Application of Bayesian Networks to Estimate River Discharges in Ecuador

Gina Alexandra Torres Alves

Additional Graduation Work (CIE5050-09)

Cover Source: https://www.flickr.com/photos/126958916@N04/33059604111/in/photostream/

Application of Bayesian Networks to Estimate River Discharges in Ecuador

By

Gina Alexandra Torres Alves

as completion of an Additional Graduation Work at the Delft University of Technology,

Student Number:	4625951	
Project Duration:	September 2017-February 2018	
Assessment Committee:	Dr. ir. O. Morales-Nápoles	TU Delft-Supervisor.
	Dr. Ir. M. Hrachowitz.	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Acknowledgements

I would like to thank to Dr. Ir. Oswaldo Morales for his advice and continuous support and understanding during the development of this project. Additionally, I would like to thank Dr. Markus Hrachowitz for his vital advice during our meetings. Without their passionate participation and input, this project could have not been successfully completed.

Finally, I would like express my gratitude to my family and friends from providing me with unfailing support and encouragement through this project. This project would not have been possible without them. Thank you.

Author

Gina Torres

Contents

1. Introduction

1.1. Background

Floods are considered the most common natural hazard and third most damaging globally after storms and earthquakes (Wilby & Keenan, 2012) they are defined as a temporary covering of water outside its normal confines (FLOODsite-Consortium, 2005; cf.Munich Reinsurance Company, 1997) and can be classified according to the event that causes it, for example: Winter rainfall floods, summer convectional storm induced floods, snow-melt floods, sea surge and tidal floods, tsunamis, rising ground water floods, dam break, urban sewer floods or reservoir control floods (Schanze, 2007). In this research attention will be given only to floods in rivers.

A flood is characterized by its water depth, flow velocity, matter fluxes and temporal and spatial dynamics and is considered a threat when it occurs in catchments that are greatly used and often influenced by man by land use, river training, etc. (Schanze, 2007). Then, given a flood scenario, is possible to quantify the damage that such flood will cause at certain location, this is mostly known as: Estimation of flood hazard, and it means that some elements at the study location will be damaged and others will not (FLOODsite-Consortium, 2005) in other words, this is defined as the occurrence of a potentially damaging flood (Schanze, 2007). The damage produced on the elements is characterized by their vulnerability (Schanze, 2007). Vulnerability can be classified in three categories depending on its principle of sustainability: social and cultural, economic and ecological (Sarewitz, 2003). Finally, the concept of flood risk arises, as it is defined as the probability of negative consequences due to floods and depends on the exposure of elements at risk to flood hazard (Schanze, 2007). Flood risk is an incorporation of the concepts of flood hazard and flood vulnerability (WBGU, 1998) and according to Kron (2005) flood risk can also be defined as the product of exposure (E) which refers to the population and economic assets affected by flooding, hazard (H) and vulnerability (V).

$$R=H*E*V$$

Flood risk analysis makes use of models based on physical processes and systems and of social responses and impacts (Schanze, 2007), in engineering projects, the use of probabilistic methodologies comes together with certain level of uncertainty so the estimation of risk analysis requires a high understanding of the fundamental probabilistic concepts by the professionals (Schanze, 2007). Additionally, flood risk analysis requires many sources of information, in which the source and quality of the data contributes to uncertainty.

Risk analysis is important given that can provide a plan for making decisions under uncertainty for future conditions over scenario analysis (Schanze, 2007). This research aims to characterize floods in rivers by using statistical tools, in which the relationship between several climate and geographical variables have been considered. Such relation is described by the method proposed by Paprotny & Morales-Nápoles (2017) based on Bayesian networks (BNs).

Bayesian Networks (BNs) is a graphical probabilistic model (Pearl, 1988; Kurowicka & Cooke, 2006) for dependent random variables from which is possible to describe the joint distribution of extreme discharges in rivers and variables of the geographical characteristics of a catchment (Paprotny & Morales-Napoles, 2017). River discharges are conditioned by climate and geographical characteristics.

1.2. Problem statement

Ecuador is a country located in South America with an interesting position within the globe; it is located in the Intertropical Convergence Zone (ITCZ) that consists of a “belt” of low pressures along the equator, therefore it is an area that is constantly threatened by hydro-meteorological conditions such as draughts, floods or the effects of “El Niño” phenomenon (FAO, 2010).

