \\ \frac{1}{MRI}, & for\, MRI \geq 10\, years \end{cases}$$

3.2.3 Compound Exceedance Probability - Pn

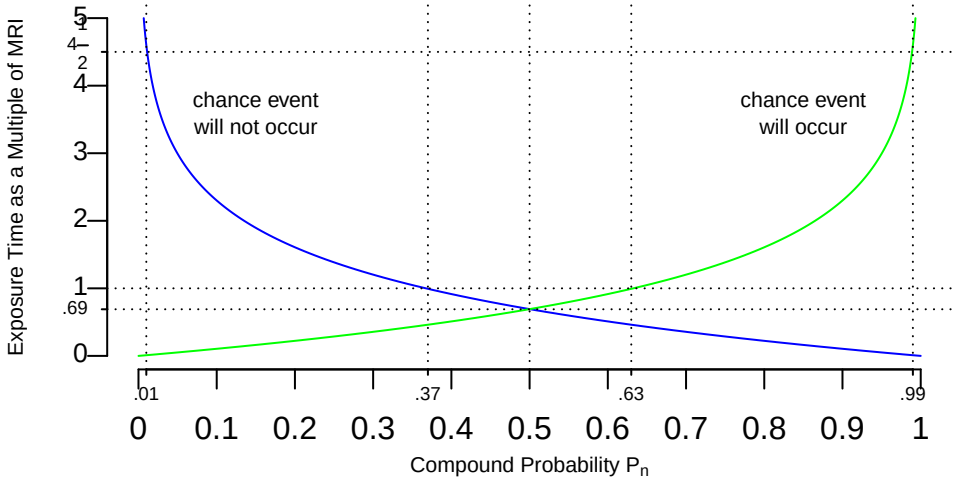

Figure 3.7: Compound Probability

If time of exposure is consider, understood as time the structure will be in use, it is possible to have a compound probability P_n, where n is the exposure period. P_n is the probability that the extreme wind speed will be equaled or exceeded at least one time in n years, and is related with the occurrence probability, but also is possible to calculate the non-occurrence compound probability (probability that the event will not occur).

$$\mathrm{P_n} = \begin{cases} 1 - \left(1 - \dfrac{1}{MRI}\right)^n, \ occurrence\, probability \\ \left(1 - \dfrac{1}{MRI}\right)^n, \ non - occurrence\, probability \end{cases}$$

If it is consider exposure time as a multiple of return period, the resulting Figure 3.7, shows that:

- When exposure time is .69% of the return period, then probability (occurrence and non-occurrence) will be 50%
- As was stated previously, when exposure time is equal to return period, then the probability that the extreme wind speed (return level) occur is 63%, and 37% for the non occurrence probability.
- If exposure time is 4.5 times the return period, there is a 99% of change that the return level will occur.

The example discussed here was presented as an instrument to introduce important concepts, nonetheless,there are specialized approaches to deal with extreme value analysis which will be discussed in Extreme Value Analysis Overview and more in detail in Peaks Over Threshold - Poisson Process. In summary, is necessary to fit the data over a specific threshold to an extreme value distribution, and P_e will be $1 - F(y)$, with F(y) as the *cdf*, and MRI as $\frac{1}{1-F(y)}$.

3.3 Extreme Value Analysis Overview

Analysis of extreme values is related with statistical inference to calculate probabilities of extreme events. Main methods to analyze extreme data are epochal, Peaks Over Threshold - POT, and extreme index. The epochal method, also known as block maxima, uses the most extreme value for a specific frame of time, typically, one year. POT is based in the selection of a single threshold value to do the analysis only with values above the threshold. But there are different POT approaches, the most common one is Generalized Pareto Distribution - POT-GPD, but also it is possible to use the Poisson process approach.

In both methods (Epochal and POT), the first step is to fit the data to an appropriate probability distribution model, among them the most used are, - Extreme Value Type I (Gumbel), Extreme Value Type II (Fréchet), Weibull, Generalized Pareto - GPD, and Generalized Extreme Value - GEV.

Distribution models are fitted based in the estimation of its parameters, commonly called location, scale and shape, nonetheless each model has its own parameters names. There are different methods to estimate parameters, among them, - method of moments (modified moments - see Kubler (1994), and L moments - see Hosking & Wallis (1997)), - method of maximum likelihood MLE, see Harris & Stocker (1998), which is problematic for GPD and GEV, - probability plot correlation coefficient, and - elemental percentiles (for GPD and GEV)

Once candidate parameters are available, it is necessary to assess the goodness of fit of the selected model, using one of the next methods, - Kolmogorov-Smirnov (KS) goodness of fit test, and - Anderson-Darling goodness of fit test. Here a visual assessment is also useful using a probability plot or a kernel density plot with the fitted *pdf* overlaid.

The main use of the fitted model is the estimation of mean return intervals - MRI, and extreme wind speeds (return levels),

$$MRI = \frac{1}{1 - F(y)}$$

with $F(y)$ as the *cdf*. If $1 - F(y)$ is the annual exceedance probability, MRI is its inverse, see Simiu & Scanlan (1996) for more details about MRI. If y is solved from previous equation using a given MRI of N-years, its value represents the Y_N wind speed return level,

$$Y_N = G\left(1 - \frac{1}{\lambda N}\right)$$

where G is the *ppf* (quantile function) and λ is the number of wind speeds over the threshold per year.

The CRAN Task View "Extreme Value Analysis" `https://cran.r-project.org/web/views/ExtremeValue.html` shows available **R** for block maxima, POT by GPD, and external indexes estimation approaches. Most important to consider are `evd`, `extremes`, `evir`, `POT`, `extremeStat`, `ismev`, and `Renext`.

3.3.1 POT-GPD

In POT using Pareto distribution, the magnitude of the observations above the threshold are assumed a) to be independent random variables with the same generalized Pareto as probability distribution, σ as scale, and ξ as tail length, and corresponding times are assumed b) to follow a one dimensional homogeneous Poisson process with γ as parameter. The *cdf* of POT-GPD is $F(y) = 1 - \left(1 - \xi\frac{y-b}{\sigma}\right)_+^{-\frac{1}{\xi}}$, where b is the threshold. In both GPD (magnitude), and $1D$ Poisson process (time), it is not possible to differentiate between thunderstorm and non-thunderstorm wind types.

In POT-GPD, to calculate return levels (RL), Y_N, corresponding to the N-years return period, next equation is used,

$$Y_N = G\left(y, 1 - \frac{1}{\lambda N}\right)$$

Where G is the quantile function (*ppf*), and the value of the probability passed to the G function, has to be modified with the λ parameter. λ is the number of wind speed events over the threshold per year.

3.4 Peaks Over Threshold Poisson Process POT-PP

According to Pintar et al. (2015) the stochastic Poisson Process - PP is mainly defined by its intensity function. As the intensity function is not uniform over the domain, the PP considered here is non-homogeneous, and due to the intensity function dependency of magnitude and time, it is also bi-dimensional. PP was described for the first time in Pickands (1971), then extended in Smith (1989).

$$\lambda\,(\mathrm{y},\mathrm{t})\begin{cases}\lambda_t(y),\ for\,t\,in\,thunderstorm\,period\\ \lambda_{nt}(y),\ for\,t\,in\,non-thunderstorm\,period\end{cases} \tag{3.5}$$

Generic Equation (3.5) shows the intensity function, which is defined in the domain $D = D_t \cup D_{nt}$, and allow to fit the PP at each station to the observed data $\{t_i, y_i\}_{i=1}^{I}$, for all the times (t_i) of threshold crossing observations, and its corresponding wind speeds magnitudes (y_i). Thus, only data above the threshold (POT) are used.

Intensity function of the PP is defined in Smith (2004),

$$\frac{1}{\psi_\mathrm{t}}\left(1+\zeta_\mathrm{t}\frac{\mathrm{y}-\omega_\mathrm{t}}{\psi_\mathrm{t}}\right)_{+}^{-\frac{1}{\zeta_\mathrm{t}}-1} \tag{3.6}$$

Where, at a given time t, parameter $shape = \zeta_t$ controls the tail length of the intensity function, and the other two parameters ω_t and ψ_t define the location and scale of the intensity function.

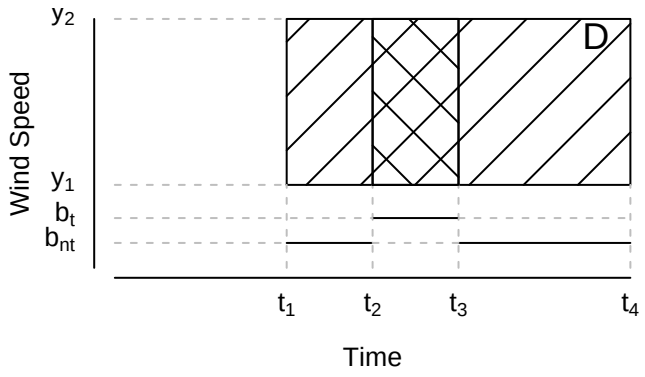

Figure 3.8: Domain off the Poisson Process - PP

Figure 3.8 represent the domain D of PP. In time, the domain represents the station service period from first sample t_1 to last sample t_4. D is the union of all thunderstorm periods D_t (from t_2 to t_3), and all non-thunderstorm periods D_{nt} (periods t_1 to t_2 and t_3 to t_4). In magnitude, only thunderstorm data above its threshold b_t, and only non-thunderstorm data above its threshold b_{nt} are used.

Thunderstorms and non-thunderstorms are modeled independently:

1. Observations in domain D follow a Poisson distribution with mean $\int_D \lambda(t, y)\, dt\, dy$
2. For each disjoint sub-domain D_1 or D_2 inside D, the observations in D_1 or D_2 are independent random variables.

Visual representation of the intensity function for PP can be seen in Figure 3.9. In vertical axis, two surfaces were drawn representing independent intensity functions for thunderstorm $\lambda_t(y)$ and for non-thunderstorm $\lambda_{nt}(y)$. The volume under each surface for its corresponding time periods and peak (over threshold) velocities, is the mean of PP.

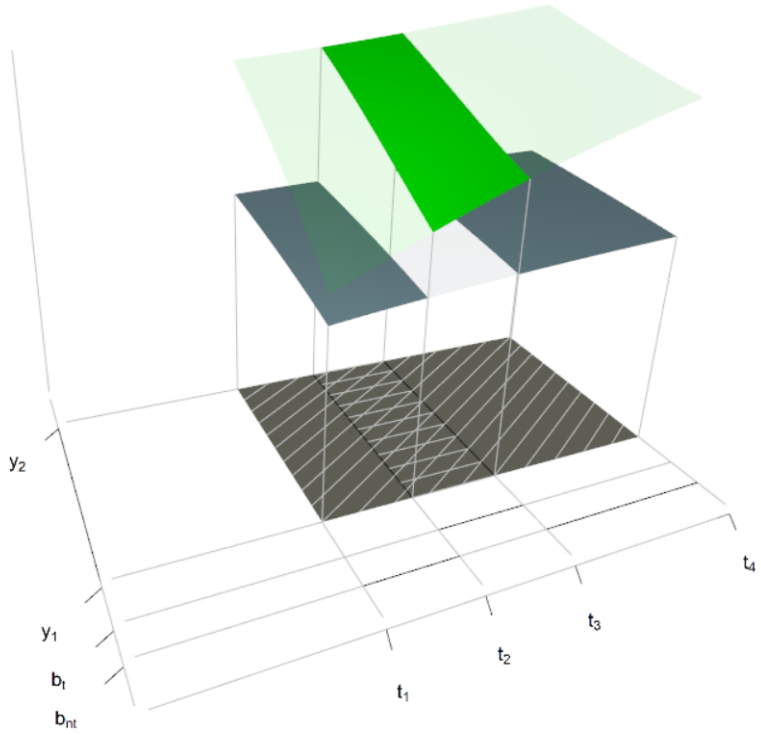

Figure 3.9: Volume under surfaces represents the mean of PP

To fit the intensity function to the data, the method of maximum likelihood is used to estimate its parameters, *scale* $= \psi$, *location* $= \omega$, and *shape* $= \zeta$, the selected vector of parameters η are the $\hat{\eta} = (\hat{\psi}, \hat{\omega}, \hat{\zeta})$ values that maximizes next function

$$\mathrm{L}(\eta) = \left(\prod_{\mathrm{i}=1}^{\mathrm{I}} \lambda\left(\mathrm{y_i}, \mathrm{t_i}\right) \right) \exp\left\{ - \int_{\mathrm{D}} \lambda\left(\mathrm{y}, \mathrm{t}\right) \mathrm{dy}\, \mathrm{dt} \right\} \tag{3.7}$$

The values of $\hat{\eta}$ need to be calculated using a numerical approach, because there is not analytical solution available.

Once the PP is fitted to the data, the model will provide extreme wind velocities (return levels), for different return periods (mean recurrence intervals).

A Y_N extreme wind velocity, called the return level (RL) belonging to the N-years return period, has a expected frequency to occur or to be exceeded (annual exceedance

probability)$P_e = \frac{1}{N}$, and also has a probability that the event does not occur (annual non-exceedance probability) $P_{ne} = 1 - \frac{1}{N}$. Y_N will be the resulting value of the G (ppf or quantile) function using a probability equal to P_{ne}. $Y_N = quantile(y, p = P_{ne}) = G(y, p = P_{ne}) = ppf(y, p = P_{ne})$. Y_N can be understood as the wind extreme value expected to be exceeded on average once every N years.

For PP, Y_N is the solution to the next equation, which is defined in terms of the intensity function,

$$\int_{\mathrm{Y_N}}^{\infty}\int_{0}^{1} \lambda\,(\mathrm{y,t})\,\mathrm{dydt} = \mathrm{A_t}\int_{\mathrm{Y_N}}^{\infty} \lambda_{\mathrm{t}}\,(\mathrm{y})\;\mathrm{dy} + \mathrm{A_{nt}}\int_{\mathrm{Y_N}}^{\infty} \lambda_{\mathrm{nt}}\,(\mathrm{y})\;\mathrm{dy} = \frac{1}{\mathrm{N}} \tag{3.8}$$

where A_t, is the multiplication of the average number of thunderstorm per year and the average length of a thunderstorm, taken to be 1 hour as defined in Pintar et al. (2015), and $A_{nt} = 365 - A_t$. The average length of a non-thunderstorm event is variable, and it is adjusted for each station to guarantee that $A_{nt} + A_t = 365$. Value 365 is used only, if operations with time in the dataset are performed in days.

The same thunderstorm event in considered to occur if the time lag distance between successive thunderstorm samples is small than six hours, and for non-thunderstorm this time is 4 days. For PP, all the measurements belonging to the same event (thunderstorm or non-thunderstorm), need to be de-clustered to leave only one maximum value. In other words, the number of thunderstorm in the time series is one plus the number of time lag distances grater than 6 hours, and for non-thunderstorm grater than 4 days.

3.4.1 Threshold Selection

POT-PP needs selection of the best threshold pairs b_t and b_{nt} (see Figure 3.8) that produces the optimal fit. Measurement of this threshold fitting is done through W statistics. If wind variable y, in a POT-PP approach, has a $cdf = U = F(y)$, then $F(y)$ is distributed as Uniform between 0 and 1 - Uniform(0,1), meaning that the transformation $W = -log(1-U)$ is an exponential random variable with mean one (1).

$$\mathrm{cdf} = \mathrm{U} = \mathrm{F(y)} = \mathrm{P(y \leq Y)} = \frac{\int_{\mathrm{b}}^{\mathrm{Y}} \lambda(\mathrm{y,t})\,\mathrm{dy}}{\int_{\mathrm{b}}^{\infty} \lambda(\mathrm{y,t})\,\mathrm{dy}} \tag{3.9}$$

The procedure to choose the best thresholds pairs based in W transformation, is described in methodology, section thresholding.

3.5 Wind Loads Requirements

As the output maps of this research will be used as input loads for infrastructure design, the methodology used for its creation, need to be consistent with Colombian official wind loads requirements. Colombian structure design code, from now the design standard, was created and it maintained by the Colombian Association of Seismic Engineering - AIS.

The design standard is mainly based in *minimum design loads and associated criteria for buildings and other structures - ASCE7-16* norm, see Engineers (2017). Under these circumstances, ASCE7-16 defines the minimum requirements of the research products. Especially

the chapter C26 - "wind loads - general requirements", C26.5 "wind hazard map", and C26.7 "Exposure" - pages 733 to 747. Wind speeds requirements of ASCE7-16 are based in the combination of independent non-hurricane analysis, and hurricane wind speeds simulations models. The focus of this research will be the analysis of non-hurricane wind data, however, existing results of hurricane studies will be used to present final maps with both components. In ASCE7-16, for non-hurricane wind speed, the procedure is mainly based on Pintar et al. (2015).

ASCE7-16 (page 734), requires the calculation of wind extreme return levels for specific return periods according to the risk category of the structure to be designed: risk category I - 300 years, risk category II - 700 years, risk category III - 1700 years, risk category IV - 3000 years. The design standard only requires 700, 1700 and 3000 years. In addition, extreme wind speeds for those MRI need to correspond to: - 3 second gust speeds, - at 33 ft (10 meters) above the ground, and - exposure category C (open space).

- Risk IV - This are 'indispensable buildings' that involve substantial risk. These structures that can handle toxic or explosive substances.
- Risk III - There is substantial risk because these structures that can handle toxic or explosive substances, can cause a serious economical impact, or massive interruption of activities if they fail.
- Risk II - Category 'by default', and correspond to structures not classified in others categories.
- Risk I - This structures represent low risk for life of people.

To standardize wind speeds to gust speeds ASCE7-16 proposes the curve Durst (see C. S. Durst (1960), and Figure 3.10). Durst curve is only valid for open terrain conditions, and it shows in axis y the gust factor $\frac{V_t}{V_{3600}}$, a ratio between any wind gust averaged at t seconds, V_t, and the hourly averaged wind speed V_{3600}, and in the axis x the duration t of the gust in seconds.

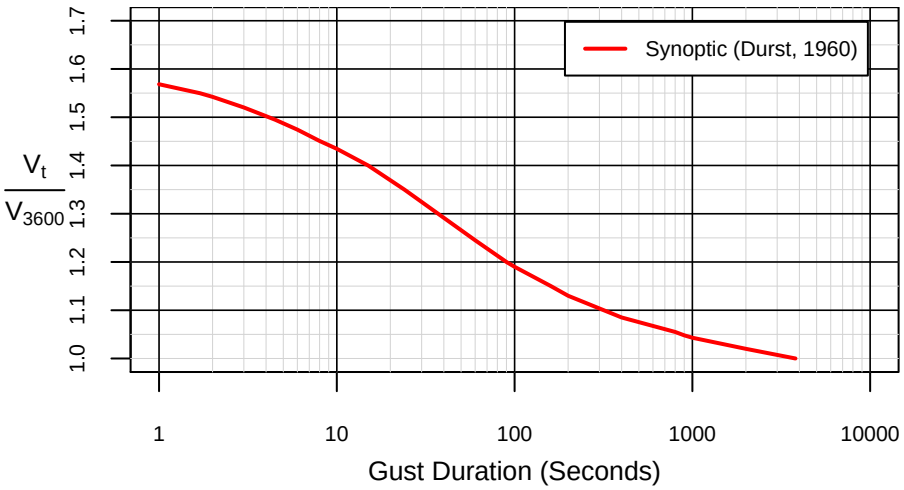

Figure 3.10: Maximum speeds averaged over t (sec), to hourly mean speed. Note: curve values taken visually from the original (use original curve for calculations!)

Chapter 4

Methodology

Figure 4.2 shows a schematic representation of the methodology, where more representative steps are identified by numbers (from 1 to 8). This research is focus in non-hurricane data, with three main elements: - data, - temporal analysis with POT-PP, and - spatial analysis to do spatial interpolation and create return levels - RL maps, for MRIs of 700, 1700, and 3000 years. Steps 1, and 3 to 7, need to be done for each available station, see Figure 4.1. With RL in each station, a continuous surface will be created, one for 700 years, next for 1700 years, and finally for 3000 years. An additional element, is the integration with existing hurricane maps to produce final maps, that will be used as input loads for infrastructure design, and will be part of the design standard.