Ecuadorian cities that are located near banks of rivers or near the sea and that have low elevation respect to sea level are more exposed to floods (MIDUVI, 2015). This research will focus on the study of the Guayas river basin, which encloses the city of Guayaquil, the most populated city of the country with 2’578.201 inhabitants where 993.123 inhabitants live in flood prone areas (INEC, 2010) that can be translated as the 46.2% of the population of the city.

As most of developing countries, Ecuador does not have much climate information available and the existing measurements hold voids or errors. Additionally, there are places with no measurements at all or entire datasets are not reliable. This reduces the number of locations in which hydrologic and/or hydraulic studies can be performed.

This research aims to apply the Bayesian Network model proposed by Paprotny & Morales-Nápoles (2017) in the Guayas river basin, and to observe how this model performs in catchments like the ones in Ecuador by comparing it to the previous application in Europe. Other applications of the model (with their corresponding adjustments) were performed in the US and in Colombia, therefore this study can be used as a first step in order to develop a model for Ecuador.

1.3. Objectives

- Evaluation of BN method in the Guayas catchment in estimating extreme annual maxima river discharges.
- Compare the performance of the BN method applied in the Guayas Basin with the previous application of the BN model in Europe.

1.4. Research questions

- How does the BN model perform in a catchment in the Guayas Basin in estimating annual maxima river discharges?
- How does the model perform in small catchments?
- How does the model perform compared with other applications? What are the main differences and similarities between the studies?

1.5. Contribution of this study

Until now in Ecuador some studies have used BN models to predict draught events or to estimate ecological water quality, among few others. However, extreme river discharges have not been estimated using this tool, therefore the contribution of this research is the application of Bayesian Networks to estimate maximum discharges in the Guayas River Basin.

As mentioned in previous sections, climate information in Ecuador is limited, in this way the BN model and its application can be used as a guideline to similar studies in other regions of the country where the European model can be adjusted to such different conditions. Moreover, calibration of the model could be achieved by comparing the results of such adjustments with similar catchments in neighboring countries like the study case presented by Nasr (2017) in Colombia, this analysis can be used as the first step in developing a BN model for Ecuador and later on used as an example to start a model for South America.

Additionally, the results from the proposed BN model can be used to obtain the maximum discharge for different return periods and such information could be used as boundary conditions for projects involving hydraulic structures. After good calibration of the model is achieved, this method can be used to predict river discharges in areas of the country were no measures are available.

However, in this project, the adjustments and calibration of the European model to Ecuador conditions are not performed but the application of such model into the Guayas Basin will be carried out. In this way, the author shows the performance of the European model in Ecuador and will assess if adjustments are needed or not, in this way sets the first step into developing a model for Ecuador.

2. Literature review

2.1. Theoretical background

The purpose of this research is to apply a model based on Bayesian Networks to estimate the annual maxima river discharges at the Guayas River Basin, therefore a literature review was carried out in this topic.

In a Bayesian Network, random variables are represented as nodes in a graph and the dependencies are represented as arcs with a defined direction (Morales Nápoles et al., 2013), in that case directed acyclic graphs (DAGs) are used to represent the joint distribution of a number of variables. A DAG consists of a set of nodes that represent random variables (discrete and/or continuous) and a set of arcs such that no directed cycle is created; therefore, a certain order of variables could be established (Morales Nápoles et al., 2013). A Bayesian network consists of a directed acyclic graph and a set of conditional distributions.

The absence of arcs guarantees a set of conditional independence facts (Hanea, Morales Napoles, & Ababei, 2015), the direct predecessors of a node are called parents and the direct successors of a node are called children. A marginal distribution is specified for each node with no parents, and a conditional distribution is associated with each child node (Hanea, Morales Napoles, & Ababei, 2015).

The relatively simple visualization of the complicated relationships between the random variables is one of the most appealing features of a BN model (Hanea, Morales Napoles, & Ababei, 2015), they are commonly used to update distributions given observations. This is referred as inference in BNs.