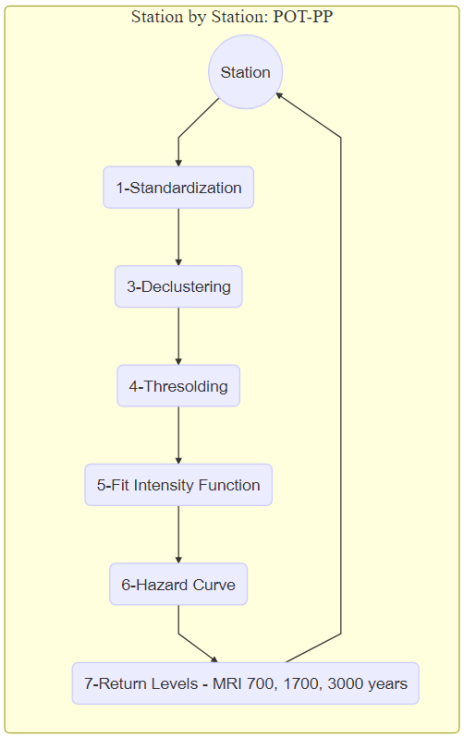

Figure 4.1: Iterative process in methodology

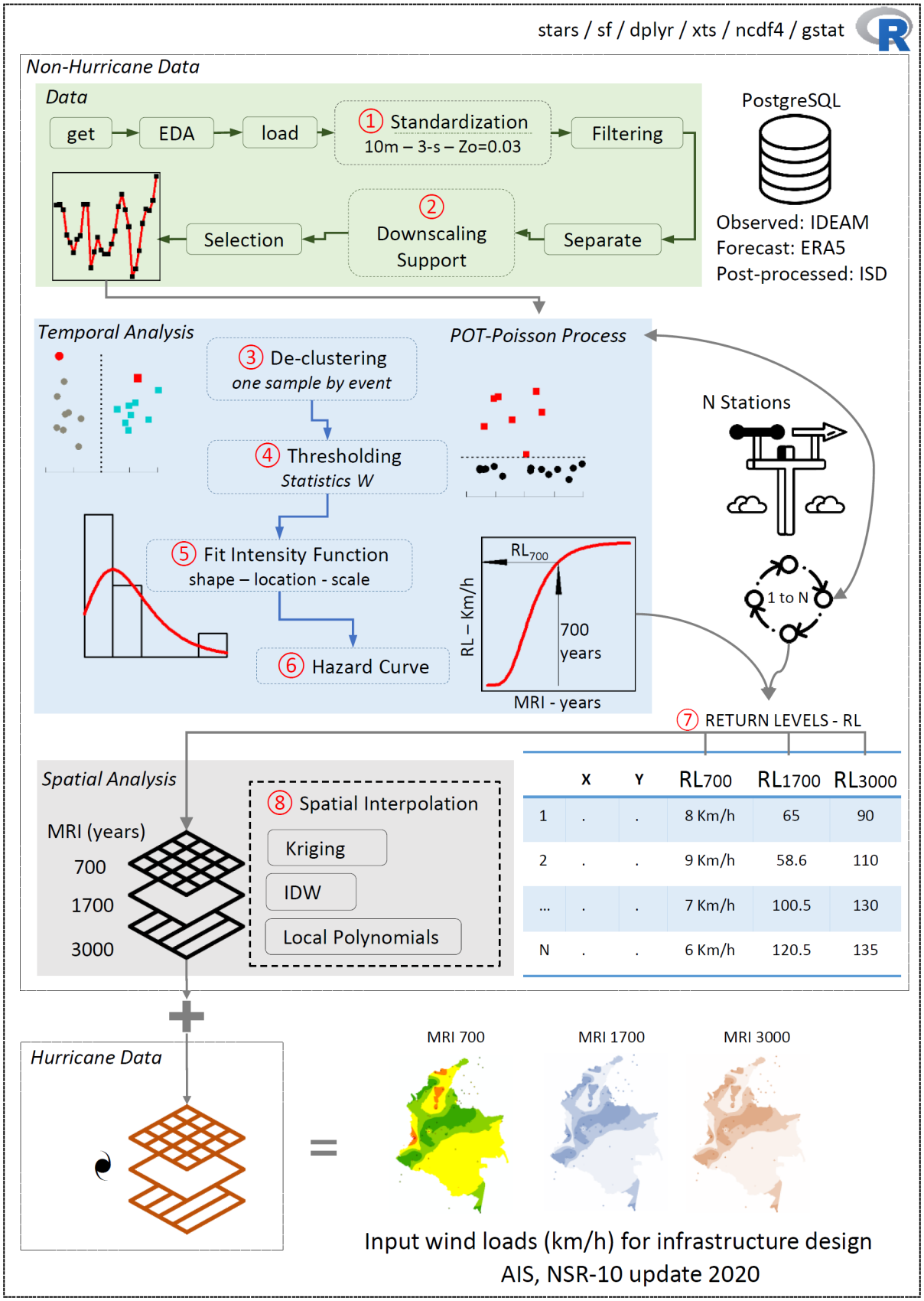

Figure 4.2: Methodology

4.1 Data Standardization

Analysis of extreme wind speeds requires data standardization as initial step. All input data must be standardized to represent three important conditions: a) anemometer height of 10 meters, b) open space terrain roughness, and c) averaging time of 3-seconds wind gust.

Parallel to the standardization activity described below, it is also important to consider for all stations involved in the analysis:

- *Separating*: As far as possible, identify each record of the time series, as thunderstorm (t) or non-thunderstorm (nt)
- *Filtering*: Remove wind speeds above $200\frac{Km}{h}$ and data pertaining to hurricane events, because the procedure with hurricane requires a different approach and need to be done independently

4.1.1 Anemometer Height - 10 m

According to the protocol for field data collection and location of methodological stations - IDEAM (2005), the anemometer (wind sensor) in installed always to a fixed height of 10 meters from the surface, as is shown in Figure 4.3, ergo, no height correction is needed.

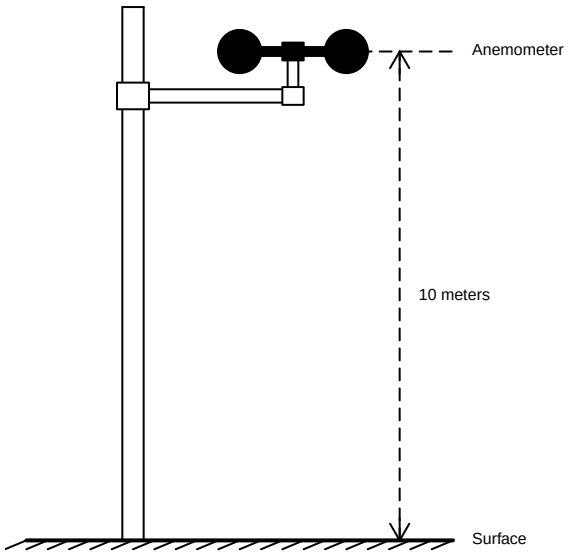

Figure 4.3: Anemometer height - 10 m

4.1.2 Surface Roughness at Open Terrain (0.03 m)

Due to the effects that the terrain has on wind speed, a correction should be applied if the station is located in a geographical space considered “not open terrain”. When terrain is open, the roughness corresponds to 0.03 meters. There are some alternative methodologies to calculate the roughness, Masters, Vickery, Bacon, & Rappaport (2010) uses the station

data, but the separation of the measurements should not exceed one minute, something difficult to obtain, and Lettau (1969) uses an empirical equation that is recommended in Engineers (2017) (page 743, equation C26.7-1), which was used here,

$$Roughness = z_0 = 0.5 * H_{ob} * \frac{S_{ob}}{A_{ob}}$$

Where H_{ob} is the average height of the obstacles, S_{ob} is the average vertical area perpendicular to the wind of the obstacles, and A_{ob} is the average area of the terrain occupied by each obstruction. Then, the empirical exponent α, gradient height z_g, and exposure coefficient K_z, corresponding to equations C26.10-3, C26.10-4, and C26.10-1.si of Engineers (2017), are used to calculate the correction factor $F_{exposition}$, verifying that z_0 units are in meters.

$$\alpha = 5.65 * z_0^{-0.133}$$

$$z_g = 450 * z_0^{0.125}$$

$$K_z = 2.01 * \left(\frac{z}{z_g}\right)$$

$$F_{exposition} = \frac{0.951434}{K_z}$$

Following NIST (2012), calculation of roughness need to be weighted according to the predominance of wind magnitude in eight directions (north, south, east, west, north-east, north-west, south-east, and south-west), see Figure 4.4, using a detailed aerial photo or satellite image inside a radius of 800 meters around the station location, as shown in Figure 4.5, with south direction highlighted.

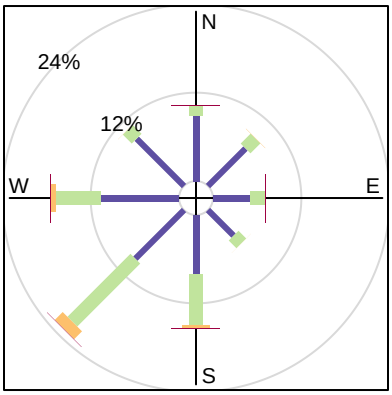

Figure 4.4: Wind rose with wind percentages in eight directions, for a generic station

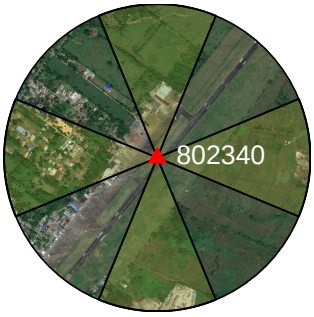

Figure 4.5: Digital imagery for 'Vanguardia' ISD station (USAF:802340), located in 'Villavicencio' airport. with four (south, north, east, and west) 45 degree sectors highlighted. Radius of the circular zone is 800 meters

Figure 4.6 shows extreme conditions for roughness, open space in left image (ISD Station 804070), closed space in center image (ISD Station 803000), and a typical example where Lettau procedure is needed. Lettau equation need to be applied to each direction and then the final z_o value is the weighted average, using historical wind percentage. See Figure 4.7 showing the strokes made to calculate the different areas for two Colombian stations. Information about wind percentage per direction at each station were obtained from IDEAM (1999).

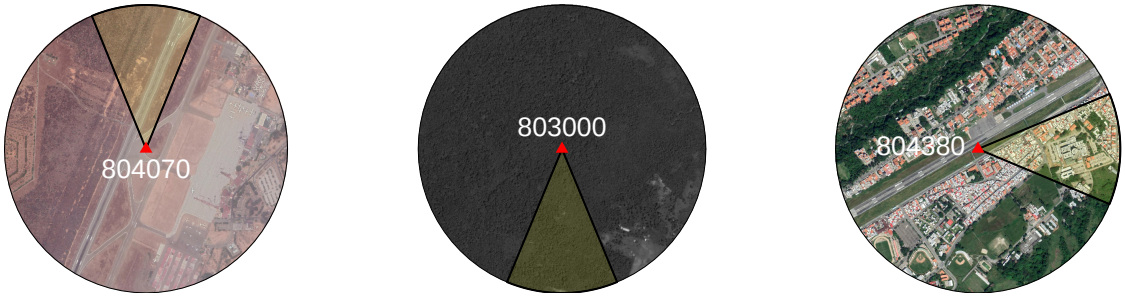

Figure 4.6: Roughness values: 0.03 for open space (left), 0.1 for closed space (center), and areas where Lettau equation is needed because roughness is different in each direction (right).

Figure 4.7: Lettau calculation. In red the area occupied by the obstacles, and in blue the perpendicular area. Source Triana (2019)

4.1.3 Averaging Time 3-s Gust

To transform hourly mean wind velocity V_{3600}, to 3-s gust velocity V_3, Engineers (2017) recommends to use C. S. Durst (1960). See Wind Loads Requirements. As the axis x represents duration t of the gust, what is done is to look there for the value 3 seconds, and read the corresponding gust factor $\frac{V_t}{V_{3600}}$, this is, the value in the axis y, then

$$V_t = V_{3\,seconds} = (gustfactor) * V_{3600\,seconds}$$

It is valid only for open terrain conditions. Durst curve shows in axis y the gust factor $\frac{V_t}{V_{3600}}$, a ration between any wind gust averaged at t seconds, V_t, and the hourly averaged wind speed V_{3600}, and in the axis x the duration t of the gust in seconds.

4.2 Downscaling Support

As it happens in this study, where it is intended to complement the local/regional wind analysis, with data from ISD (output data of a model for extreme winds), and ERA5 reanalysis dataset (large scale forecast data), it is required to probe by means of *comparisons* (exploratory data analysis and/or statistical measures) that those sources (modeled and forecast) are similar to IDEAM field measurements.

The proposed mechanism in the search for downscaling support is, a) the creation of *common time series graphs*, where time series overlays for all data sources, are expected to be similar, and b) the elaboration of *scatter plots graphics*, which are generated matching two sources by time (sorted in ascending order by wind velocity), and that, visually will allow to evaluate about data similarity between two sources, when all the points fall very close to a 45 degree line. In both cases, the strategy for station matching, could be one of the following:

1. *Manual matching*, doing a detailed analysis station by station (only for ISD and IDEAM). While it is true that ISD is based on IDEAM, their names and locations are somewhat different, for this reason, it is necessary to read information available from each source, and decide station by station, about its correspondence.
2. *Intersection matching*, between ISD and IDEAM point stations and ERA5 cells. All ISD and IDEAM stations falling inside a ERA5 cell, will be compared between them.

4.3 Peaks Over Threshold - Poisson Process (POT-PP)

Similar to how the adjustment of statistical data to a normal distribution works in order to make inferences considering deviations from the mean, here only some part of the data (those that are extreme - over a high threshold - POT), need to be fitted to a PP considering extreme deviations from the mean. While in the first case (normal distribution) the inferences are for events similar to the samples, in this case, when working with extreme value theory, the

inferences will be for more extreme events than any previously observed or measured. In the theoretical framework section are described the main elements of POT-PP.

In summary, POT means only to work with extreme values, and PP means to adjust data to a *pdf*, which depends on an intensity function $\lambda(t, y)$, where t is time, y is wind extreme velocity. As is shown in Figure 3.8, in a POT-PP approach with domain D, all the observations follow a Poisson distribution with mean $\int_D \lambda(t, y)\, dt\, dy$. Main advantage of POT-PP is that it is designed to consider storm and not-storm events independently (for each disjoint sub-domain D_1 or D_2 inside D, the observations in D_1 or D_2 are independent random variables), but in the end use them both for the inferences,

$$\mathrm{pdf} = \mathrm{f(t, y|\eta)} = \frac{\lambda(\mathrm{t, y})}{\int_{\mathrm{D}} \lambda(\mathrm{t, y})\, \mathrm{dt\, dy}} \tag{4.1}$$

4.3.1 Declustering

To make the assumptions of PP more justifiable, it is important to have only one sample per event, the highest one. For instance, if a hypothetical storm started at 11:30 in the morning and ended at 12:30 in the afternoon, and the time series for that event has thirty wind measurements (one each two minutes), it is necessary to leave only the stronger or maximum value, and this process is called de-clustering (see Figure 4.8). POT-PP defines that all the adjacent observations separated by six hours (6) or less in the case of thunderstorm events, and four (4) days or less, in the case of non-thunderstorm events belong to the same cluster.

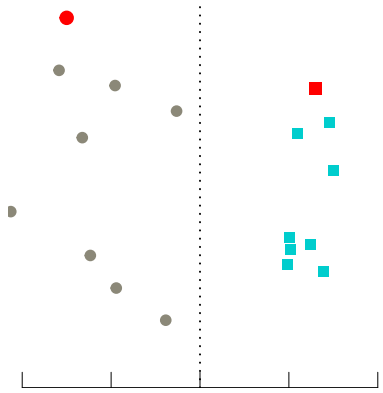

Figure 4.8: De-clustering in PP. Two thunderstorm clusters are shown. Separation between adjacent observations inside the clusters are always equal or less than six hours. Distance between the last event in the first cluster and the first event in the second cluster is larger than six hours. Only red samples are used to fit the PP, but in addition a POT (thresholding) process is needed

4.3.2 Thresholding

As the POT model requires to work only with the most extreme values in the time series, it is necessary to select a threshold to filter out small values. Selection of threshold value imply two effects in the model. Bias is high when a low threshold is selected (many exceedances) because the asymptotic support is weak. Opposite situation happens for high thresholds where variance is potentially high, so according to Davison & Smith (1990), it is needed to select a threshold value, consistent with model structure.

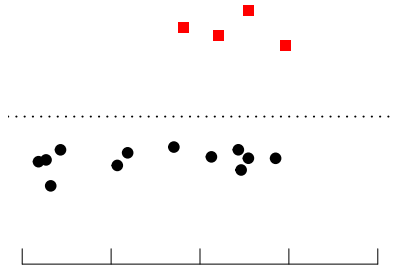

Figure 4.9: POT - Thresholding

Selection of the thresholds pairs, one for thunderstorm, and one for non-thunderstorm, is based in W transformation described in threshold selection section. W-statistic is done comparing the ordered empirical result of applying $W = -log(1-U)$ to the data, axis y in Figure 4.10, with the theoretical quantiles of an exponential variable with uniform distribution between 0 and 1, axis x in same figure. W-statistic is the highest vertical distance between the 45º line and the points in the graphic. The best thresholds pairs returns the minimum value for W-statistics after testing, in an iterative process, with many threshold pairs combinations.

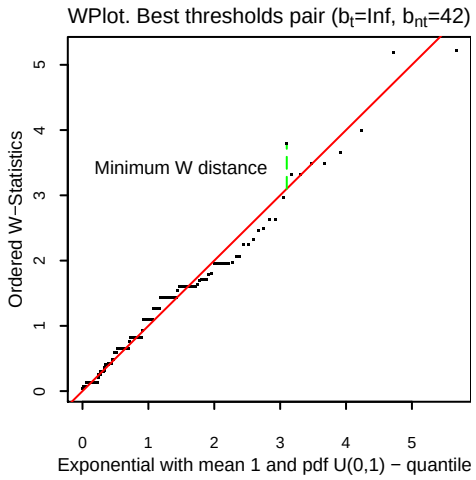

Figure 4.10: POT - Thresholding

4.3.3 Exclude No-Data Periods

PP requires to remove long periods of time when stations were not recording or failing. Proposed time in Pintar et al. (2015) is 180 days, namely, to remove all the gaps from the time series larger than six months.

4.3.4 Fit Intensity Function

Probability density function *pdf*, and cumulative distribution function *cdf*, of the PP, depend of the intensity function, and are shown in Equation (4.1), and Equation (3.9), respectively.

To facilitate the estimation of the parameters for the PP intensity function, parameter $shape = \zeta_t$ is taken to be zero in Equation (3.6), then doing the limit, the resulting intensity function is the same as the the GEV type I or Gumbel distribution,

$$\frac{1}{\psi_{\mathrm{t}}} \exp\left\{\frac{-(\mathrm{y}-\omega_{\mathrm{t}})}{\psi_{\mathrm{t}}}\right\} \tag{4.2}$$

In this study, used intensity functions are shown in next Equation (4.3).

$$\lambda\,(\mathrm{y},\mathrm{t}) \begin{cases} \dfrac{1}{\psi_s} \exp\left(\dfrac{-(y-\omega_s)}{\psi_s}\right), & for\ t\ in\ thunderstorm\ period \\ \dfrac{1}{\psi_{nt}} \exp\left(\dfrac{-(y-\omega_{nt})}{\psi_{nt}}\right), & for\ t\ in\ non--thunderstorm\ period \end{cases} \tag{4.3}$$

As is shown in 4.11, the fitting process involve finding the best group of parameters of the intensity function, in such a way that the red curve (*pdf* of the PP, based in intensity function) be as tight as possible to the shape of the data histogram. As is described in POT-PP, optimal parameters to do the fitting process of the intensity function are calculated using *maximum likelihood*.

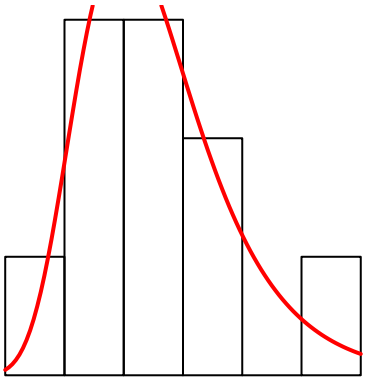

Figure 4.11: POT - PP intensity function fitting process

4.3.5 Hazard Curve - Return Levels - RL

If Equation (3.8), Y_N is solved using estimated parameters of the intensity function, and a hazard curve is constructed as shown in Figure 4.12, where axis x represents annual exceedance probability $P_e = \frac{1}{N}$, and axis y represents the RL Y_N for the corresponding N-years return period, then it will be possible to have the extreme return wind velocity level for any given return period going from axis x to axis y through the curve.

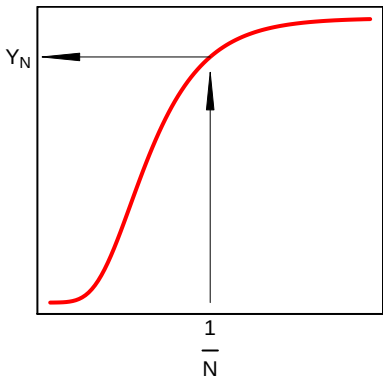

Figure 4.12: POT - PP fitting process

Two alternatives approaches for RL

There is an equation that allows direct calculations of return levels, and also it is possible to use the quantile function of Gumbel when shape parameter equals to zero, but it is important to emphasize that Equation (4.4), and the use of Gumbel quantile function for RL calculations, is only valid when the analysis of POT-PP includes only one type of event (thunderstorm or non-thunderstorm), and the average estimated duration time of the event in a year is considered to be one (independent of the units in which time is processed), namely, the values for parameters A_t or A_{nt} of Equation (3.8) are equal to one.

Instead of solving Equation (3.8), next Equation (4.4) can be used replacing directly the PP parameters and the N return periods to create the hazard curve and get RL.

$$\mathrm{Y_N} = \frac{\psi}{\zeta}\left[-\log\left(\frac{\mathrm{N}-1}{\mathrm{N}}\right)\right]^{-\zeta} - \frac{\psi}{\zeta} + \omega \tag{4.4}$$

As for this research $\zeta = 0$, return levels Y_N can be calculated using the Gumbel quantile function, but using $(1 - \frac{1}{N})$ as probability.

4.4 Spatial Interpolation

Probabilistic (Kriging) and deterministic (IDW, local polynomials) techniques are used to create maps for return levels with same return period. Interpolation with Kriging requires verification of minimum procedures to ensure proper use of the method, for instance,

- Structural analysis, which includes data normality check, for example, with Kolmogorov Smirnov or Shapiro Wilk goodness of fit tests, and if needed, data transformation to ensure data normality, e.g. using Box-Cox, and in addition, trend analysis to verify the need for trend modeling, in subsequent steps
- Semivariance Analysis: Use of available tools like cloud semivariogram, experimental semivariogram, directional semivariograms to verify isotropy or anisotropy, and different theoretical semivariograms, to ensure the best model of spatial autocorrelation, as a preliminary step to interpolation.
- Kriging Predictions: Use of different types of Kriging predictors, like simple, ordinary, universal, based on the results of the structural analysis.
- Cross Validation: Use of statistics like root mean square, average standard error, mean standardized, and root mean square standardized, that allow to measure the quality of the predictions and the magnitude of the errors.

Possible advantage of deterministic methods, is a better assessment of the local variability of spatial autocorrelation. It can also be considered with IDW or local polynomials a detailed assessment of structural analysis and cross validation. At the end of the spatial interpolation analysis all the predictions can be compared to select the most suitable result.

Main references in this research related to this matter, using **R software** are E. Pebesma & Graeler (2019), Pebesma (2004), and Gräler, Pebesma, & Heuvelink (2016). For the implementation of spatial statistics using vector or raster format, see E. Pebesma (2019a), E. Pebesma (2019b), and Pebesma (2018).

4.5 Integration with Hurricane Data

Engineers (2017) propose the equation C26.5-2 for combination of statistically independent events, of non-hurricane and hurricane wind speed data.

$$\mathrm{P_e(y > Y_N) = 1 - P_{NH}(y < Y_N) * P_H(y < Y_N)} \tag{4.5}$$

Where $P_e(y > Y_N)$ is the annual exceedance probability for the combined wind hazards, $P_{NH}(y < Y_N)$ is the annual non-exceedance probability for non-hurricane winds, and $P_H(y < Y_N)$ is the annual non-exceedance probability for hurricane winds.

To understand Equation (4.5), it is important to remember that to calculate return level Y_N, for a given N-year return period, the exceedance probability $\frac{1}{N}$ of Y_N is calculated. Then, the non-exceedance probability for Y_N is $\left(1 - \frac{1}{N}\right)$. The procedure consist in the creation of a new hazard curve, calculating all $P_e(y > Y_N)$ values for different Y_N return levels, combining hazard curves from non-thunderstorm and thunderstorm data.

Equation (4.5) can be expressed in terms of only exceedance probabilities, $P_e = 1 - (1 - P_{nh}) * (1 - P_h)$, where P_{nh} is the the annual exceedance probability for non-hurricane winds, and P_h is the annual exceedance probability for hurricane winds. A graphical explanation of the procedure to calculate the combined P_e for the return level $30\frac{Km}{h}$, is shown in next

Figure 4.13. For each cell in the study area, it is necessary to calculate a new combined hazard curve, this is, all the P_e values corresponding all different return levels (see right table in Figure 4.13).

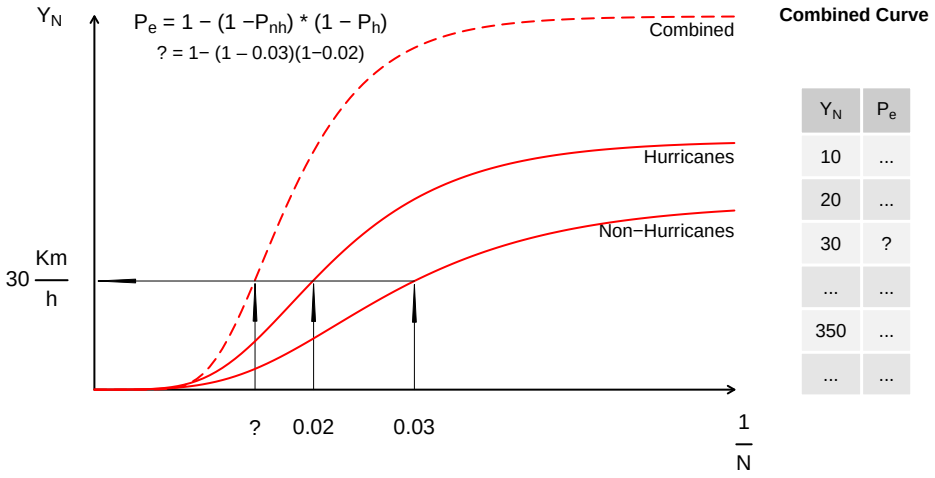

Figure 4.13: Integration Hurricane and Non-Hurricane Data

After the combined hazard curve is created, a new process of spatial interpolation need to be accomplished. In case of absence of hazard curves for stations, but availability of return levels maps, it becomes necessary to recreate hazard curves cell by cell, to apply Equation (4.5). In this case are required as many maps as possible for different return periods, in order to estimate detailed enough hazard curves from return level values (cell values).

Chapter 5

Results and Discussion

In this section, will be shown first, the data source comparison (post standardization process) to face the downscaling issue by using ERA5 and ISD database, then, the resulting process of fitting a POT-PP in the ISD station 801120, which includes revision of intensity function parameters, goodness of fit, hazard curve, return levels, and comparison with POT-GPD results, next, non-hurricane maps outputs, which includes results for ISD and ERA5 stations, after that, output maps combining hurricane and non-hurricane results will be displayed, and finally, a detailed discussion of the retults and future work is highlighted.

5.1 Data Standardization and Downscaling Support

Looking for a statistical justification in the use of ISD (model) and ERA5 databases (forecast), as input data for this study, and considering the *downscaling approach* described in the downscaling supportsection of methodology, data sources ISD and IDEAM were standardized to enable comparison. Standardization consisted of transforming the data to be equivalent to V_3 3-s gust, 10 meters anemometer height, and open space roughness. In the comparison process, for coincident stations by spatial location, it was checked if the velocity values (standardized) in the three sources, for equal dates, were similar in magnitude.

5.1.1 Data Standardization

None of the sources required anemometer height standardization. Lettau (1969) was used for roughness standardization of ISD and IDEAM, applying the method station by station. Gust velocities standardization was done using Durst curve, and in order to obtain V_3 from Durst curve, it was required to start from V_{3600} (average hourly speed), or from a different wind gust speed, for instance V_5 5-s gust.

For ERA5:

- Variable *10m wind gust - 10fg* of ERA5 data source does not need any standardization,

because it comes standardized from the source.

For ISD:

- Wind velocity from ISD comes from source as V_5, that is, five seconds gust wind velocity. To standardize from V_5 to V_3, using Durst curve, the correction factor is 1.03.
- Wind velocity V_5 from 74 ISD stations, was standardized station by station, using procedure described in Surface Roughness at Open Space section, and Averaging Time 3-s Gust section.

For IDEAM:

- As the original variables obtained from IDEAM, do not represent gust speeds, it was necessary to start from *average hourly speed* V_{3600}, to obtain 3-s gust V_3. To standardize from V_{3600} to V_3, using Durst curve, the correction factor is 1.51.
- It was not possible to obtain the *average hourly speed* V_{3600} from IDEAM, see Table 2.2, but from *instantaneous wind velocity each 2 minutes - VV_AUT_2* it is possible to obtain a **good** estimator of V_{3600}, and from *instantateous wind velocity each 10 minutes - VV_AUT_10* it is possible to obtain a **poor** estimator of V_{3600}.

5.1.2 Data Comparison

The available IDEAM data allowed two comparison processes, with quality data for few stations, and with low quality data, but available for all stations.

In both cases, to make the use of ISD and ERA5 viable, its time series are expected to be as similar as possible to IDEAM (field measurements). To verify this, two types of graphics were constructed:

1. **Time series overlay** for the three sources. Not very effective method due to the large amount of data that makes the graphics unreadable.
2. **Scatter plot graphics** comparing two different sources. Matching values by time, were sorted in ascending order, and put together on a scatter plot. The expected behavior in case of similarity in the data, is that all the points fall in a 45° line

IDEAM VV_AUT_2 - Quality Data Comparison

IDEAM VV_AUT_2 was available for twenty (20) stations, of which only twelve (12) were *perfectly equivalent* to ISD stations, see Table 5.1, and map in left panel of Figure 5.1. VV_AUT_2 dataset was transformed to V_{3600} (average hourly speed), averaging all 20 values available per hour. For twelve matching stations, wind velocity V_{3600} (transformation of VV_AUT_2), was standardized station by station, using procedure described in Surface Roughness at Open Space section and Averaging Time 3-s Gust, and finally, for the same twelve ISD and IDEAM standardized stations, a comparison was done against matching ERA5 stations (the corresponding cell in ERA5 that has within ISD and IDEAM locations).

Table 5.1: Quality Data Comparison

ISD ID	IDEAM ID	ERA5 ID, (col,row), [lon,lat]
803980	48015050	3320, (37, 68), [-70, -4.25]
803700	52055230	2309, (6, 48), [-77.75, 0.75]
802110	26125061	1582, (14, 33), [-75.75, 4.5]
802100	26125710	1533, (14, 32), [-75.75, 4.75
801120	23085270	1240, (15, 26), [-75.5, 6.25]
801100	27015330	1240, (15, 26), [-75.5, 6.25]
800970	16015501	909, (27, 19), [-72.5, 8]
800940	23195502	1102, (24, 23), [-73.25, 7]
800630	13035501	749, (14, 16), [-75.75, 8.75]
800360	28025502	416, (24, 9), [-73.25, 10.5]
800350	15065180	221, (25, 5), [-73, 11.5]
800280	29045190	312, (18, 7), [-74.75, 11]

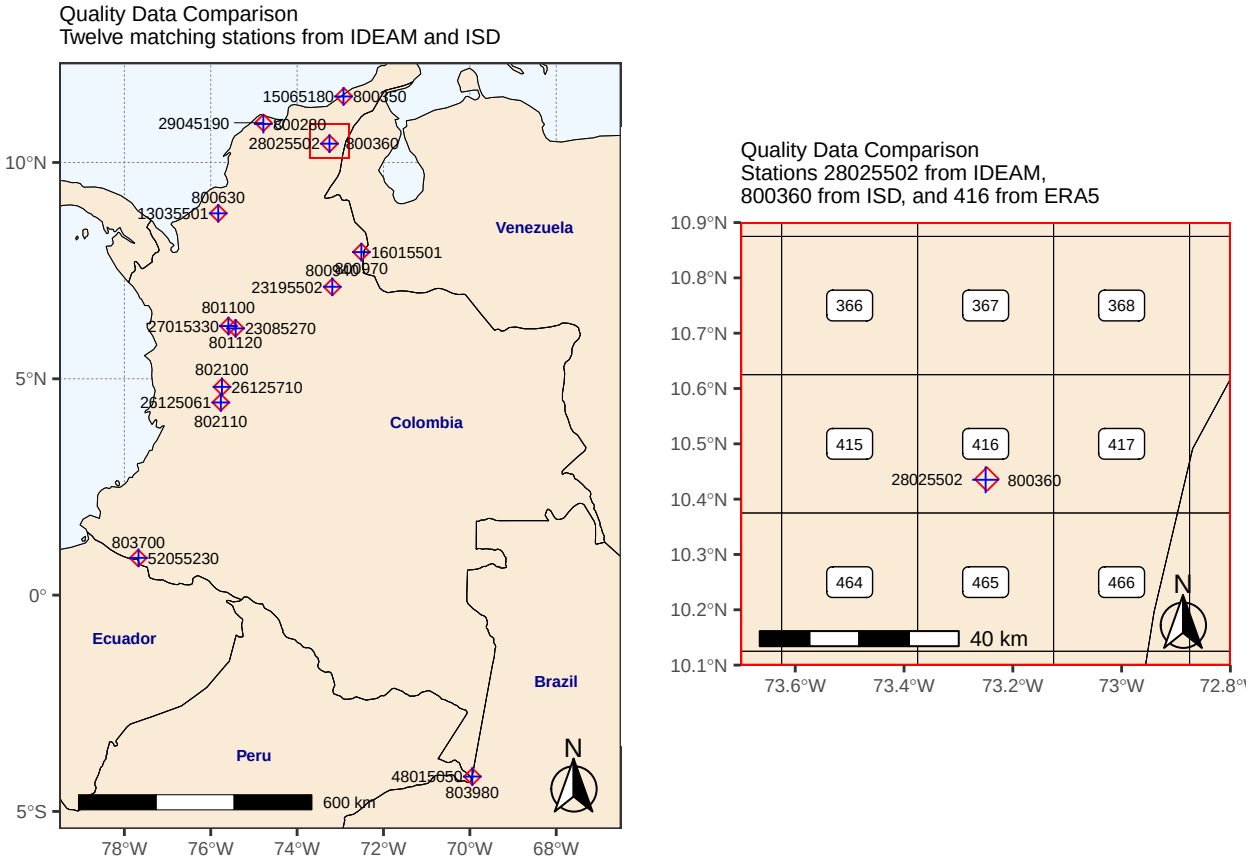

Figure 5.1: IDEAM VV_AUT_2 - Quality Data Comparison.

The stations described in each row of the previous Table 5.1, were compared by generating scatter plots and common time series graphics. Stations 28025502 from IDEAM, 800360 from ISD, and 416 (cell with center point in $-73.25°$ longitude, and $10.5°$ latitude) from ERA5,

see map in right panel of Figure 5.1, showed high correspondence, see Figure 5.2, because green regression line (empirical) is very similar to 45° line (theoretical). Unfortunately, in the other eleven stations, there was no high equivalence between sources.

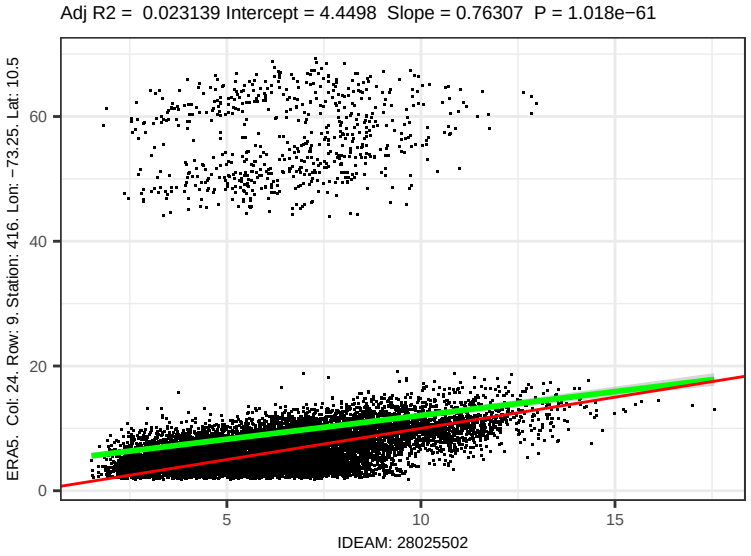

Figure 5.2: Quality Data Comparison. High similarity between sources

IDEAM VV_AUT_10 - Non Quality Data Comparison (available in all IDEAM stations)

VV_AUT_10 was available for 204 stations, and despite that V_{3600} calculated from this source, is not an accurate or quality estimator, the standardization procedure was done to allow an additional comparison process, whose results are shown in the map displayed in Figure 5.3. Downscaling support was 'Good' comparing IDEAM and/or ISD stations with twenty-three (23) ERA5 stations (2261, 1971, 2066, 2020, 2260, 1875, 2213, 2637, 1442, 1583, 1501, 1582, 1381, 1493, 1485, 1397, 1338, 1055, 511, 1644, 515, 221, 1038), and 'Very Good' comparing IDEAM and/or ISD with five (5) ERA5 stations (265, 360, 78, 312, 416).

'Very Good' downscaling results for this non quality data comparison, are shown bellow:

- Table 5.2, shows in each row compared stations.
- Figure 5.4, shows an example of a very good time series plot for the ERA5 station 78 vs IDEAM stations 15075501 and 15079010.
- Figure 5.5, shows four different very good scatter plots, a) IDEAM 15015120 vs ERA5 265, b) IDEAM 29004520 vs ERA5 312, c) IDEAM 15079010 vs ERA5 78, and d) IDEAM 15075501 vs ERA5 78. Red line in each graphic represent the desired 45° line, where all points should fall, if the data sources would be exactly the same (theoretical behavior when there is equivalence of sources), and green line represents the linear regression line (empirical or real behavior when making the comparison).

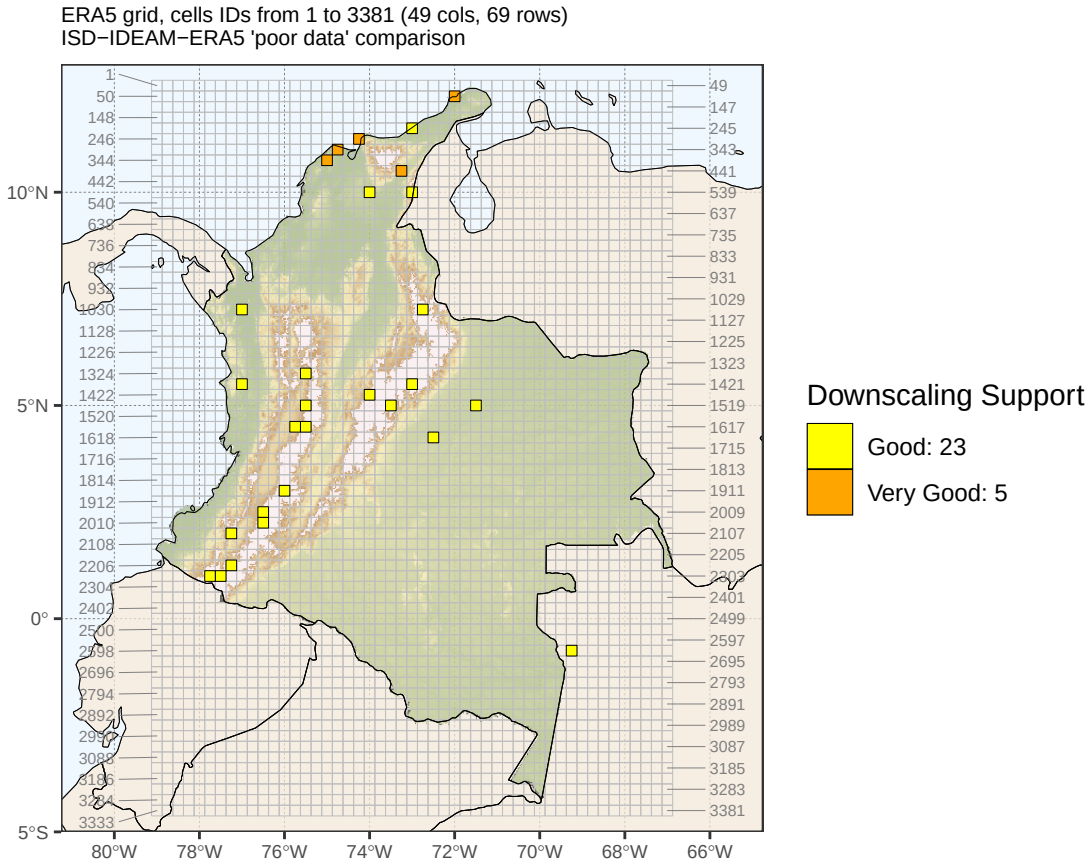

Figure 5.3: IDEAM VV_AUT_10 - Non Quality Data Comparison. Two different types of downscaling support: 'Good' and 'Very Good'

Table 5.2: Non quality data comparison. 'Very Good' downscaling support.