Non-parametric Bayesian Networks (NPBNs) associate nodes with random variables for which no marginal distribution assumption is made, and arcs with one-parameter conditional copulae (Joe, 1997), parametrized by Spearman's rank correlations. The main result of NPBNs states that a particular choice of conditional copulae together with the one-dimensional marginal distributions and the conditional independence statements implied by the graph uniquely determine the joint distribution (Hanea, Kurowicka, & Cooke, 2006). Zero correlation entails the independent copula (Hanea, Morales Napoles, & Ababei, 2015).

To estimate the variable of interest the identification of the factors that contribute to its computation has to be defined. Additionally, previous knowledge of how these factors are related or influence each other is required in order to establish a cause and effect relationship that will finally result in the estimation of the annual maxima river discharges.

In this project the determination of the involved variables and its relationship is already set in the same way as in the work of Paprotny & Morales-Napoles, (2017) see Figure 1. The nodes are presented as histograms where the numbers indicate the mean and standard deviation of each variable, the arcs and the values on them represent the conditional rank correlation coefficients.

This model, initially constructed for Europe, will be applied in the Guayas Basin. However; the datasets for each variable will correspond to the case study basin.

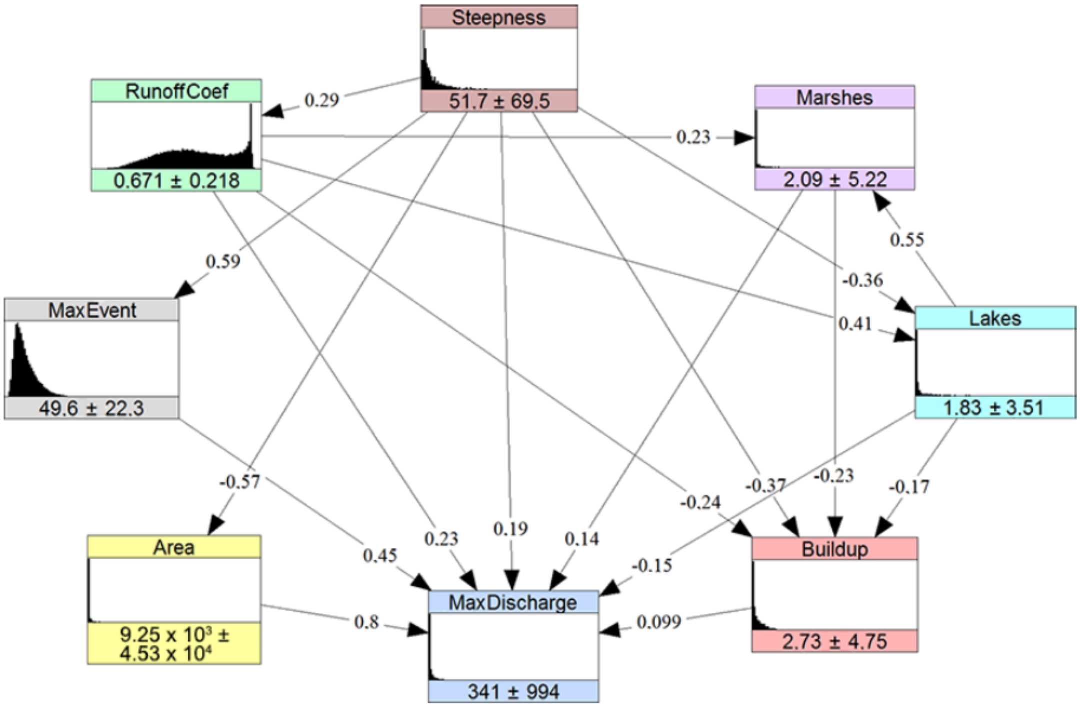


Figure 1 Bayesian network for river discharges in Europe.

3. Area of Study

This chapter presents a description of the area of study of this project.

3.1. Description of The Area of Study

3.1.1. Geographical position

Ecuador is American country located in South America. It limits in the north with Colombia and in the South and East with Peru, and to the West with the Pacific Ocean. The Andes mountain range divides the territory from North to South.

The country is classified in four different regions: Coast, The Galapagos Islands, the east Amazon basin and a central belt called “Sierra” which includes the Andes Mountain System. Ecuador is located in the Intertropical Convergence Zone (ITCZ) that consists of a “belt” of low pressures along the equator, therefore is an area that is constantly threatened by hydro-meteorological conditions such as draughts, floods or the effects of “El Niño” phenomenon (FAO, 2010).