ISD ID	IDEAM ID	ERA5: ID, (col,row), [lon,lat]
NA	16015501	78, (29, 2), [-72, 12.25]
NA	15079010	78, (29, 2), [-72, 12.25]
NA	15075501	78, (29, 2), [-72, 12.25]
NA	15015120	265, (20, 6), [-74.25, 11.25]
NA	29004520	312, (18, 7), [-74.75, 11]
800280	29045190	312, (18, 7), [-74.75, 11]
NA	29045000	360, (17, 8), [-75, 10.75]
NA	28025502	416, (24, 9), [-73.25, 10.5]
800360	28035060	416, (24, 9), [-73.25, 10.5]

For some ISD stations in previous Table 5.2, the value 'NA' means that for the corresponding ERA5 and IDEAM station (same row), there is not a ISD station located inside the ERA5 cell space ($0.25° * 0.25°$).

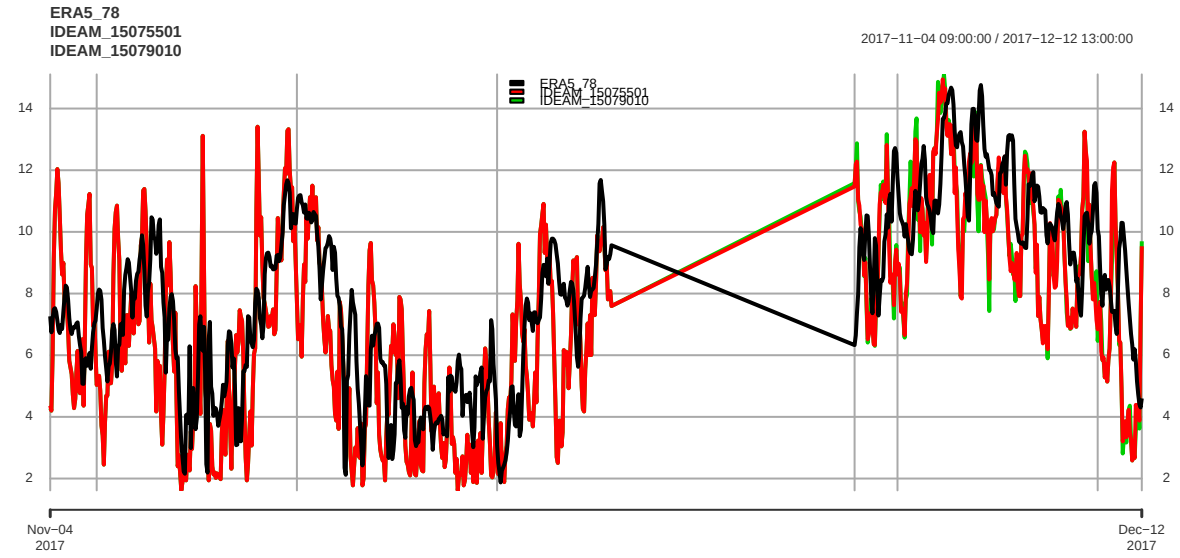

Figure 5.4: Non Quality Data Comparison. Time Series Graphic for 'Very Good' Downscaling Support

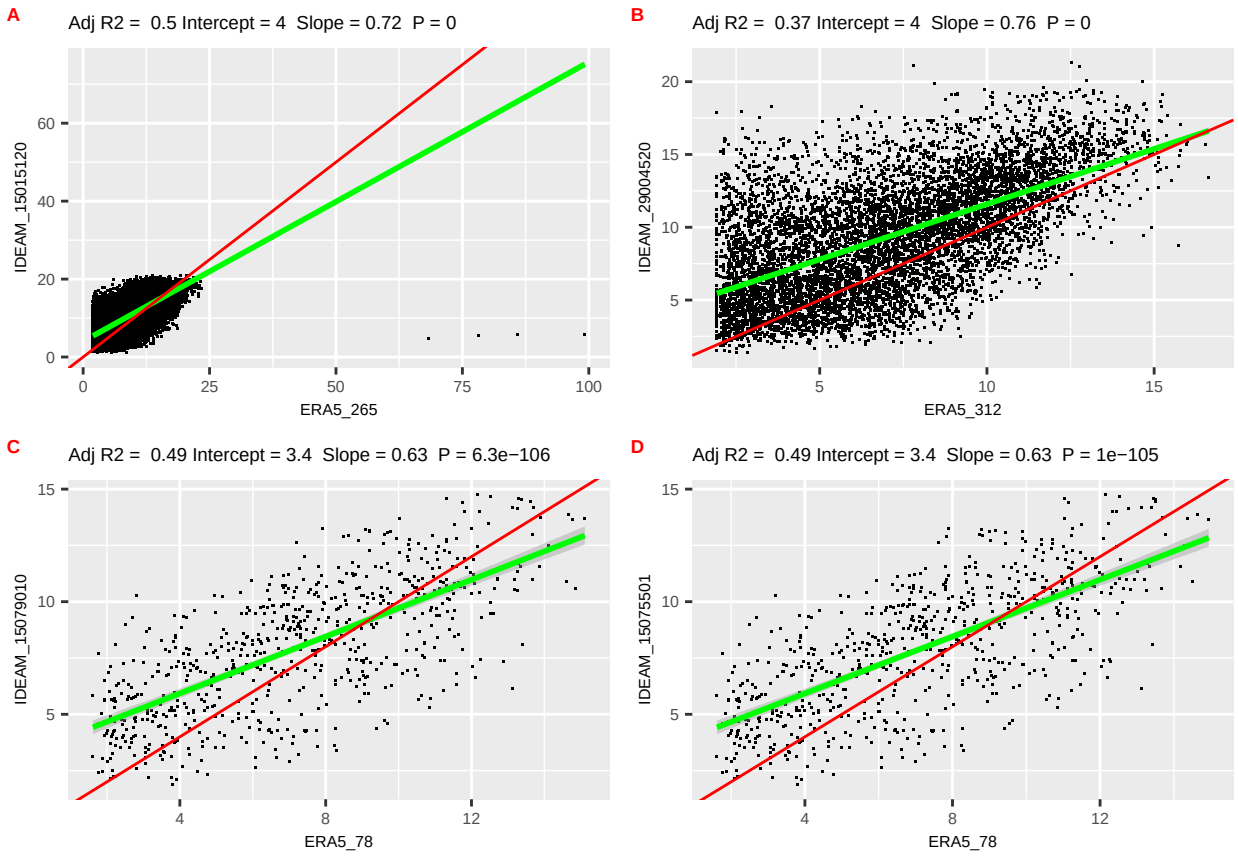

Figure 5.5: Non Quality Data Comparison: Scatter plots for 'Very Good' Downscaling Support

5.2 POT-PP for ISD Station 801120

Figure 5.6 shows the satellite image (source Google Earth) of ISD station 801120, located in the international airport 'José María Córdova', municipality of Rio Negro (Antioquia, Colombia), with latitude 6.125°, and longitude −75.423° WGS84 coordinates. Red circle represents an influence radius of 800 meters. Table 5.3 shows different calculations related to correction factors applied to this station, using procedure described in sections Surface Roughness at Open Space, and Averaging Time 3-s Gust.

Figure 5.6: Location of ISD station 801120

Table 5.3: Corrections factors for ISD station 801120

Variable	Value
Roughness - Z_o	0.05
Empirical exponent - α	8.38
Gradient height - z_g	310.56
Exposure coefficient - K_z	0.88
$F_{exposition}$	1.07
Gust factor for V_3	1.03

5.2.1 Raw Data, De-clustering, and Thresholding

As storm information is not available for any of the data sources, all the data for the station was classified as *non-thunderstorm*. According to POT-PP method described in methodology, the first process applied to original time series -raw data-, is declustering, and then, thresholding.

Non-thunderstorm raw data for ISD station 801120 has 2931 records, from 1986-12-06 12:00:00 to 2019-03-01 12:00:00, corresponding to a total amount time in days of 11739, and to an average number of events per year of 18.9, which means that the average duration of an event is 19.3 days (average size in days of a cluster). After declustering, and thresholding processes, the number of records decreases to 181. Time series graphics are shown in

Figure 5.7, showing the data before (left) and after (right) applying the mentioned processes. Detailed yearly statistics are reported in Table 5.4, also including summary for before (left), and after (right).

Table 5.4: Yearly statistics of raw data and declustered data for ISD station 801120

	Raw Data				Declustered Data			
Year	Count	Mean	Min	Max	Count	Mean	Min	Max
1986	63	45.2	27.9	163.3	7	106.4	43.8	163.3
1987	192	36.1	26.7	87.6	10	61.0	45.0	87.6
1988	234	43.8	26.7	90.4	23	64.2	45.0	90.4
1989	256	44.2	27.9	103.6	19	64.4	45.0	103.6
1990	250	44.9	26.7	103.6	21	67.2	45.0	103.6
1991	149	38.7	26.7	127.5	20	58.6	45.0	127.5
1992	126	35.2	26.3	81.7	9	52.6	43.8	81.7
1993	109	36.3	26.3	79.7	13	53.5	43.8	79.7
1994	124	36.8	26.7	79.7	12	56.1	45.0	79.7
1995	89	33.3	26.7	111.5	2	77.7	43.8	111.5
1996	70	35.6	26.7	87.6	6	65.7	43.8	87.6
1997	71	36.6	26.7	119.5	4	86.9	49.0	119.5
1998	65	33.8	27.9	61.4	2	54.6	47.8	61.4
1999	48	31.7	26.7	47.8	1	47.8	47.8	47.8
2000	69	33.4	26.7	87.6	3	68.3	55.8	87.6
2001	62	29.9	26.7	39.8	0	NA	NA	NA
2002	94	33.3	26.7	71.7	5	54.2	43.8	71.7
2003	78	31.5	26.7	71.7	1	71.7	71.7	71.7
2004	60	31.9	26.7	51.8	2	48.4	45.0	51.8
2005	59	33.3	26.7	94.4	2	69.1	43.8	94.4
2006	55	32.6	26.7	164.1	1	164.1	164.1	164.1
2007	25	29.8	26.7	39.0	0	NA	NA	NA
2008	13	36.1	26.7	96.4	1	96.4	96.4	96.4
2009	36	31.6	26.7	82.1	1	82.1	82.1	82.1
2010	31	43.0	27.9	119.5	8	83.0	61.4	119.5
2011	32	29.2	26.7	41.0	0	NA	NA	NA
2012	82	31.9	26.7	87.6	4	64.5	43.0	87.6
2013	91	29.7	26.7	37.0	0	NA	NA	NA
2014	95	30.1	26.7	47.8	1	47.8	47.8	47.8
2015	129	30.3	26.7	51.8	1	51.8	51.8	51.8
2016	33	30.7	26.7	87.6	1	87.6	87.6	87.6
2017	18	31.3	26.7	67.7	1	67.7	67.7	67.7
2018	22	31.0	26.7	39.8	0	NA	NA	NA
2019	1	28.7	28.7	28.7	0	NA	NA	NA

It can be seen in the Table 5.4, that de-clustered data has zero records for some years. This situation is due to that all the data for each one of those years (2001, 2007, 2011, 2013, 2018, and 2019), belonged to a cluster that started the previous year, or finished the next year, and the unique chosen maximum value (the value representative for the cluster) was found in the previous or next year, but not in mentioned years of zero records.

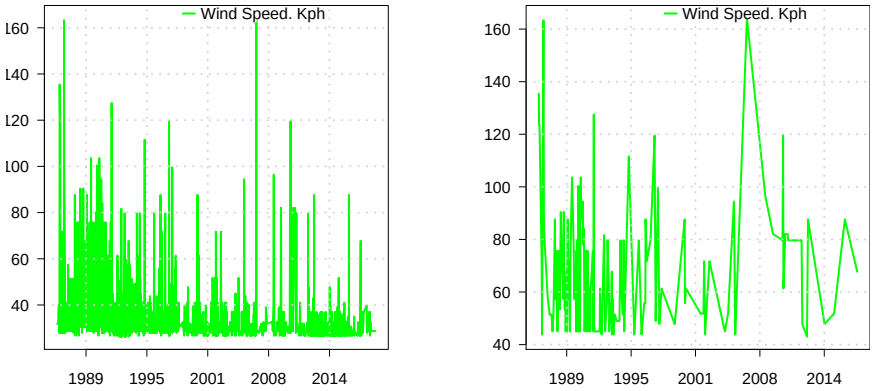

Figure 5.7: Non-Thunderstorm Time Series for ISD station 801120. Left: Raw Data. Right: De-clustered Data

Using de-clustered data, and considering that it is only necessary to calculate optimal threshold for non-thunderstorm data, because there is no records classified as thunderstorm in any data source, many non-thunderstorm thresholds were tested, to choose the best one using the W statistic, as described in section thresholding of the methodology. Figure 5.8 shows a very good fit in resulting W-Statistic plot, for optimal non-thunderstorm threshold $b_{nt} = 42\frac{Km}{h}$, with a minimum W distance of 0.47, for ISD station 801120, where empirical values (black points) are very close or similar to theoretical values (red line).

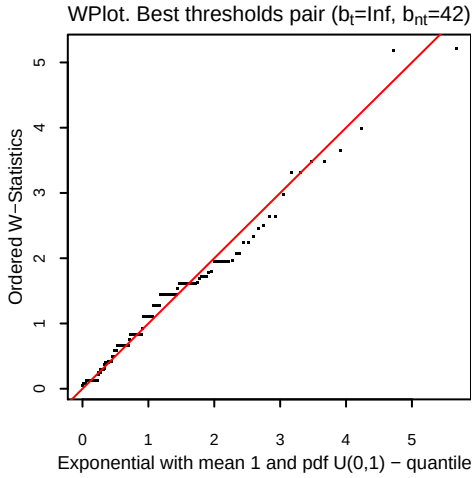

Figure 5.8: POT - Thresholding

5.2.2 Fitted *pdf* and *cdf*, and Goodness of Fit

Equation (3.6), defined in section POT-PP of the methodological framework, was used as intensity function, $\lambda(t, y) = \lambda_{nt}(y)$, for POT-PP. When shape ζ_t is equal to zero, as it is in this study, an equivalent intensity function is described in Equation (4.3) defined in terms of the parameters location (ω_t), and scale (ψ_t). Related *pdf* and *cdf* functions are referenced

in Equations (4.1), where the domain D constraint the data above the threshold b, and the time to a non-thunderstorm period, and (3.9) respectively.

- Intensity function: $\frac{1}{\psi_{nt}} \exp\left(\frac{-(y-\omega_{nt})}{\psi_{nt}}\right)$
- pdf: $f(t,y) = \frac{\lambda(t,y)}{\int_D \lambda(t,y)\,dt\,dy}$
- cdf: $F(t,y) = P(y \leq Y) = \frac{\int_b^Y \lambda(y,t)\,dy}{\int_b^\infty \lambda(y,t)\,dy}$

After fitting the intensity function to the domain D, the resulting parameters for ISD station 801120, are location ω_t equal to -55.62, and scale ψ_t equal to 23.4. Figure 5.9 shows the histogram and fitted *pdf* in panel A, Q-Q plot (theoretical quantiles against empirical ones) in panel B, empirical cumulative distribution against fitted *cdf* in panel C, and P-P plot (theoretical probabilities against empirical ones) in panel D. In all four panels, it can be seen that there is a very good visual correspondence between empirical data (points and histogram) and theoretical adjustment (lines).

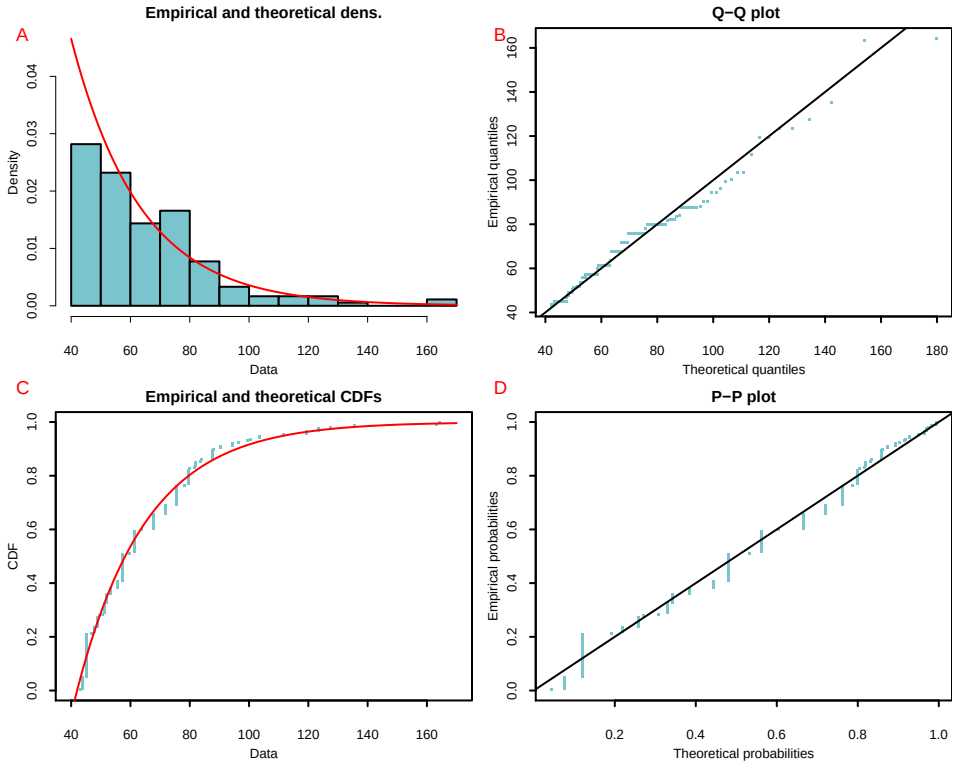

Figure 5.9: Graphic Diagnosis Of Goodness of Fit. Station 801120

Results of formal goodness of fit statistics for 'Kolmogorov-Smirnov D', 'Cramer-von Mises T' and 'Anderson-Darling A' are 0.089, 0.21, and 1.68 respectively. For a proposed null hypothesis, which indicates that the data conforms to a POT-PP, all resulting p-values using statistics D, T and A, confirm that there is no statistical evidence to to reject stated hypothesis. Resulting p-value for statistic D is 0.11. Another available criteria to measure

the quality of the fitted process are 'Akaike's Information Criterion', and 'Bayesian Information Criterion', with values 1505.2, and 1508.4 respectively. The Root Mean Square Error (RMSE), calculated using theoretical versus empirical *cdf*, is 0.023.

5.2.3 Hazard Curve and Return Levels - RL

Hazard curve is the solution to Equation (3.8), but eliminating from it elements related to thunderstorms, the equation is simplified to $A_{nt} \int_{Y_N}^{\infty} \lambda_{nt}(y)\, dy = \frac{1}{N}$, where A_{nt} is the average time of non-thunderstorm events by year, and Y_N is the return level or extreme wind velocity, corresponding to the N-years return period or MRI. Replacing in this equation the intensity function $lambda_{nt}$, and solving Y_N for all possible values of MRI, will provide hazard curve displayed in Figure 5.10.

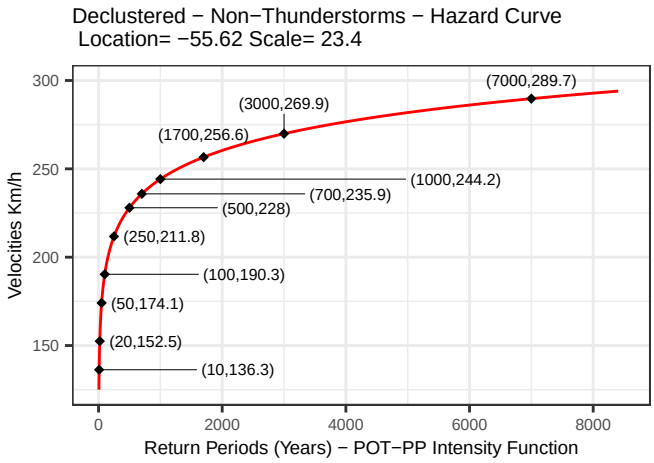

Figure 5.10: Hazard Curve. Station 801120

Table 5.5: Return Levels -RL for typical Mean Return Intervals - MRI. ISD station 801120

MRI	Return Level
10	136.30
20	152.48
50	174.10
100	190.32
250	211.76
500	227.98
700	235.85
1000	244.20
1700	256.61
3000	269.90
7000	289.73

Return levels of interest for this research, correspond to 700, 1700 and 3000 years of MRI, however, due to the mechanism of integration with existing hurricane study information, described in methodology, it is necessary to extract for all stations values related to typical return periods, as shown in the Table 5.5.

5.2.4 Comparison with POT-GPD and Common Extreme Value Distributions

To enable a comparison between, a) POT-PP (previous section), b) POT-GPD, and c) the fitting process of common extreme value distributions (GPA, GEV, GUM) without using POT method, this is, using the generic concept of hazard function *hf* (see theoretical framework), a whole automation process was done to calculate return levels and errors using mentioned alternatives, bearing in mind that in all cases *maximum likelihood* was used to calculate the parameters.

Resulting return levels and errors for POT-GPD, using R packages extRemes - Gilleland (2019), ismev - Janet E. Heffernan with R port & Alec G. Stephenson. (2018), evd - Stephenson (2002), Renext - Deville & IRSN (2016), evir - Pfaff & McNeil (2018), and fExtremes - Wuertz, Setz, & Chalabi (2017), are reported in Table 5.6. Similarly are shown it Table 5.7, return levels calculated from the adjustment of the probability distributions GPA, GEV, and Gumbel.

Table 5.6: POT-GPD. Return Levels in Kph

PACKAGE	RETURN LEVELS FOR TYPICAL MRIs											ERROR
	10	20	50	100	250	500	700	1000	1700	3000	7000	RMSE
extRemes	155.6	169.3	187.2	200.4	217.6	230.3	236.4	242.8	252.2	262.1	276.6	0.057
ismev	155.5	169.3	187.1	200.4	217.5	230.1	236.2	242.6	252.0	261.9	276.4	0.057
evd	155.6	169.3	187.2	200.4	217.6	230.3	236.4	242.7	252.2	262.1	276.6	0.057
Renext Renouv	155.6	169.3	187.2	200.4	217.6	230.3	236.4	242.7	252.2	262.1	276.6	0.057
evir	155.0	168.5	185.8	198.6	215.1	227.3	233.1	239.2	248.2	257.6	271.3	0.058
fExtremes	155.5	169.3	187.2	200.4	217.5	230.2	236.3	242.6	252.0	261.9	276.5	0.057
Renext 2 parameters	200.8	203.9	206.5	207.8	208.9	209.4	209.6	209.7	209.9	210.1	210.3	0.337

Table 5.7: Common Extreme Value Distributions. Return Levels in Kph

EVD		RETURN LEVELS FOR TYPICAL MRIs											ERROR
NAME		10	20	50	100	250	500	700	1000	1700	3000	7000	RMSE
gpa	Generalized Pareto	149.6	160.6	174.2	183.9	195.8	204.2	208.2	212.2	218.0	223.9	232.2	0.048
gev	Generalized Extreme Value	172.5	198.8	239.2	274.8	329.5	377.8	403.5	432.7	479.9	536.0	631.7	0.058
gum	Gumbel	140.9	152.1	167.0	178.2	193.0	204.3	209.7	215.5	224.1	233.3	247.0	0.067

5.3 Wind Maps

Using calculated return levels in all available ISD stations, continuous maps covering the study area, were created using spatial interpolation techniques as described in methodology. As calculated return levels for ERA5, represent predefined square cells of 0.25° decimal degrees, no interpolation process was necessary in this reanalysis dataset.

5.3.1 Existing Hurricane Maps

The Colombian consulting firm Ingeniar ltda, following the methodology described in CIMNE (2015), and CIMNE (2017), has provided raster wind maps for return periods 10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, and 7000 years, resulting from the probabilistic study of winds due to hurricanes in the Colombian Caribbean Sea and the surrounding continental zone. Figure 5.11 shows three of mentioned maps.

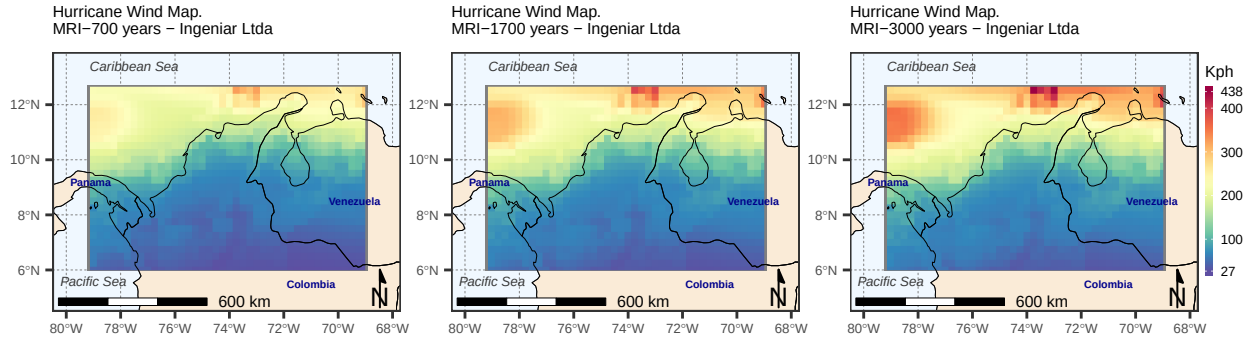

Figure 5.11: Ingeniar Hurricane Wind Maps.

5.3.2 Non-Hurricane Maps

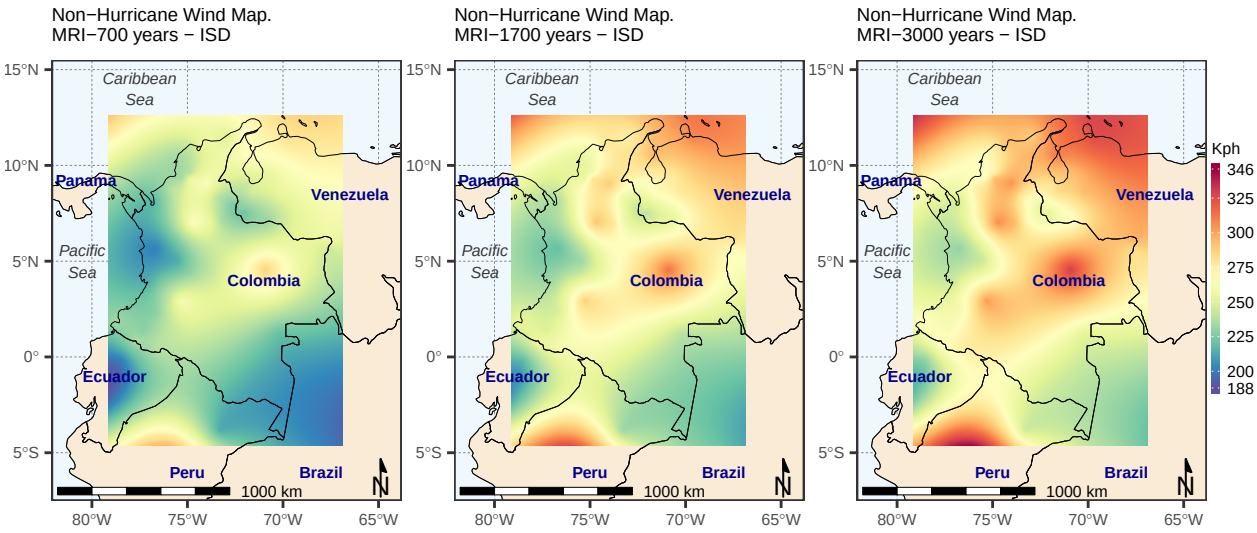

Figure 5.12: ISD Non-Hurricane Wind Maps.

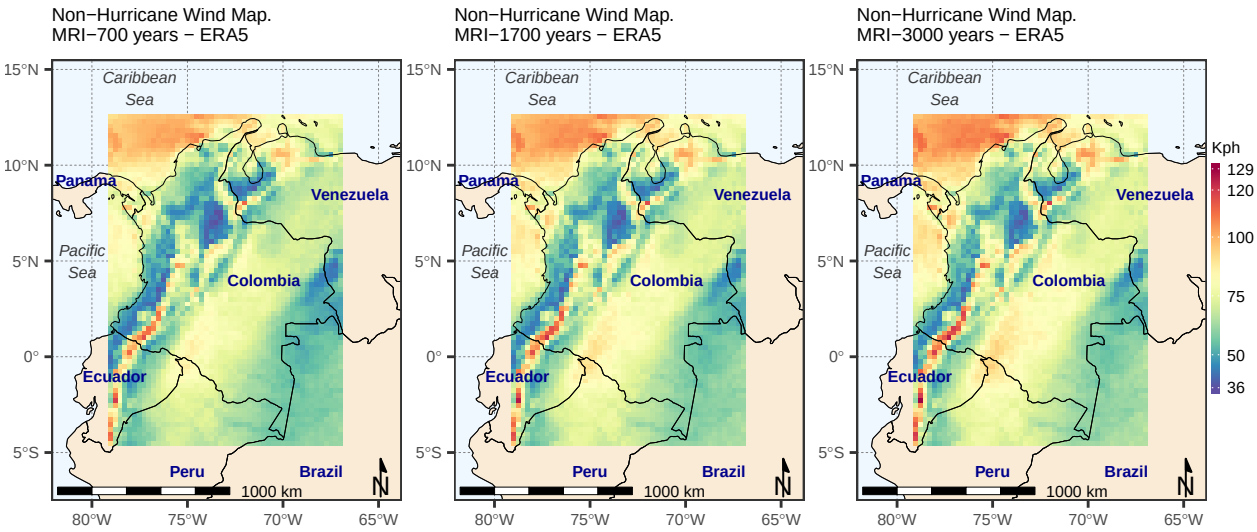

Figure 5.13: ERA5 Non-Hurricane Wind Maps.

5.3.3 Combined Maps

Following the procedure described in integration with Hurricane data, final wind maps are created, combining existing data for hurricane studies, and non-hurricane maps produced in this study.

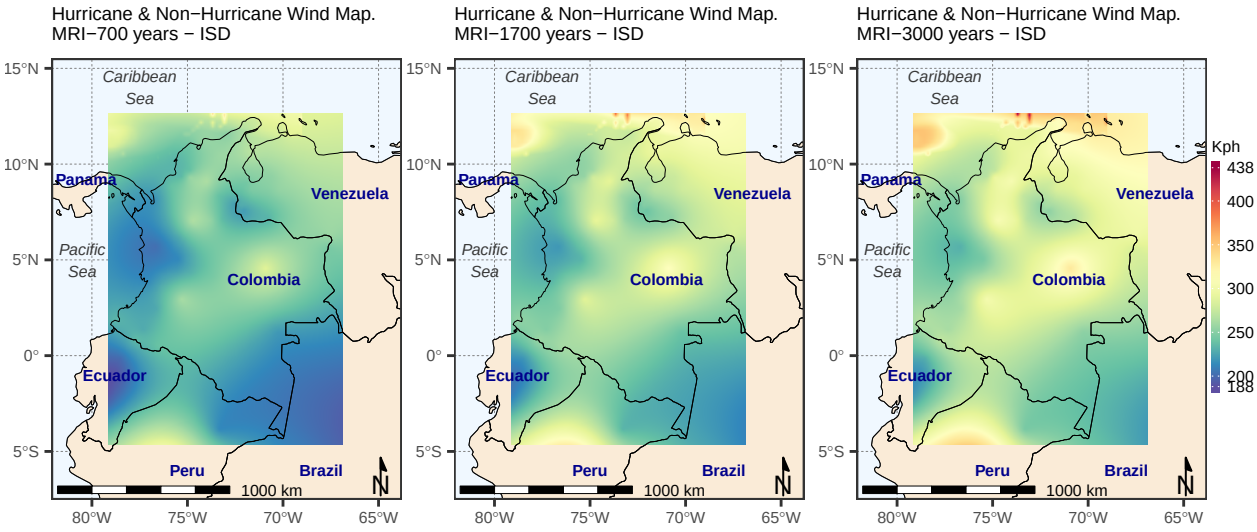

Figure 5.14: ISD Hurricane & Non-Hurricane Wind Maps.

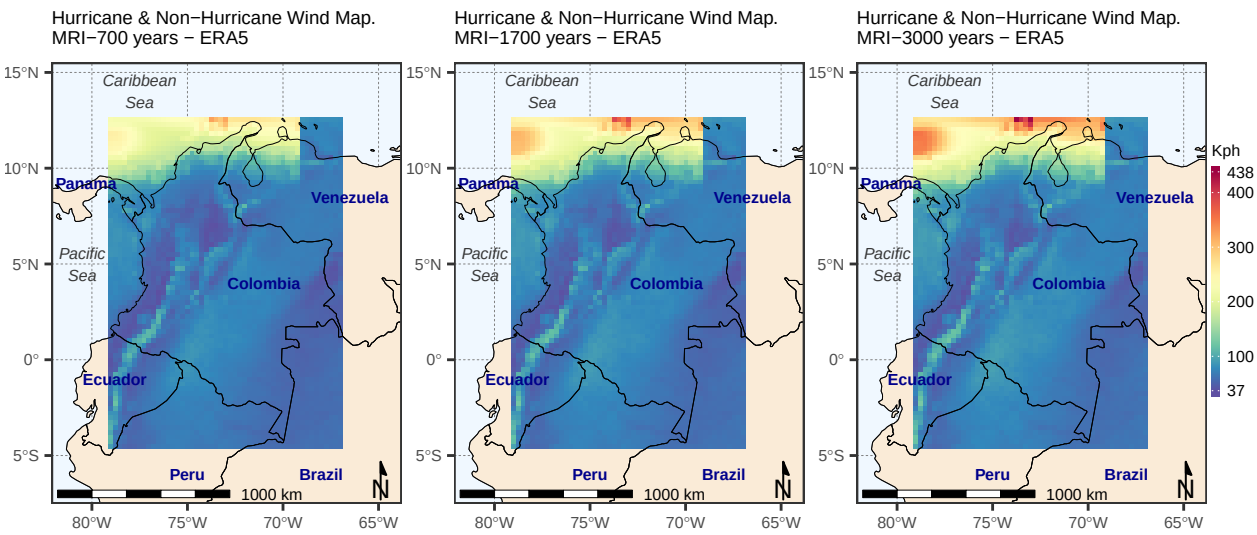

Figure 5.15: ERA5 Hurricane & Non-Hurricane Wind Maps.

5.4 Final Discussion and Future Work

Regarding the comparison of the data, it must be remembered that the basis of comparison, that is, the one that represents the truth in the field - IDEAM field measurements, was not fully available, what disturbed the process since before starting it. On the other hand, there are many uncertainties with respect to the model that represents the ISD database, because first, the available documentation does not specify whether it is an average or a gust data, second, the comparative graphs showed that ISD database did not represent a continuous variable (vertical or horizontal stripes in scatter plots), and finally, the comparisons against IDEAM never showed good results.

With respect to ERA5 database, although the comparative results showed greater similarity, it should be remembered that each record in the time series does not represent a point value, on the contrary, it represents a square cell of 0.25 decimal degrees. The IDEAM stations with which the comparison was made, can fall into any location of the cell, and constitute only a very local condition, that is not represented by an averaged forecast for the whole cell size, this considering that Colombia is a tropical region with a widely diverse territory (mountains, plains, rivers, forests, etc) and climate. So the possible similarity between IDEAM and ERA5 is limited by this condition.

The main difference between POT-PP and POT-GPD is that in the latter, wind quantities are adjusted to a GPD, and the time is adjusted to a Poisson Process (1D), while in the former, time and magnitude are adjusted to a Poisson Process (2D). If there is no weather classification available (storm and no storm) in the wind time series, POT-PP loses its advantages and resembles in potential and scope to POT-GPD, because the intensity function varying only in magnitude, becomes similar to a GPD. For this reason, POT-PP method is really useful, if historical classification of time series is available, record by record, in storm and non-storm.

This classification by time of historical series, is useful because it allows to define more precisely the average rate of events per year (Poisson process rate), which in POT-PP is represented by the average amount of events time per year, this is, components A_t and A_{nt} of Equation (3.8) - $A_t \int_{Y_N}^{\infty} \lambda_t(y)\, dy + A_{nt} \int_{Y_N}^{\infty} \lambda_{nt}(y)\, dy = \frac{1}{N}$ -, used to calculate return levels Y_N.

By the lack of thunderstorm and non-thunderstorm information, is impossible to calculate which part of the annual time belongs to storm and which to not-storm. As the available data were all assumed as non-thunderstorm data, this average time of events per year will always result in a fixed wrong value of 365 days, the maximum possible value. For ISD, this condition is reflected in high and unlikely final results.

However, the same condition did not affect the results in the ERA5 database in the same way. Although also in ERA5, all the data were classified as "non-thunderstorm", and the average time of events was always 365 days, an additional condition made the final result more realistic. Contrary to what happens in the ISD database, where time series have many gaps and there is lack of information, used ERA5 database has the full time series, hour by hour, from 1979 to 2019. Following the theory behind POT-PP, this implies that there is only one cluster for the whole time series, which would leave a single data after the de-clustering, canceling the entire subsequent process. Our proposed solution was to work only with one data per week, the maximum, which implies that the de-clustering process will have no effect, since it is based on 4-day gaps, see Declustering in Methodology section, resulting in more events above the threshold, exactly 48 events for each one of the 40 years of history, which translates into greater averaging of the final wind data.

One of the objectives of the investigation was to compare different methods in the calculation of return levels, and that was achieved using in all cases, both POT-GPD, and POT-PP. The Poisson for POT-GPD, does not accept data classification by time (storm or non-storm), because it is one-dimensional and data must represent a single general type of event: wind. Despite that, it is important to emphasize that the shortcomings in the calculation of the Poisson process rate, similarly affected the application of both methods, so in all cases, the results have the same limitation, and the use of POT-GPD does not represent best quality in the results.

To improve the quality of extreme wind analysis in future research, the inclusion of seasonal effects is recommended, this can be done in two ways, first, using a separate POT-PP for each season, and second, model the Poisson process parameters (location, scale and shape) as sinusoidal functions of time. Finally, it is possible to include more formal statistics (not only graphics), to face the downscaling challenge.

Chapter 6

Conclusions

Final maps using ERA5 forecast database, see Combined Maps in Results and Discussion section, representing return periods of 700, 1700, and 3000 years, are the extreme velocities needed as input load for the design of structures of different use category in the study area. Nevertheless, by one hand, full data from the IDEAM source is needed to enable the validation of downscaling support, on the other hand, it is essential to include in the study the classification of thunderstorm and non-thunderstorm data to achieve more realistic results, and finally, an additional conservative calibration process is needed, where to each municipality is assigned only a wind velocity, in order to define final values that will be part of the structure design norm.

Other general conclusions of the investigation are:

- In the abscense of wind field measurements, alternatives data sources as ISD and ERA5 can be a viable source of data to calculate extreme wind events, but always must be searched for statistic or graphic support for the downscaling issue, and at the end a process of calibration is needed for each particular case.
- A powerful R tool is available to do extreme value analysis using POT-PP and POT-GPD approaches
- Results of this research, could be the starting point of a process to update wind maps in countries with information issues.
- Output results of this research will contribute to reduce the risk of infrastructure collapse, representing a favorable impact in human lives, material losses, and disaster prevention.