Ecuador has an area of 283,561km2 (Oratlas, 2017) and is the 4th smallest country in the subcontinent. Is the 10th most populated country in America with approximately 16 million habitants and the most densely populated of South America.

The Guayas river basin is located in the coastal region of the country (*Figure 2*). To the north it limits with a spur of the Andes mountain range that extends to the west. To the east, the water divisor line of the western Andes mountain range.

This basin has a maximum height of 6310 m.a.s.l. and flows at sea level into the Pacific Ocean. Its geographic position extents between the parallels 00º14’ S, 02º27’ S and the meridians 78º36’ W, 80º36’W.

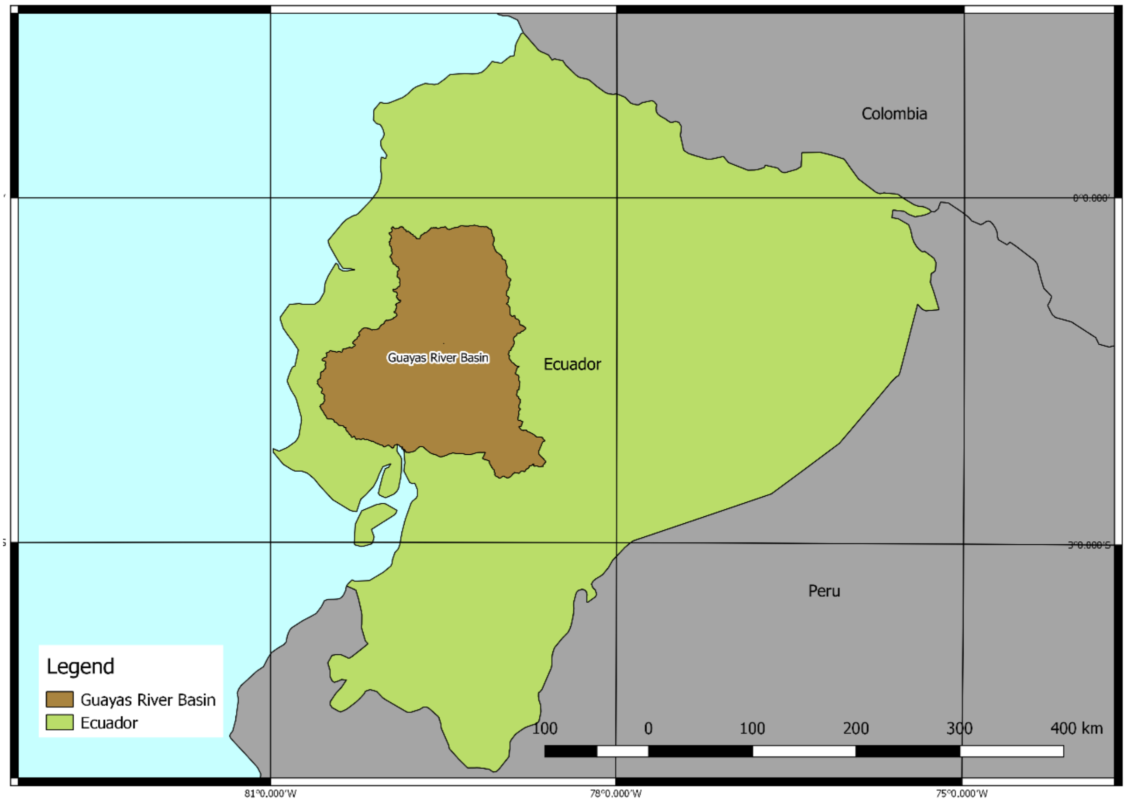


Figure 2 Guayas River Basin

The reason this catchment is chosen is due to its importance relative to the country's population and economy. With an area of 40000km2 approximately, the Guayas basin encloses the city of Guayaquil, which is the biggest city within the catchment and in the country, where 993.126 inhabitants are located in zones that are susceptible to flooding, corresponding to 46,2% of the population of the city (MIDUVI, 2015).