For a detailed analysis of the results, refer to Results and Discussion section, and for a discussion about the project and its relevant topics, refer to Final Discussion

Appendix A

Research R Code - Digital Files

Table A.1: Research R Code. ftp://ftp.geocorp.co/windthesis/. User anonymous@geocorp.co (no password).

Folder Tree - Ftp Links	Description
code	Folder with R code. ALL CODE CREATED BY DR. ADAM PINTAR IS NOT PUBLISHED.
|-pot_pp	Folder with POT-PP R code. Based in Dr Adam Pintar code (respected copyright).
| |-function_lib.r	POT-PP Functions. Author of declustering and thresholding functions is Dr Adam Pintar.
| |-plot_nt.r	Plot non-thunderstorm graphics.
| |-plot_t.r	Plot thunderstorm graphics.
| |-plot_t_nt.r	Plot graphics with thunderstorm and non-thunderstorm data, in simultaneous.
| |-stats_graphs_dnt.r	Statistics and graphics for non-thunderstorm declustered data.
| |-stats_graphs_dt.r	Statistics and graphics for thunderstorm declustered data.
| |-stats_raw_data.r	Statistics for raw data.
| |-stats_raw_data_nt.r	Statistics for non-thunderstorm raw data.
| |-stats_raw_data_t.r	Statistics for thunderstorm raw data.
| |-tnt_csv_1perday.r	Create CSV (thunderstorm and non-thunderstorm) with one data (the maximum) per day.
| |-era5	Folder with specific code for ERA5 data.
| | |-pot_pp_era5.r	POT-PP for ERA5 data. Based in Dr Adam Pintar code.
| | |-maps	Folder with specific code to calculate return levels and plot maps for ERA5 data.
| | | |-return_levels.r	Calculate return levels for ERA5 data.
| | | |-plot_maps.r	Join return levels to cells and plot ERA5 maps.
| |-isd	Folder with specific code for ISD data.
| | |-pot_pp_isd.r	POT-PP for ISD data. Based in Dr Adam Pintar code.
| | |-maps	Folder with code to calculate return levels, do spatial interpolation, and plot maps. ISD data.
| | | |-rl_10_nh.r	Calculate return levels and do spatial interpolation. MRI 10, non-hurricane data.
| | | |-rl_20_nh.r	Calculate return levels and do spatial interpolation. MRI 20, non-hurricane data.
| | | |-rl_50_nh.r	Calculate return levels and do spatial interpolation. MRI 50, non-hurricane data.
| | | |-rl_100_nh.r	Calculate return levels and do spatial interpolation. MRI 100, non-hurricane data.
| | | |-rl_250_nh.r	Calculate return levels and do spatial interpolation. MRI 250, non-hurricane data.
| | | |-rl_500_nh.r	Calculate return levels and do spatial interpolation. MRI 500, non-hurricane data.
| | | |-rl_700_nh.r	Calculate return levels and do spatial interpolation. MRI 700, non-hurricane data.
| | | |-rl_1000_nh.r	Calculate return levels and do spatial interpolation. MRI 1000, non-hurricane data.
| | | |-rl_1700_nh.r	Calculate return levels and do spatial interpolation. MRI 1700, non-hurricane data.
| | | |-rl_3000_nh.r	Calculate return levels and do spatial interpolation. MRI 3000, non-hurricane data.
| | | |-rl_7000_nh.r	Calculate return levels and do spatial interpolation. MRI 7000, non-hurricane data.
| | | |-rl_combined.r	Integrate return levels from hurricane and non-hurricane data.
| | | |-plot_maps.r	Plot ISD maps.
|-downscaling	Folder with code to compare all data sources, looking for downscaling support.
| |-qualitydata	Folder with code to compare using quality data from IDEAM (variable VV_AUT_2).
| | |-VV_AUT_2_1.r	Using predefined list of matching stations (ISD vs IDEAM). ERA5 match is by intersection (1).
| | |-VV_AUT_2_2.r	Using predefined list of matching stations (ISD vs IDEAM). ERA5 match is by intersection (2).
| | |-VV_AUT_2_3.r	Using predefined list of matching stations (ISD vs IDEAM). ERA5 match is by intersection (3).
| |-nonqualitydata	Folder with code to compare using non-quality data from IDEAM (variable VV_AUT_10).
| | |-VV_AUT_10.r	All stations from ISD or IDEAM that intersects one ERA5 cell are compared.

Appendix B

Results - Digital Files

Table B.1: Results. Digital files in FTP site ftp://ftp.geocorp.co/windthesis/. User anonymous@geocorp.co (no password).

Folder Tree - Ftp Links	Description
downscaling	Downscaling Support
|-qualitydata	Quality data comparison (graphics in PDF)
|-nonqualitydata	Non quality data comparison (graphis in PDF)
|-ideam_stations.csv	Ideam Stations
potpp	POT-PP input and output files
|-era5	ERA5 files
| |-FittedModel_*.pdf	ERA5 POT-PP output graphics. See Table B.4.
| |-fitted_model_result.xlsx	Return levels ERA5 (all stations). See Table B.5.
| |-raw_data_station_*_fitted.xlsx	ERA5 POT-PP output parameters by station. See Table B.2.
| |-raw_data_station_*_statistics.xlsx	ERA5 POT-PP time (year, month, week) statistics by station. See Table B.3.
| |-maps	ERA5 raster and vector output data
| | |-era5grid_left_right.*	ERA5 stations shapefile (IDs from left to right, then down)
| | |-era5grid_left_right_pol.*	ERA5 cells shapefile (IDs from left to right, then down)
| | |-era5grid_up_down.*	ERA5 stations shapefile (IDs from top to down, then right)
| | |-era5grid_up_down_pol.*	ERA5 cells shapefile (IDs from top to down, then right)
| | |-rl4326_points_nh_combined.*	ERA5 stations shapefile with all return levels
| | |-combined	ERA5 final wind maps (non-hurricanes + hurricanes). See Table B.6.
| | |-nonhurricanes	ERA5 POT-PP non-hurricane wind maps. See Table B.6.
|-isd	ISD files
| |-01 estaciones - 76 ok isd.txt	ISD list of used stations
| |-01 estaciones - isd - error.txt	One ISD station not working
| |-FittedModel_*.pdf	ISD POT-PP output graphics. See Table B.4.
| |-fitted_model_result.xlsx	Return levels ISD (all stations). See Table B.5.
| |-isd_stations.xlsx	ISD stations
| |-raw_data_station_*_fitted.xlsx	ISD POT-PP output parameters by station. See Table B.2.
| |-raw_data_station_*_statistics.xlsx	ISD POT-PP time (year, month, week) statistics by station. See Table B.3.
| |-maps	ISD raster and vector output data
| | |-rl_nh_h_combined_allcells4326.*	ISD stations shapefile with all return levels
| | |-combined	ISD final wind maps (non-hurricanes + hurricanes). See Table B.7.
| | |-nonhurricanes	ISD POT-PP non-hurricane wind maps. See Table B.7.
| |-raw_data	ISD non-thunderstorm time series (standardized)
|-correction_factors_isd_ideam.xlsx	Correction factors for standardization (ISD and IDEAM)
final_presentation.pdf	Slides for final thesis defence
final_document.pdf	Thesis final report
outfile_nc4c_zip9.nc	ERA5 data. Variable fg10 (3-seconds wind gust)

Table B.2: Content of the output Excel Book 'raw_data_station_*_fitted.xlsx', where '*' is replaced by the Station ID. One file by station.

Excel Sheet Name	Description	Important
nt_evd-fgev_fGumbel	Non thunderstorm. Fitting Gumbel using evd::fgev	Do not use
nt_bbmle-mle2	Non thunderstorm. Fitting Gumbel using bbmle::mle2	Do not use
nt_nll-optim	Non thunderstorm. Fitting Gumbel using negative likelihood and optim	Do not use
nt_fitdistrplus-fitdist	Non thunderstorm. Fitting Gumbel using fitdistrplus::fitdist	Do not use
nt_extRemes	Non thunderstorm. Calculation of return levels POT-GPD, using extRemes::fevd	Do not use
nt_distLquantile_quant	Non thunderstorm. Calculation of return levels and RMSE (POT-GPD and EVDs), using extremeStat::distLquantile	To compare with POT-PP
nt_distLquantile_parameters	Non thunderstorm. Calculation of fitting parameters POT-GPD and EVD, using extremeStat::distLquantile	To compare with POT-PP
nt_distLextreme_returnlev	Non thunderstorm. Calculation of return levels POT-GPD and EVD, using extremeStat::distLextreme	To compare with POT-PP
nt_distLextreme_parameter	Non thunderstorm. Calculation of fitting parameters POT-GPD and EVD, using extremeStat::distLextreme	To compare with POT-PP
nt_POT-GPD-Equivalent	Non thunderstorm. For POT-PP and using POT-GPD equivalent. Calculation of Goodness of Fit and RMSE	Use as Goodness of Fit of POT-PP

Table B.3: Content of the output Excel Book 'raw_data_station_*_statistics.xlsx', where '*' is replaced by the Station ID. One file by station.

Excel Sheet Name	Description
all_years	Raw data time series statistics by year
all_months	Raw data time series statistics by month
all_weeks	Raw data time series statistics by week
all_gaps30days	Raw data gaps of 30 days of more
nt_years	Non-thunderstorm time series statistics by year
nt_months	Non-thunderstorm time series statistics by month
nt_weeks	Non-thunderstorm time series statistics by week
nt_gaps30days	Non-thunderstorm gaps of 30 days of more
IMP.VALS	Main statistics of dataset after declustering and thresholding
declu_nt_years	Non-thunderstorm time series statistics by year, after declustering and thresholding
declu_nt_months	Non-thunderstorm time series statistics by month, after declustering and thresholding
declu_nt_weeks	Non-thunderstorm time series statistics by week, after declustering and thresholding
declu_nt_gaps30days	Non-thunderstorm gaps of 30 days of more, , after declustering and thresholding

Table B.4: Content of the output graphics PDF file 'Fitted-Model_*.pdf', where '*' is replaced by the Station ID. One file by station.

Graphic	Description
Page 1	Time Series Plot for Raw.Data
Page 2	Time Series Plot for Non-Thunderstorm ('nt')
Page 3	Log-Likelihood(Gumbel) - Optim (nll-optim)
Page 4	Declustered - Non-Thunderstorm - fitdistrplus-fitdist(gumbel)
Page 5	Declustered Non-Thunderstorm ('nt') Time Series
Page 6	W-Statistic Plot for best pair of thresholds

Page 7	Declustered - Non-Thunderstorm - Density Function from Intensity Function of Poisson Process
Page 8	Declustered - Non-Thunderstorm - POT-GPD Equivalent
Page 9	Declustered - Non-Thuderstorm - Cumulative Distribution Function from Intensity Function of Poisson Process
Page 10	Declustered - Non-Thunderstorms - Hazard Curve. Poisson Process Intensity Function
Page 11	Declustered - Non-Thunderstorms - Hazard Curve. Gumbel like tail Intensity Function of Poisson Process
Page 12	Non-Thunderstorms. Gumbel Density Function, but using parameters of Poisson Process
Page 13	Non-Thunderstorms. Gumbel Cumulative Distribution, but using parameters of Poisson Process
Page 14	Declustered Non-Thunderstorm. Fitted Gumbel density function using parameters of Poisson Process
Page 15	Declustered - Non-Thunderstorms - Hazard Curve. Gumbel Quantile Function using parameters of Poisson Process

Table B.5: Content of the output Excel Book 'fitted_model_result.xlsx' (sheet pp_pintar). One file by dataset (ISD, ERA5).

Column ID	Columns Name	Important	Description
1	id		Consecurive Row ID
2	t_thresh	Not Available	Threshold for thunderstorm data
3	t_mu_location	Not Available	Location for thunderstorm data
4	t_psi_scale	Not Available	Scale for thunderstorm data
5	nt_thresh		Threshold for non-thunderstorm data
6	nt_mu_location		Location for non-thunderstorm data
7	nt_psi_scale		Scale for non-thunderstorm data
8	distance_w		Minimun W distance to choose best threshold pairs
9	station		Station ID
10 to 20	t_MRI_poissonprocessintfunc	Not Available	Thunderstorm Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using Poisson Process Intensity Function
21 to 31	t_MRI_gumbeltailintfunc	Not Available	Thunderstorm Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using PP Gumbel Tail Intensity Function
32 to 42	t_MRI_gumbelquantilefunc	Not Available	Thunderstorm Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using PP Gumbel Quantile Function
43 to 53	nt_MRI_poissonprocessintfunc	Used to create maps!	Non-thunderstorm Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using Poisson Process Intensity Function
54 to 64	nt_MRI_gumbeltailintfunc	Do not use	Non-thunderstorm Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using PP Gumbel Tail Intensity Function
65 to 75	nt_MRI_gumbelquantilefunc	Do not use	Non-thunderstorm Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using PP Gumbel Quantile Function
76 to 86	tnt_MRI_poissonprocessintfunc	Not Available	Combined (t and nt) Return levels for MRIs (10, 20, 50, 100, 250, 500, 700, 1000, 1700, 3000, 7000), using Poisson Process Intensity Function

Table B.6: ERA5 output maps

File	Description
rl_nonhurricanes_4326_10_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 10 years
rl_nonhurricanes_4326_20_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 20 years
rl_nonhurricanes_4326_50_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 50 years
rl_nonhurricanes_4326_100_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 100 years
rl_nonhurricanes_4326_250_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 250 years
rl_nonhurricanes_4326_500_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 500 years
rl_nonhurricanes_4326_700_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 700 years

rl_nonhurricanes_4326_1000_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 1000 years
rl_nonhurricanes_4326_1700_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 1700 years
rl_nonhurricanes_4326_3000_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 3000 years
rl_nonhurricanes_4326_7000_st.tif	ERA5 POT-PP non-hurricane wind map. MRI 7000 years
rl_combined_4326_10_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 10 years
rl_combined_4326_20_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 210 years
rl_combined_4326_50_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 50 years
rl_combined_4326_100_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 100 years
rl_combined_4326_250_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 250 years
rl_combined_4326_500_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 500 years
rl_combined_4326_700_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 700 years
rl_combined_4326_1000_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 1000 years
rl_combined_4326_1700_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 1700 years
rl_combined_4326_3000_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 3000 years
rl_combined_4326_7000_st.tif	ERA5 final wind map (hurricane + non-hurricane). MRI 7000 years

Table B.7: ISD output maps

File	Description
nh_10.tif	ISD POT-PP non-hurricane wind map. MRI 10 years
nh_20.tif	ISD POT-PP non-hurricane wind map. MRI 20 years
nh_50.tif	ISD POT-PP non-hurricane wind map. MRI 50 years
nh_100.tif	ISD POT-PP non-hurricane wind map. MRI 100 years
nh_250.tif	ISD POT-PP non-hurricane wind map. MRI 250 years
nh_500.tif	ISD POT-PP non-hurricane wind map. MRI 500 years
nh_700.tif	ISD POT-PP non-hurricane wind map. MRI 700 years
nh_1000.tif	ISD POT-PP non-hurricane wind map. MRI 1000 years
nh_1700.tif	ISD POT-PP non-hurricane wind map. MRI 1700 years
nh_3000.tif	ISD POT-PP non-hurricane wind map. MRI 3000 years
nh_7000.tif	ISD POT-PP non-hurricane wind map. MRI 7000 years
isd_combined_4326_10_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 10 years
isd_combined_4326_20_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 210 years
isd_combined_4326_50_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 50 years
isd_combined_4326_100_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 100 years
isd_combined_4326_250_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 250 years
isd_combined_4326_500_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 500 years
isd_combined_4326_700_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 700 years
isd_combined_4326_1000_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 1000 years
isd_combined_4326_1700_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 1700 years
isd_combined_4326_3000_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 3000 years
isd_combined_4326_7000_st.tif	ISD final wind map (hurricane + non-hurricane). MRI 7000 years

Appendix C

ERA5 Data Download and Integration

The European Center for Medium-Range Weather Forecasts - ECMWF had implemented the Climate Data Storage - CDS `https://cds.climate.copernicus.eu/`, where all its datasets can be downloaded, however there is a straightforward way to get ERA5 data through Python library CDSAPI. Before to use CDSAPI, it is necessary to research about names and meanings of ERA5 variables using the official *data documentation* web page `https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation`, or the *parameter database* `https://apps.ecmwf.int/codes/grib/param-db`, that includes all ECWMF data sources, not only ERA5.

Next block of code shows the use of CDSAPI (Python) to download ERA5 variable *10fg - 10 meters wind gust*, from 1979 to 1991 for Colombia area. The most important keywords there, allow to define ERA5 data source and product type, variable name *'variable'*, format netCDF *'format'*, area of interest *'area'*, using WGS88 coordinates in the format *north, west, south, east*, cell size *'grid'* in decimal degrees, and all the keywords related to time *'year'*, *'month'*, *'day'*, *'time'*.

```
import cdsapi
c = cdsapi.Client()
c.retrieve(
'reanalysis-era5-single-levels',
{
   'product_type':'reanalysis',
   'format':'netcdf',
   'variable':'10m_wind_gust_since_previous_post_processing',
   'year':[
       '1979','1980','1981',
       '1982','1983','1984',
       '1985','1986','1987',
       '1988','1989','1990',
```

```
        '1991'
    ],
    'month':[
        '01','02','03',
        '04','05','06',
        '07','08','09',
        '10','11','12'
    ],
    'time':[
        '00:00','01:00','02:00',
        '03:00','04:00','05:00',
        '06:00','07:00','08:00',
        '09:00','10:00','11:00',
        '12:00','13:00','14:00',
        '15:00','16:00','17:00',
        '18:00','19:00','20:00',
        '21:00','22:00','23:00'
    ],
    'day':[
        '01','02','03',
        '04','05','06',
        '07','08','09',
        '10','11','12',
        '13','14','15',
        '16','17','18',
        '19','20','21',
        '22','23','24',
        '25','26','27',
        '28','29','30',
        '31'
    ],
    'area':[12.5, -79.1, -4.5, -66.8], # North, West, South, East.
    'grid':[0.25,0.25]
},
'10fg_1979_1991_netcdf_.25x.25.nc')
```

Next Table C.1, shows all Python scripts used to download variables *10fg - 10 meters wind gust* and *fsr - forecast surface roughness*, for the study area, along the period 1979 to 2019. Last file in the table, hold a summary of commands to manipulate NetCDF files.

After downloading separate netCDF files, there are different tools available, to manipulate them, for instance, Climate Data Operators - CDO `https://code.mpimet.mpg.de/projects/cdo/`, NetCDF command line utilies `https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_working_with_netcdf_files.html`, and NetCDF operator - NCO `http://nco.sourceforge.net/`.

Table C.1: Python Scripts to download ERA5 data, and commands to join netCDF files. ftp://ftp.geocorp.co/windthesis/. User anonymous@geocorp.co (no password).