3.1.2. Physical, biological and social characteristics

One of the highlighted environments in Guayas river basin is the Gulf of Guayaquil which is the zone of more artisanal and industrial fishing production of Ecuador (Montaño Armijos & Sanfeliu Montolío, 2008).

The weather in this region is influenced by 3 factors that modify the weather in a seasonal way during the whole year:

- Atmospheric continental circulation identified by the trend winds of the SE.
- The Pacific Ocean as permanent generator of the humid air mases that summed with the effects of the maritime currents (Humboldt and El Niño) are the biggest regulators of the seasonal effects of the weather.
- The Andean foothills that with their height, relieve, and orientation channel the humid masses.

Due to this factors the littoral-coastal zone has a distinguished seasonality that causes an imbalance of the precipitation. From January to May is the wet season or rainfall season where flooding through large periods. From June to December is the dry season characterized by the lack of rain during the months of September and October.

The land use in this catchment is mainly agricultural (Highly technified intensive agricultural systems), the principal activities carried are bananas, rice, coffee, cacao, corn, African palm, tropical fruits as mango, oranges, melon, sugar cane etc.

4. Methodology

4.1. General Methodology

This chapter shows an overview of the methodology carried out in order to estimate the river discharges from the BN model. In the following chart, the main steps are described and later a brief explanation of each is given.

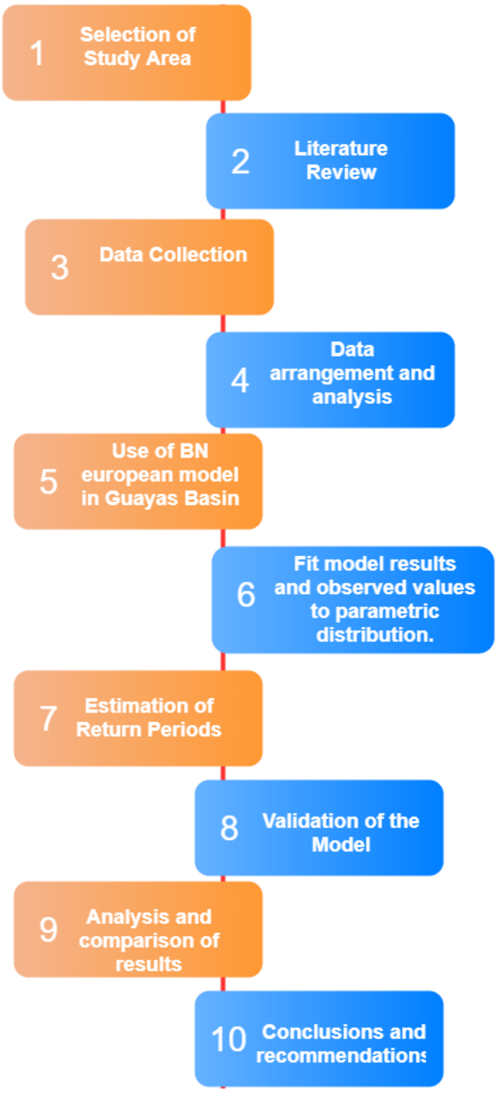

1 Selection of Study Area
2 Literature Review
3 Data Collection
4 Data arrangement and analysis
5 Use of BN european model in Guayas Basin
6 Fit model results and observed values to parametric distribution.
7 Estimation of Return Periods
8 Validation of the Model
9 Analysis and comparison of results
10 Conclusions and recommendations

Figure 3 Chart of General Methodology

A literature review is presented in chapter 2 in order to set the background knowledge for the methods applied in this project. Then, a description of the case study area is presented in Chapter 3.

Explanation about the methodology used for data collection is presented in section 4.2 and the results of the arrangement and analysis of the information is presented in chapter 5 and a description of the methodology used for the setting up the BN model is presented in section 4.3.

The results from the BN model and the measurements are fitted to parametric distributions following the methodology described in section 4.4, together with the procedure to obtain return periods. The validation of the model is presented in section 4.5.

Chapter 5 presents the analysis and comparison of results; the methodology to perform this task is presented in section 4.6**.** Finally, Chapter 6 contains conclusions and recommendations.

4.2. Methodology for Data Gathering and Processing

Before the model is ready to be used, different types of data are required. The reason for the selection of the variables and the sources from which the information was obtained was set in such way to maintain similarity with the studies of Paprotny & Morales-Nápoles (2017), Couasnon (2017) and Nasr (2017). In this section, a description of the procedure for obtaining the required data is presented.