Folder Tree - Ftp Links	Description
downloadingEra5	Python scripts to download ERA5 data
|-10fg	10fg: 10 meters wind gust ERA5 variable
| |-10fg_1979_1991_netCDF_0.25x0.25.py	Get 10fg variable from 1979 to 1991
| |-10fg_1992_2004_netCDF_0.25x0.25.py	Get 10fg variable from 1992 to 2004
| |-10fg_2005_2013_netCDF_0.25x0.25.py	Get 10fg variable from 2005 to 2013
| |-10fg_2014_2018_netCDF_0.25x0.25.py	Get 10fg variable from 2014 to 2018
| |-10fg_2019_oct_netCDF_0.25x0.25.py	Get 10fg variable from jan to sep of 2019
|-fsr	fsr: forecast surface roughness ERA5 variable
| |-fsr_1979_1991_netCDF_0.25x0.25.py	Get fsr variable from 1979 to 1991
| |-fsr_1992_2004_netCDF_0.25x0.25.py	Get fsr variable from 1992 to 2004
| |-fsr_2005_2013_netCDF_0.25x0.25.py	Get fsr variable from 2005 to 2013
| |-fsr_2014_2018_netCDF_0.25x0.25.py	Get fsr variable from 2014 to 2018
| |-fsr_2019_oct_netCDF_0.25x0.25.py	Get fsr variable from jan to oct of 1991
|-netcdfcommands.txt	Commands to join NETCDF files

Next CDO command, will join by time all netCDF files inside a folder, where *-f* defines the file format, *-b* defines the data format, and *-z* defines the compression level (from 1 to 9). The resulting file name is *outfile_nc4c_zip9.nc*. Then whit *cdo griddes* is possible to review output file content.

```
cdo -f nc4c -b F32 -z zip_9 mergetime *.nc outfile_nc4c_zip9.nc
cdo griddes outfile_nc4c_zip9.nc
```

Using NetCDF command line utilities, next commands will explore the content of a file, and change file format.

```
ncdump -h file.nc
nccopy -k 'nc4' -d 9 file.nc file_d9.nc
```

In Linux, next NCO commands will extract variable p0001 from file.nc to p0001.nc, rename variables p0001 to fsr in file p0001.nc, and view the result.

```
ncks -v p0001 file.nc p0001.nc
ncrename -v p0001,fsr p0001.nc
ncview p0001.nc
```

Appendix D

Thesis Document R Code

This appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` or `echo=FALSE, message=FALSE, warning=FALSE` chunks tag) to help with readability and/or setup.

In Chapter 2 - Data:

1. Install/Load Packages

```
# List of packages required for this analysis
pkg <- c("dplyr", "sf", "ggplot2","rnaturalearth", "rnaturalearthdata", "ggspatial", "kableExtra", "ncdf4", "stars", "magick", "RcmdrMisc", "knitr",
 "ggrepel", "grid", "gridExtra", "cowplot", "xts", "bookdown", "lubridate", "devtools")
# Check if packages are not installed and assign the
# names of the packages not installed to the variable new.pkg
new.pkg <- pkg[!(pkg %in% installed.packages())]
# If there are any packages in the list that aren't installed,
# install them
if (length(new.pkg))
  install.packages(new.pkg, repos = "http://cran.rstudio.com")
# Load packages (thesisdown will load all of the packages as well)
library(thesisdown)
library(dplyr)
library(sf)
library(ggplot2)
library(rnaturalearth)
library(rnaturalearthdata)
library(ggspatial)
library(knitr)
library(kableExtra)
library(ncdf4)
library(stars)
library(magick)
library(RcmdrMisc)
library(ggrepel)
library(grid)
library(gridExtra)
library(cowplot)
library(xts)
library(lubridate)
```

2. Load IDEAM and ISD Stations

```
#Load IDEAM and ISD Stations
con1 = src_postgres(dbname = "winddata", host = "localhost", port = 5432, user = "user1", password = "user1")

#Get Ideam Stations Table
originalfields4 = c("objectid", "codigo1", "nombre", "latitud", "longitud", "categoria")
originalfields4 = paste(originalfields4, collapse= ", ", sep = "")
query4 = paste("select", originalfields4, "from ideam_all_stations", "where inpqrs2 = 'YES'", sep=" ")
ideam_stations = as_tibble(tbl(con1, sql(query4)))
Encoding(ideam_stations$categoria) <- "UTF-8"
Encoding(ideam_stations$nombre) <- "UTF-8"

originalfields3 = c("id", "usaf", "station_name", "latitud", "longitud")
originalfields3 = paste(originalfields3, collapse= ", ", sep = "")
query3 = paste("select", originalfields3, "from isd_all_stations where usaf_isd_dataua != ''", sep=" ")
isd_stations = as_tibble(tbl(con1, sql(query3)))
#Create simple features from Ideam Stations
```

```
ideam_stations = st_as_sf(ideam_stations, coords = c("longitud", "latitud"), crs = 4326)
#Create simple features from ISD stations
isd_stations = st_as_sf(isd_stations, coords = c("longitud", "latitud"), crs = 4326)
```

3. Plot IDEAM Stations

```
#Plot IDEAM Stations
theme_set(theme_bw())
world <- ne_countries(scale = "medium", returnclass = "sf")
world_points<- st_centroid(world)
world_points <- cbind(world, st_coordinates(st_centroid(world$geometry)))


colombia = world_points$name == "Colombia"
panama = world_points$name == "Panama"
peru= world_points$name == "Peru"
brazil= world_points$name == "Brazil"
venezuela= world_points$name == "Venezuela"
ecuador= world_points$name == "Ecuador"

ggplot(data = world) +
  geom_sf(fill= "antiquewhite") +
  geom_text(data= world_points[venezuela,],aes(x=-67, y=8.5, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-79.2, y=9.2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-78.2, y=-1, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75, y=-4, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-67, y=-2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 13, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 4) +
  annotate(geom = "text", x = -80, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 4) +
  geom_sf(data = ideam_stations, size=1, aes(shape=categoria, color=categoria), show.legend = "point") +
  scale_color_discrete(name = 'Category', labels = c("Agrometeorological", "Ordinary Climatic", "Main Climatic", "Mareographic",
   "Special Meteorological", "Main Synoptic")) +
  scale_shape_discrete(name = 'Category', labels = c("Agrometeorological", "Ordinary Climatic", "Main Climatic", "Mareographic",
   "Special Meteorological", "Main Synoptic")) +
  annotation_scale(location = "bl", width_hint = 0.5) +
  annotation_north_arrow(location = "br", which_north = "true", pad_x = unit(0.05, "in"), pad_y = unit(0.05, "in"),
   style = north_arrow_fancy_orienteering) +
  coord_sf(xlim = c(-82.1, -63.8), ylim = c(-7.5, 15.5), expand = FALSE) +
  xlab("Longitude") +
  ylab("Latitude") +
  ggtitle("IDEAM Stations") +
  theme(panel.grid.major = element_line(color = gray(.5), linetype = "dashed", size = 0.5), panel.background = element_rect(fill = "aliceblue"))
#Plote time series - one IDEAM Station
originalfields = c("21205791")
newfields = paste ("X", originalfields, sep="")
originalfields = paste('"', originalfields, '"', sep = "")
newfields = paste('"', newfields, '"', sep = "")

fiedls_query = paste(originalfields, "as", newfields, sep = " ")
fiedls_query = c(paste('"', "mydatetime", '"', sep = ""), fiedls_query)
fiedls_query = paste (fiedls_query, "", sep= "", collapse=", ")

wherestring = c("21205791")
wherestring = paste('"', wherestring, '"', sep = "")
wherestring = paste(wherestring, "IS NOT NULL", sep = " ")
wherestring = paste(wherestring, collapse = " OR " , sep = " ")
query = paste("select", fiedls_query, "from ideam_vvmx_60", "where", wherestring, sep=" ")

all_vvmx_aut_60 = as_tibble(tbl(con1, sql(query)))
timestamp_all_vvmx_aut_60 <- as.POSIXct(as_tibble(select(all_vvmx_aut_60, mydatetime))$mydatetime,format="%Y-%m-%d %H:%M:%S", tz="UTC")

statideam_xts = na.omit(xts(x=select(all_vvmx_aut_60, "X21205791"), order.by = timestamp_all_vvmx_aut_60))

plot.xts(statideam_xts, main = "Station ID: 21205791\nWind Velocity [m/s]", major.ticks="years", format.labels = "%b-%d\n%Y", legend.loc = "top",
  col="green", cex.main=0.3, cex=0.4, cex.axis=0.9, mar = c(2.5,1,0,1), oma = c(0,0,0,0))
```

4. Plot ISD Stations

```
#Plot ISD Stations
theme_set(theme_bw())
world <- ne_countries(scale = "medium", returnclass = "sf")
world_points<- st_centroid(world)
world_points <- cbind(world, st_coordinates(st_centroid(world$geometry)))

colombia = world_points$name == "Colombia"
panama = world_points$name == "Panama"
peru= world_points$name == "Peru"
brazil= world_points$name == "Brazil"
venezuela= world_points$name == "Venezuela"
ecuador= world_points$name == "Ecuador"

ggplot(data = world) +
  geom_sf(fill= "antiquewhite") +
  geom_text(data= world_points[venezuela,],aes(x=-67, y=8.5, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-79.2, y=9.2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-78.2, y=-1, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75, y=-4, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-67, y=-2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 13, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 4) +
```

```
  annotate(geom = "text", x = -80, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 4) +
  geom_sf(data = isd_stations, size=1, aes(color= "ISD Stations"), shape=2, show.legend = "point") +
  scale_color_manual(values = c("ISD Stations" = "black"), name="") +
  annotation_scale(location = "bl", width_hint = 0.5) +
  annotation_north_arrow(location = "br", which_north = "true", pad_x = unit(0.05, "in"), pad_y = unit(0.05, "in"),
   style = north_arrow_fancy_orienteering) +
  coord_sf(xlim = c(-82.1, -63.8), ylim = c(-7.5, 15.5), expand = FALSE) +
  xlab("Longitude") +
  ylab("Latitude") +
  ggtitle("Integrated Surface Database - ISD") +
  theme(panel.grid.major = element_line(color = gray(.5), linetype = "dashed", size = 0.5), panel.background = element_rect(fill = "aliceblue"))
#Plot - ISD Station
originalfields1 = c("802590")
newfields1 = paste ("X", originalfields1, sep="")
originalfields1 = paste('"', originalfields1, '"', sep = "")
newfields1 = paste('"', newfields1, '"', sep = "")
fiedls_query1 = paste(originalfields1, "as", newfields1, sep = " ")
fiedls_query1 = c(paste('"', "mydatetime", '"', sep = ""), fiedls_query1)
fiedls_query1 = paste (fiedls_query1, "", sep= "", collapse=", ")

wherestring1 = c("802590")
wherestring1 = paste('"', wherestring1, '"', sep = "")
wherestring1 = paste(wherestring1, "IS NOT NULL", sep = " ")
wherestring1 = paste(wherestring1, collapse = " OR " , sep = " ")
query1 = paste("select", fiedls_query1, "from isd_lite_unstack", "where", wherestring1, sep=" ")

isdlite = as_tibble(tbl(con1, sql(query1)))

timestamp_isdlite <- as.POSIXct(as_tibble(select(isdlite, mydatetime))$mydatetime,format="%Y-%m-%d %H:%M:%S", tz="UTC")

statisd_xts = na.omit(xts(x=select(isdlite, "X802590"), order.by = timestamp_isdlite))

plot.xts(statisd_xts, main = "Station ID: 802590\nWind Velocity [m/s]", major.ticks="years", format.labels = "%b-%d\n%Y", legend.loc = "top",
 col="green", cex.main=0.3, cex=0.4, cex.axis=0.9, mar = c(2.5,1,0,1), oma = c(0,0,0,0))
```

5. Load ERA5 Stations

```
#Load ERA5 netCDF dataset - variable fg10
ncname <- "outfile_nc4c_zip9"
filename <- paste("./data/", ncname, ".nc", sep = "")
ncin <- nc_open(filename)
lon <- ncvar_get(ncin, "longitude")
nlon = dim(lon)
lat <- ncvar_get(ncin, "latitude")
nlat = dim(lat)
ntime <-  dim(ncvar_get(ncin, "time"))
variablename <- "fg10"
fg10.units <- ncatt_get(ncin, variablename, "units")
fg10.units
lonlat.unstack <- expand.grid(lon=as.numeric(lon), lat=as.numeric(lat))
#Create ERA5 centers (point with lat, lon, and value, as cell index)
era5colpoints = st_as_sf(lonlat.unstack, coords=1:2, crs=st_crs(4326))
era5colpoints$value = 1:(nlon*nlat)
#Define stars object to match with ERA5 bounding box.
#Cell centers of stars object, need to be same cell centers of ERA5
pointsbbox = st_bbox(era5colpoints)
cellsize = lonlat.unstack$lon[2]- lonlat.unstack$lon[1]
mybbox = st_bbox(c(pointsbbox$xmin - (cellsize/2), pointsbbox$xmax + (cellsize/2), pointsbbox$ymax + (cellsize/2), pointsbbox$ymin - (cellsize/2)),
 crs = st_crs(4326))
era5colraster.st = st_rasterize(era5colpoints, st_as_stars(mybbox, nx = nlon, ny = nlat, values = era5colpoints$value))
#Load ERA5 polygon vectors, representing cells of ERA5
file_era5_sf_pol = "./data/era5grid_left_right_pol.shp"
era5_4326_sf_pol = st_read(dsn=file_era5_sf_pol)
pts <- do.call(rbind, st_centroid(st_geometry(era5_4326_sf_pol)))
x = pts[,1]
y = pts[,2]
era5_4326_sf_pol$x = x
era5_4326_sf_pol$y = y
era5_4326_sf_pol_filter_corners = era5_4326_sf_pol %>% filter(DN %in% c(1, 49, 3333, 3381))
era5_4326_sf_pol_filter_corners_left = era5_4326_sf_pol %>% filter(DN %in% c(1, 3333))
era5_4326_sf_pol_filter_corners_right = era5_4326_sf_pol %>% filter(DN %in% c(49, 3381))
```

6. Plot ERA5 Stations

```
#Plot ERA5 Stations
theme_set(theme_bw())
world <- ne_countries(scale = "medium", returnclass = "sf")
world_points<- st_centroid(world)
world_points <- cbind(world, st_coordinates(st_centroid(world$geometry)))

colombia = world_points$name == "Colombia"
panama = world_points$name == "Panama"
peru= world_points$name == "Peru"
brazil= world_points$name == "Brazil"
venezuela= world_points$name == "Venezuela"
ecuador= world_points$name == "Ecuador"

big = ggplot(data = world) +
  geom_sf(fill= "antiquewhite") +
```

```
  geom_sf(data=era5colpoints, size=0.1, aes(color = "Stations"), shape=".",  show.legend = "point")+
  scale_color_manual(values = c("Stations" = "black"), name="ERA5", guide = guide_legend(override.aes = list(fill= c(NA), linetype = c("blank"),
   shape = c("."))) +
  geom_sf(data = era5_4326_sf_pol_filter_corners, color = "black", aes(fill="Cells"), size=0.1, alpha=1, show.legend = "polygon") +
  scale_fill_manual(values = c("Cells" = NA), name="", guide = guide_legend(override.aes = list(fill = c(NA), shape = c(NA), size=0.1))) +
  geom_rect(mapping=aes(xmin=-79.252968100, xmax=-78.247031900, ymin=11.832846362, ymax=12.667153638), color="red", alpha=0, size=0.1) +
  geom_rect(mapping=aes(xmin=-67.752968100, xmax=-66.747031900, ymin=11.832846362, ymax=12.667153638), color="red", alpha=0, size=0.1) +
  geom_rect(mapping=aes(xmin=-79.258632089, xmax=-78.241367911, ymin=-4.671851259, ymax=-3.828148741), color="red", alpha=0, size=0.1) +
  geom_rect(mapping=aes(xmin=-67.758632089, xmax=-66.741367911, ymin=-4.671851259, ymax=-3.828148741), color="red", alpha=0, size=0.1) +
  geom_text(data= world_points[venezuela,],aes(x=-66.3, y=8.5, label=name), color = "darkblue", fontface = "bold", size=2, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-79.7, y=9.2, label=name), color = "darkblue", fontface = "bold", size=2, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-79.5, y=-1.5, label=name), color = "darkblue", fontface = "bold", size=2, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75.5, y=-5.2, label=name), color = "darkblue", fontface = "bold", size=2, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-66, y=-2, label=name), color = "darkblue", fontface = "bold", size=2, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 14, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 2) +
  annotate(geom = "text", x = -80, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 2) +
  geom_text_repel(data = era5_4326_sf_pol_filter_corners_left, size=2, aes(x=x, y=y, label = DN), direction="y", segment.size=0.1,
   segment.color= "grey50", color="grey50", nudge_x=-1, hjust=1, box.padding=0.1) +
  geom_text_repel(data = era5_4326_sf_pol_filter_corners_right, size=2, aes(x=x, y=y, label = DN), direction="y", segment.size=0.1,
   segment.color= "grey50", color="grey50", nudge_x=+1, hjust=0, box.padding=0.1) +
  coord_sf(xlim = c(-81.1, -64.8), ylim = c(-5, 13), expand = FALSE) +
  xlab("") +
  ylab("") +
  ggtitle("ERA5 Reanalysis - Forecast") +
  theme(plot.title = element_text(size=8)) +
  theme(axis.text.x= element_text(size=7)) +
  theme(axis.text.y= element_text(size=7)) +
  theme(panel.grid.major = element_line(color = gray(.5), linetype = "dashed", size = 0.1)) +
  theme(panel.background = element_rect(fill = "aliceblue")) +
  theme(legend.title = element_text(size=8)) +
  theme(legend.text=element_text(size=8)) +
  theme(legend.key.size = unit(0.5,"line")) +
  theme(plot.margin=unit(c(0,0,0,0),"cm")) +
  theme(axis.text.x = element_text(margin =  margin(t =2, b = -10))) +
  theme(axis.text.y = element_text(margin =  margin(r =2, l = -10)))

corner1lt = ggplot(data = world) +
  geom_sf(fill= "antiquewhite",  size=0.1) +
  geom_sf(data = era5_4326_sf_pol, colour="black", fill=NA, size=0.1) +
  geom_text(data= world_points[venezuela,],aes(x=-66.5, y=8.5, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-80.5, y=9.2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-79.2, y=-1, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-68, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[colombia,],aes(x=-71, y=4, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 14, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 4) +
  annotate(geom = "text", x = -80.5, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 4) +
  geom_sf(data = st_cast(world, "MULTILINESTRING"),  size=0.1)+
  geom_sf_text(data = era5_4326_sf_pol, aes(label = DN), size=2) +
  coord_sf(xlim = c(-79.252968100, -78.247031900), ylim = c(11.832846362, 12.667153638), expand = FALSE) +
  xlab("") +
  ylab("") +
  ggtitle("") +
  theme(panel.grid = element_blank()) +
  theme(panel.background = element_rect(fill = "aliceblue")) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank()) +
  theme(axis.ticks = element_blank()) +
  theme(plot.margin=grid::unit(c(0,0.2,0,0),"cm")) +
  theme(panel.border = element_rect(colour = "red"))+
  theme(axis.ticks.length=unit(0, "null")) +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(plot.title = element_blank())

corner2rt = ggplot(data = world) +
  geom_sf(fill= "antiquewhite",  size=0.1) +
  geom_sf(data = era5_4326_sf_pol, colour="black", fill=NA, size=0.1) +
  geom_text(data= world_points[venezuela,],aes(x=-66.5, y=8.5, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-80.5, y=9.2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-79.2, y=-1, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-68, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[colombia,],aes(x=-71, y=4, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 14, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 4) +
  annotate(geom = "text", x = -80.5, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 4) +
  geom_sf(data = st_cast(world, "MULTILINESTRING"),  size=0.1)+
  geom_sf_text(data = era5_4326_sf_pol, aes(label = DN), size=2) +
  coord_sf(xlim = c(-67.752968100, -66.747031900), ylim = c(11.832846362, 12.667153638), expand = FALSE) +
  xlab("") +
  ylab("") +
  ggtitle("") +
  theme(panel.grid = element_blank()) +
  theme(panel.background = element_rect(fill = "aliceblue")) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank()) +
  theme(axis.ticks = element_blank()) +
  theme(plot.margin=grid::unit(c(0,0,0,0.2),"cm")) +
  theme(panel.border = element_rect(colour = "red"))+
  theme(axis.ticks.length=unit(0, "null")) +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(plot.title = element_blank())
```

```
corner3lb = ggplot(data = world) +
  geom_sf(fill= "antiquewhite",  size=0.1) +
  geom_sf(data = era5_4326_sf_pol, colour="black", fill=NA, size=0.1) +
  geom_text(data= world_points[venezuela,],aes(x=-66.5, y=8.5, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-80.5, y=9.2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-79.2, y=-1, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-68, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[colombia,],aes(x=-71, y=4, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 14, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 4) +
  annotate(geom = "text", x = -80.5, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 4) +
  geom_sf(data = st_cast(world, "MULTILINESTRING"),  size=0.1)+
  geom_sf_text(data = era5_4326_sf_pol, aes(label = DN), size=2) +
  coord_sf(xlim = c(-79.258632089, -78.241367911), ylim = c(-4.671851259, -3.828148741), expand = FALSE) +
  xlab("") +
  ylab("") +
  ggtitle("") +
  theme(panel.background = element_rect(fill = "aliceblue")) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank()) +
  theme(axis.ticks = element_blank()) +
  theme(plot.margin=grid::unit(c(0,0.2,0,0),"cm")) +
  theme(panel.border = element_rect(colour = "red"))+
  theme(axis.ticks.length=unit(0, "null")) +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(plot.title = element_blank())

corner4rb = ggplot(data = world) +
  geom_sf(fill= "antiquewhite",  size=0.1) +
  geom_sf(data = era5_4326_sf_pol, colour="black", fill=NA, size=0.1) +
  geom_text(data= world_points[venezuela,],aes(x=-66.5, y=8.5, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[panama,],aes(x=-80.5, y=9.2, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[ecuador,],aes(x=-79.2, y=-1, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[peru,],aes(x=-75, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[brazil,],aes(x=-68, y=-6, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  geom_text(data= world_points[colombia,],aes(x=-71, y=4, label=name), color = "darkblue", fontface = "bold", size=3, check_overlap = FALSE) +
  annotate(geom = "text", x = -77.5, y = 14, label = "Caribbean\nSea", fontface = "italic", color = "grey22", size = 4) +
  annotate(geom = "text", x = -80.5, y = 5, label = "Pacific\nSea", fontface = "italic", color = "grey22", size = 4) +
  geom_sf(data = st_cast(world, "MULTILINESTRING"),  size=0.1)+
  geom_sf_text(data = era5_4326_sf_pol, aes(label = DN), size=2) +
  coord_sf(xlim = c(-67.758632089, -66.741367911), ylim = c(-4.671851259, -3.828148741), expand = FALSE) +
  xlab("") +
  ylab("") +
  ggtitle("") +
  theme(panel.background = element_rect(fill = "aliceblue")) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank()) +
  theme(axis.ticks = element_blank()) +
  theme(plot.margin=grid::unit(c(0,0,0,0.2),"cm")) +
  theme(panel.border = element_rect(colour = "red"))+
  theme(axis.ticks.length=unit(0, "null")) +
  theme(axis.ticks.margin=unit(0, "null")) +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(plot.title = element_blank())

grid.arrange(big, arrangeGrob(corner1lt, corner2rt, corner3lb, corner4rb), ncol=2, widths=c(2.2,1))
```