The variables considered are depicted in the following table:

Name	Description	Units
Area	Area of the Catchment	Km^2
MaxEvent	Annual Maximum of daily precipitation and snowmelt	mm/day
RunoffCoeff	Maximum runoff coefficient of the catchment	-
Steepness	Catchment Steepness	m/km
Marsh	Area of the Catchment covered by marshes	%
Lake	Area of the Catchment covered by lakes	%
Builtup	Area of the Catchment covered by built-up	%
MaxDischarge	Annual Maximum of daily River Discharges	m^3/s

Table 1 Name and description of the variables

The variable MaxDischarge is assumed to be linked with all the other remaining variables from Table 1; by calculating these variables at catchment level is possible to predict MaxDicharge. Each variable represents a node in the BN (See Figure 1).

4.2.1. Measurements of River Discharge

River discharge measurements are very important for the model; validation of the model can be performed by making use of this information. The data for river discharge measurements was obtained from The Global Runoff Data Centre (GRDC) which is an international data Centre operating under the auspices of the World Meteorological Organization (WMO).

The locations of the chosen stations within the catchment are also of great important because they define the extent of the sub-catchments upon their position and with them the area and steepness of the catchment are defined.

The GRDC has nineteen stations in Ecuador from which only five belong to the Guayas Basin: Quevedo, La Capilla, Vinces-DCP, Catarama, Lechugal (See Figure 5). To have access to this information the GRDC request a form where the required stations have to be specified and some other documents.

The measurements are presented as daily time series accompanied by other characteristics proper of each station. It was found that each station considers different time periods and therefore the number of measurements varied from station to station.

River discharge measurements were also available from the national data bases from Ecuador, however, they were not used in order to keep similarity with the projects mentioned in the previous section.

As the boundaries for the stations were not provided, the catchments had to be delineated using computer software. The Guayas Basin was divided in five sub-basins using the stations available from GRDC. The delineation of the catchments was obtained through QGIS and GRASS by creating a flow accumulation raster map, a drainage direction raster map, and an outlet point that corresponds to the location of the stations. The areas resulting from this procedure are in accordance with the data provided by the GRDC.

Later, the datasets were analyzed and processed for each station and given that the data was measured daily, the maximum annual discharge (Q_{max}) could be computed. This was done using a Matlab script.

The Q_{max} values were used later to be compared with the results from the BN model and to validate its performance.

4.2.2. Geographical Information and Terrain

Geographical and terrain information is available from Ecuador's national databases like the Military Geographical Institute (IGM by its acronym in Spanish) in a scale of 1:50000. However, as with river measurements, other sources were considered to keep compatibility with the work of Paprotny and Morales (2017).

Terrain information was obtained from the USGS database, specifically through HydroSHEDS (Lehner, Verdin, & Jarvis, 2006), these products provides a DEM (Digital Elevation Model) from the Shuttle Radar Topography Mission (SRTM) void filled at 3 arc-second (~90m) spatial resolution.

A DEM is very important for this study because several variables of the BN are calculated from it, which are: catchment area, runoff coefficient and steepness (slope). The latter determine if a terrain is flat or hilly and it was calculated using the following equation:

$$Steepness = \frac{H_{\max} - H_{\min}}{\sqrt{A}}$$

Where:

- $H_{\max}$: Maximum elevation of the catchment [m].
- $H_{\min}$: Minimum elevation of the catchment [m].
- A: Area of the catchment [km^2].

Finally, a terrain classification (**land cover**) is needed, this variable is also computed using the DEM; three classes are considered in this study: Built-up, lakes and marshes. The information was obtained from The Global Lakes and Wetlands Database (GLWD) which has a combination of sources for lakes and wetlands on global scale (1:1 to 1:3 million resolution) from which the result is a database that focuses in three levels (Lehner B. a., 2004):

1. Large lakes and reservoirs (GLWD-1): Comprises polygons –shapefiles- large lakes with area larger or equal to 50 km^2 and reservoirs with storage capacity larger or equal to 0.5 km^3. Spatial resolution of 1:1 to 1:3 million, in geographic projection degrees' latitude and longitude.
2. Smaller water bodies (GLWD-2): Comprises polygons of permanent open water bodies with surface area larger or equal to 0.1 km^2 excluding the water bodies form level 1. Spatial resolution of 1:1 to 1:3 million, in geographic projection degrees' latitude and longitude.
3. Wetlands. (GLWD-3) comprises lakes reservoirs rivers and different wetland types in a form of raster map. Of 30 seconds resolution (~900m).

This information together with the sub-basin delineations allows to estimate the coverage per catchment relative to the area of the catchment expressed as percentage.

4.2.3. Climate Data

One of the most important variables for the BN model are defined by climate data. Those variables are precipitation, runoff and snowmelt. For this project only historical data was considered; the information was obtained from the WRCP CORDEX data base through the ESGF (Earth System Grid Federation) at the DKRZ website (German Climate Computing Center).

The data sets at the DKRZ are available through regions, for the case of the Guayas basin, the data corresponds to the South American region. The corresponding domain is SAM-CORDEX that uses the SMHI-RCA4 regional climate model with realization r12i1p1. A domain is the name assigned to each CORDEX regions (14 regions in total). The historical time line dates from 1950 until 2005 in daily time frequency.

This source was chosen because CORDEX information was also used by Paprotny & Morales-Napoles; Couasnon and Nasr (2017). The variables are represented by abbreviations,for precipitation (pr), snowmelt (snm) and for runoff (mrro). The information is downloaded as NCDF files and on a 0.44^{o} rotated grid (spatial resolution of 50km approximately).

One of the biggest challenges in this project was to obtain the information from the NetCDF files that later was extracted using Matlab scripts. The maximum per year per catchment was calculated for each variable and finally the snowmelt and precipitation were added together to form the MaxEvent node of the BN model. In the Guayas Basin there is almost no presence of snowmelt however, it was still included in this study to keep concordance with the other applications mentioned through this document.

As the data that first was extracted from the files corresponded to the entire South American domain the area was reduced just to the points that fell into the sub-basins of the Guayas River basin. To visualize the data, the grid and the points of each grid cell were plotted in QGIS, using Thiessen polygons.

Next, the maximum value per year per catchment for each variable was computed using Matlab scripts. From this data processing procedure, the MaxEvent and run off are stored. Finally, the runoff coefficient is calculated as follows:

RunoffCoeff=Maximum Runoff/MaxEvent.

4.2.4. Summary of Data Gathering

As shown in the previous sections, the information for the variables involved in the BN model come from different sources, all of them requiring different approaches for its visualization, gathering, analysis and processing. In the following table a summary of the variables and its corresponding source is shown:

Data	Node	Source	Timeline	Resolution
Annual Maximum Daily discharge per year per station	MaxDischarge	GRDC	1962-2005 (Varies per basin)	-
Area of the Catchments	Area Steepness	GRDC & Catchment Delineation		
Digital Elevation Model (DEM)	Area Steepness	SRTM	2008	3 arc second (~90m)
Land Use (Land cover)	Buildup Lakes Marshes	WWF		1:1 to 1:3million 30 arc second (~900m)
Precipitation Snowmelt Total Runoff	MaxEvent RunoffCoef	WRCP CORDEX		0.44o (~50km)

Table 2 Summary of sources for the variables of the BN model.

In section 4.2.1 was mentioned that each river discharge station had different number of measurements, additionally, the climate variables had a different amount of measurements from the river discharges. The following table provides an overview:

Station	River Discharge Measurements	Precipitation, Run Off and Snowmelt
Quevedo	1962-1977;1979-2005	1950-2005
La Capilla	1970-1980;1982; 1984-2005	1950-2005
Vinces-DCP	1964-1998; 2000-2005	1950-2005
Catarama	1982-1997; 2001	1950-2005
Lechugal	1974; 1976-1977; 1979-1997; 2002-2005	1950-2005

Table 3 Data available for the variables of the BN model

The following table shows a summary the operations performed in each variable.