In Chapter 3 - Theoretical Framework:

1. Install/Load Packages

```
# List of packages required for this analysis
pkg <- c("RcmdrMisc")
# Check if packages are not installed and assign the
# names of the packages not installed to the variable new.pkg
new.pkg <- pkg[!(pkg %in% installed.packages())]
# If there are any packages in the list that aren't installed,
# install them
if (length(new.pkg))
  install.packages(new.pkg, repos = "http://cran.rstudio.com")
# Load packages (thesisdown will load all of the packages as well)
library(RcmdrMisc)
```

2. Plot Gumbel *pdf*

```
par(mar=c(2.5,2.5,2,0))
par(oma=c(0,0,0,0))
par(mgp=c(1.5,0.5,0))
location = 100
scale = 40
.x <- seq(0, 300, length.out=1000)
pdfG <- function(x) {
  1/location *exp(-(x-location)/scale)*exp(-exp(-(x-location)/scale))
  }
.y = pdfG(.x)
plot(.x, .y, col="green", lty=4, xlab="Velocities Km/h", ylab="Density Function - Gumbel Distribution", cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7,
 main=paste("Gumbel - Density Function Gumbel Distribution\n", "Location=", round(location,2), " Scale=", round(scale,2)), type="l", cex.sub=0.6)
```

```
par(mar=c(2.5,2.5,2,0))
par(oma=c(0,0,0,0))
par(mgp=c(1.5,0.5,0))
location = 100
scale = 40
.x <- seq(0, 300, length.out=1000)
dfG = dgumbel(.x, location=location, scale=scale)
plot(.x, dfG, col="red", lty=4, xlab="Velocities Km/h", ylab="Density Function - Gumbel Distribution", cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7,
 main=paste("Gumbel - Density Function Gumbel Distribution\n", "Location=", round(location,2), " Scale=", round(scale,2)), type="l", cex.sub=0.6)
```

3. Plot Gumbel *cdf*

```
par(mar=c(2.5,2.5,2,0))
par(oma=c(0,0,0,0))
par(mgp=c(1.5,0.5,0))
location = 100
scale = 40
.x <- seq(0, 300, length.out=1000)
cdfG <- function(x) {
  exp(-exp(-(x-location)/scale))
  }
.y = cdfG(.x)
plot(.x, .y, col="green", lty=4, xlab="Velocities Km/h", ylab="Probability", cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, type="l",
 main=paste("Gumbel - Cumulative Distribution Function\n", "Location=", round(location,2), " Scale=", round(scale,2)), cex.sub=0.6)
```

4. Plot Gumbel *ppf*

```
par(mar=c(2.5,2.5,2,0))
par(oma=c(0,0,0,0))
par(mgp=c(1.5,0.5,0))
location = 100
scale = 40
.x <- seq(0, 1, length.out=1000)
ppfG <- function(x) {
  location - (scale*log(-log(x)))
  }
.y = ppfG(.x)
plot(.x, .y, col="green", lty=4, ylab="Velocities Km/h", xlab="Probability", cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6,
 main=paste("Gumbel - Percent Point Function\n", "Location=", round(location,2), " Scale=", round(scale,2)), type="l")
```

5. Plot Gumbel *hf*

```
par(mar=c(2.5,2.5,2,0))
par(oma=c(0,0,0,0))
par(mgp=c(1.5,0.5,0))
location = 100
scale = 40
.x <- seq(0, 1500, length.out=1000)
hfG <- function(x) {
  (1/scale)*(exp(-(x-location)/scale))/(exp(exp(-(x-location)/scale))-1)
  }
.y = hfG(.x)
plot(.x, .y, col="green", lty=4, xlab="Velocities Km/h", ylab="Hazard", cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6,
 main=paste("Gumbel - Hazard Function\n", "Location=", round(location,2), " Scale=", round(scale,2)), type="l", xlim=c(0,500))
```

6. Compound Exceedance Probability - Pn

```
par(mar=c(3,3,0,0))
par(oma=c(0,0,0,0))
par(mgp=c(2,1,0))
plot(1, type="n", xlab=expression(paste("Compound Probability ", P[n])), ylab="Exposure Time as a Multiple of MRI",
 xlim=c(0,1), ylim= c(0,2500), xaxt ="n", yaxt="n", bty="n", cex.lab=0.7)

y1 = c(0, 500,1000,1500,2000,2500)
text(y=y1,  x=par("usr")[1], labels = y1/500, srt = 0, pos = 2, xpd = TRUE, cex=0.8)
y1 = 500*0.69
text(y=y1,  x=par("usr")[1], labels = ".69", srt = 0, pos = 2, xpd = TRUE, cex=0.6)
y1 = 2250
text(y=y1,  x=par("usr")[1], labels = expression(paste("4",frac(1,2))), srt = 0, pos = 2, xpd = TRUE, cex=0.6)

npn <- function(x) (1-(1/500))^x #Event will not occur
n = seq(from=0, to=2500, by=1)
mynpn = npn(n)

lines(x=mynpn,y=n, col= "blue")

pn <- function(x) 1-(1-(1/500))^x #Event will occur
mypn= pn(n)
lines(x=mypn,y=n, col= "green")

text(x=c(0.01, 0.37,0.63, 0.99),  par("usr")[3], labels = c(".01",".37",".63",".99") , srt = 0, pos = 1, xpd = TRUE, cex=0.6)

axis(1, at= seq(from=0, to=1, by= 0.1), labels=seq(from=0, to=1, by= 0.1), tick=TRUE, col.axis="black", cex=0.8)

axis(2, at=c(0, 500,1000,1500,2000,2500),labels=FALSE, tick=TRUE, col.axis="black")
axis(2, at=c(345, 2250),labels=FALSE, tick=TRUE, col.axis="black", tck=-0.015)
```

```
axis(1, at=c(0.01,.37,.63,0.99),labels=FALSE, tick=TRUE, col.axis="black", tck=-0.015)

abline(v=c(0.01, 0.37,0.5, 0.63,0.99), lty="dotted")
abline(h=c(345,500, 2250), lty="dotted")
text(x=0.15,  y=1800, labels = "chance event\nwill not occur", cex=0.7)
text(x=0.85,  y=1800, labels = "chance event\nwill occur", cex=0.7)
```

In Chapter 4 - Methodology:

1. Combine Hazard Curve

```
plotit<-function(){

  par(mar=c(2,2,0,0))
  par(oma=c(0,0,0,0))
  par(bg=NA)
  plot(1, xlab='', ylab='', type='n', yaxt='n', xaxt='n', tck=0, xlim=c(0,200), ylim=c(0,0.05), bg = 'transparent', bty="n")

  arrows(0,0,0,0.05, length=0.04)
  arrows(0,0,200,0, length=0.04)

  text(x = par("usr")[2] - 5,  y = par("usr")[3] - 0.005, labels = expression(frac(1, N)), xpd = NA, srt = 0, cex = 0.7)

  text(x = par("usr")[1] - 6,  y = 0.05, labels = expression(Y[N]), xpd = NA, srt = 0, cex = 0.7)

  text(x = 50.2,  y = par("usr")[3] - 0.003, labels = "?", xpd = NA, srt = 0, cex = 0.7)

  text(x = 68.5,  y = par("usr")[3] - 0.003, labels = "0.02", xpd = NA, srt = 0, cex = 0.7)

  text(x = 100,  y = par("usr")[3] - 0.003, labels = "0.03", xpd = NA, srt = 0, cex = 0.7)

  text(x = par("usr")[1] - 10,  y = 0.015, labels = expression(paste("30 ", frac(Km, h))), xpd = NA, srt = 0, cex = 0.7)

  text(x = par("usr")[2] - 2,  y = 0.048, labels = "Combined", xpd = NA, srt = 0, pos = 2, cex = 0.6)

  text(x = par("usr")[2] - 2,  y = 0.031, labels = "Hurricanes", xpd = NA, pos = 2, srt = 0, cex = 0.6)

  text(x = par("usr")[2] - 2,  y = 0.021, labels = "Non-Hurricanes", xpd = NA, pos = 2, srt = 0, cex = 0.6)

  myexp = expression(paste(P[e], " = 1 - (1 -", P[nh], ") * (1 - ", P[h], ")"))

  text(x = par("usr")[1] + 60,  y = 0.049, labels = myexp, xpd = NA, srt = 0, cex = 0.7)

  text(x = par("usr")[1] + 60,  y = 0.045, labels = "? = 1- (1 - 0.03)(1-0.02)", xpd = NA, srt = 0, cex = 0.6)

  location = 65
  scale = 20
  .x <- seq(0, 1500, length.out=1000)
  hfG <- function(x) {
    (1/scale)*(exp(-(x-location)/scale))/(exp(exp(-(x-location)/scale))-1)
  }
  curve(hfG, add=T, col="red", lwd=1, lty=5)
  Arrows (x0=50.2, y0=0, x1=50.2, y1=(hfG(50.2)-0.003), arr.type="triangle", arr.width=0.04, lwd=0.1)

  location = 80
  scale = 30
  .x <- seq(0, 1500, length.out=1000)
  curve(hfG, add=T, col="red", lwd=1)
  Arrows (x0=68.5, y0=0, x1=68.5, y1=(hfG(68.5)-0.003), arr.type="triangle", arr.width=0.04, lwd=0.1)

  location = 100
  scale = 40
  .x <- seq(0, 1500, length.out=1000)
  curve(hfG, add=T, col="red", lwd=1)
  Arrows (x0=100, y0=0, x1=100, y1=(hfG(100)-0.003), arr.type="triangle", arr.width=0.04, lwd=0.1)
  Arrows (x0=100, y0=hfG(100), x1=7, y1=hfG(100) , arr.type="triangle", arr.width=0.04, lwd=0.1)
}

z.plot1<-function(){plotit()}

mydataframe = data.frame(v = c(10, 20, 30, "...", 350, "..."), Pe = c("...", "...", "?", "...", "...", "..."))

names(mydataframe) <- c(expression(Y[N]), expression(P[e]))

tt <- ttheme_default(base_size = 7, colhead=list(fg_params = list(parse=TRUE)))
tbl <- tableGrob(mydataframe, rows=NULL, theme=tt)

plot_grid(z.plot1, tbl, ncol = 2, rel_widths = c(4,1), labels=c("", "Combined Curve"), label_size = 7, hjust=-0.13)
```

Appendix E

User Manual

References

ADB. (2014). *Guidelines for wind resource assessment: Best practices for countries initiating wind development*. Asian Development Bank. Retrieved from https://www.ebook.de/de/product/30686652/guidelines_for_wind_resource_assessment.html

Beirlant, J., Goegebeur, Y., Teugels, J., & Segers, J. (2004). *Statistics of extremes: Theory and applications*. John Wiley & Sons, Ltd. http://doi.org/10.1002/0470012382

CIMNE, I. (2015). *Update on the probabilistic modelling of natural risks at global level: Global risk model* (technical report). The United Nations Office for Disaster Risk Reduction - UNISDR. Retrieved from https://www.preventionweb.net/english/hyogo/gar/2015/en/bgdocs/CIMNE-INGENIAR,%202014a.pdf

CIMNE, I., ITEC. (2017). *Metodología de modelación probabilista de riesgos naturales* (technical report No. ERN-CAPRA-T1-3). CAPRA- Probabilistic Risk Assessment Initiative. Retrieved from https://ecapra.org/sites/default/files/documents/ERN-CAPRA-R6-T1-3%20-%20Modelos%20de%20Evaluaci%C3%B3n%20de%20Amenazas.pdf

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer London. http://doi.org/10.1007/978-1-4471-3675-0

Comarazamy, D. (2005). *Disaster mitigation in health facilities. Wind effects. Structural issues.* Pan American Health Organization. Retrieved from http://www.disaster-info.net/viento/english/guiones/structural.pdf

Council, N. R. (1994). Hurricane hugo, puerto rico, the virgin islands, and charleston, south carolina, september 17-22, 1989. In (pp. 247–257). Washington, DC: National Academies Press. http://doi.org/10.17226/1993

C. S. Durst, B. A., O. B.E. (1960). Wind speeds over short periods of time. *The Meteorological Magazine*, *89*(1056), 181–187. Retrieved from https://www.depts.ttu.edu/nwi/Pubs/ReportsJournals/ReportsJournals/Windspeeds.pdf

Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*(3), 393–442. Retrieved from http://www.jstor.org/stable/2345667

Deville, Y., & IRSN. (2016). *Renext: Renewal method for extreme values extrapolation*. Retrieved from https://CRAN.R-project.org/package=Renext

Engineers, A. S. O. C. (2017). *Minimum design loads and associated criteria for buildings and other structures (asce7-16)*. American Society of Civil Engineers. Retrieved from https://www.ebook.de/de/product/35017614/american_society_of_civil_engineers_minimum_design_loads_and_associated_criteria_for_buildings_and_other_structures_7_16.html

European Centre For Medium-Range Weather Forecasts. (2017). ERA5 reanalysis. UCAR/NCAR - Research Data Archive. http://doi.org/10.5065/D6X34W69

Gilleland, E. (2019). *ExtRemes: Extreme value analysis*. Retrieved from https://CRAN.R-project.org/package=extRemes

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, *8*(1), 204–218. Retrieved from https://journal.r-project.org/archive/2016/RJ-2016-014/index.html

Haigh, I. D., & Wahl, T. (2019). Advances in extreme value analysis and application to natural hazards. *Natural Hazards*, *98*(3), 819–822. http://doi.org/10.1007/s11069-019-03628-x

Harris, J. W., & Stocker, H. (1998). Maximum likelihood method. In *Handbook of mathematics and computational science* (p. 824). Springer-Verlag.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis*. Cambridge University Press. http://doi.org/10.1017/cbo9780511529443

IDEAM. (1999, June). Aeronautical information. Annual wind regime. Web Page. Retrieved from http://bart.ideam.gov.co/cliciu/rosas/viento.htm

IDEAM. (2005). *Protocolo toma de datos de campo y emplazamiento de estaciones meteorológicas*.

Janet E. Heffernan with R port, O. S. functions written by, & Alec G. Stephenson., R. documentation provided by. (2018). *Ismev: An introduction to statistical modeling of extreme values*. Retrieved from https://CRAN.R-project.org/package=ismev

Kubler, J. (1994). *Computational Statistics &Amp; Data Analysis*, *18*(4), 473–474. Retrieved from https://EconPapers.repec.org/RePEc:eee:csdana:v:18:y:1994:i:4:p:473-474

Lettau, H. (1969). Note on aerodynamic roughness-parameter estimation on the basis of roughness-element description. *Journal of Applied Meteorology*, *8*(5), 828–832. http://doi.org/10.1175/1520-0450(1969)008%3C0828:NOARPE%3E2.0.CO;2

Masters, F. J., Vickery, P. J., Bacon, P., & Rappaport, E. N. (2010). Toward objective, standardized intensity estimates from surface wind speed observations. *Bulletin of the American Meteorological Society*, *91*(12), 1665–1682. http://doi.org/10.1175/2010bams2942.1

Ministerio de Vivienda, C. y T. (2010). *Reglamento colombiano de construccción sismo*

resistente - nsr-10. Carrera 20 # 84-14, oficina 502, Bogotá.: Asociación Colombiana de Ingeniería Sísmica. Comisión Asesora Permanente.

NIST. (2012, February). Standardized extreme wind speed database for the united states. Web Page. Retrieved from `https://www.itl.nist.gov/div898/winds/NIST_TN/nist_tn.htm`

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. `http://doi.org/10.32614/RJ-2018-009`

Pebesma, E. (2019a). *Sf: Simple features for r*. Retrieved from `https://CRAN.R-project.org/package=sf`

Pebesma, E. (2019b). *Stars: Spatiotemporal arrays, raster and vector data cubes*.

Pebesma, E., & Graeler, B. (2019). *Gstat: Spatial and spatio-temporal geostatistical modelling, prediction and simulation*. Retrieved from `https://CRAN.R-project.org/package=gstat`

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers & Geosciences*, *30*, 683–691.

Pfaff, B., & McNeil, A. (2018). *Evir: Extreme values in r*. Retrieved from `https://CRAN.R-project.org/package=evir`

Pickands, J. (1971). The two-dimensional poisson process and extremal processes. *Journal of Applied Probability*, *8*(4), 745–756. `http://doi.org/10.2307/3212238`

Pintar, A. L., Simiu, E., Lombardo, F. T., & Levitan, M. L. (2015). *Simple guide for evaluating and expressing the uncertainty of NIST MeasuremenMaps of non-hurricane non-tornadic wind speeds with specified mean recurrence intervals for the contiguous united states using a two-dimensional poisson process extreme value model and local regressiont results*. National Institute of Standards; Technology.

Rezapour, M., & Baldock, T. E. (2014). Classification of hurricane hazards: The importance of rainfall. *Weather and Forecasting*, *29*(6), 1319–1331. `http://doi.org/10.1175/waf-d-14-00014.1`

Roberts, S. (2012). *Wind wizard: Alan g. Davenport and the art of wind engineering*. Princeton University Press. Retrieved from `https://books.google.de/books?id=e2eYDwAAQBAJ`

Simiu, E., & Scanlan, R. H. (1996). *Wind effects on structures : Fundamentals and applications to design* (3rd ed.). New York : John Wiley. Retrieved from `http://lib.ugent.be/catalog/rug01:001267836`

Smith, A., Lott, N., & Vose, R. (2011). The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, *92*(6), 704–708. `http://doi.org/10.1175/2011bams3015.1`

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to

trend detection in ground-level ozone. *Statistical Science*, *4*(4), 367–377. http://doi.org/10.1214/ss/1177012400

Smith, R. L. (2004). Extreme values in finance, telecommunications, and the environment (chapman & hall/crc monographs on statistics and applied probability). In B. F. inkenstädt & H. Rootzén (Eds.), (pp. 1–78). Chapman; Hall/CRC. Retrieved from https://www.amazon.com/Telecommunications-Environment-Monographs-Statistics-Probability/dp/1584884118?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1584884118

Stephenson, A. G. (2002). Evd: Extreme value distributions. *R News*, *2*(2), 0. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Wuertz, D., Setz, T., & Chalabi, Y. (2017). *FExtremes: Rmetrics - modelling extreme events in finance*. Retrieved from https://CRAN.R-project.org/package=fExtremes