Variable	Operation
MaxDischarge	None
Area	Area delineation (Software)
RunoffCoeff	Maximum(Total Runoff)/MaxEvent
MaxEvent	Maximum (Rainfall+Snowmelt)
Steepness	[Maximum(Elevation)-Minimum(Elevation)]/sqrt(Area)
Builtup	100*Area of Builtup/Area
Lakes	100*Area of Lake/Area
Marshes	100*Area of Marsh/Area

Table 4 Summary of the operation performed to obtain the variables in the BN model.

4.3. Methodology to Estimate Extreme River Discharges Through the BN Model

In this section the methodology to obtain the extreme rivers discharges using the European BN model is explained.

4.3.1. The BN Model

The BN model is composed by eight variables as shown in Table 1 and aims to estimate the annual maxima river discharges at five sub-basins of the Guayas Basin. The model used for this research is the European model presented by Paprotny & Morales-Napoles (2017) and carried out through a Matlab script which has not been altered in any way except the input data. In the following paragraphs, a brief description of how the algorithm works is presented. The results of this research are expected to show the performance of the model in the Guayas Basin.

The first step to construct the model is to define the DAG of the BN in which the parent-child relation is stablished for all the variables; following Figure 1:

- Pa(Steepness): None.
- Pa(Area): Steepness.
- Pa(MaxEvent): Steepness.
- Pa (Runoff Coefficient): Steepness.
- Pa(Lakes): Steepness, Runoff Coefficient.
- Pa(Marshes): Lakes, Runoff Coefficient.
- Pa(Built-up): Steepness, Runoff Coefficient, Marshes, Lakes.
- Pa(Discharge): Area, MaxEvent, Runoff Coefficient, Steepness, Lakes, Marshes, Built-up.

**Pa( ) refers to the “parent” variable.

It can be noticed that "Discharge" is child of all the 7 remaining variables, meaning that all of them influence in its estimation.

Then the conditional rank correlation coefficient for each arc is computed, this is accomplished by assuming a Gaussian copula (normal copula). The results vary between 0 and 1 and are presented as the correlation matrix between all the variables, however, another way to visualize the correlation between variables is trough scheme ball graphs as shown in the following figure.

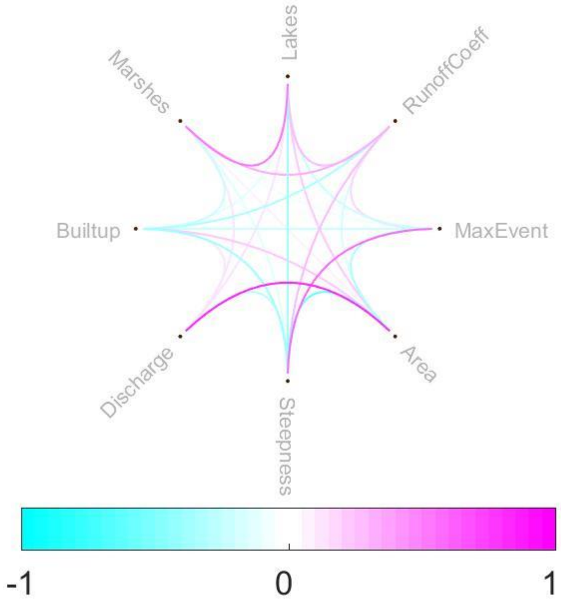


Figure 4 Scheme Ball for the Correlation Matrix of the European BN.

In a schema ball the variables are joined by straight lines or curves. The plot includes a color scale that indicates the correlation between the variables defined between -1 for negative correlation (cyan) and 1 for positive correlation (magenta), 0 means no correlation (white), so the correlation matrix of the variables determines the shading of the colors in the graph; a brighter color means more correlation between the variables (either positive or negative).

Next, the river discharges are estimated, this procedure is also known as inference. This mean that the estimation of discharge is based on information of the other variables (nodes), these nodes are conditioned (or forced) to certain values that in this case corresponds to the information gathered for the Guayas Basin. The aim is to infer values of max discharge using the quantification of the European model but at the same time is conditionalized to information of the Guayas Basin.

The maximum discharge sampling is performed from the joint normal distribution (copula) and then transformed to its original margins. The distribution of this sample can be described by the mean and standard deviation, from this values the uncertainty related to the maximum discharges is characterized.

After following these steps the performance of the BN model is evaluated by using the (available) measurements and the inferred discharges. To do this, performance indicators are used and are depicted in section 4.5.

4.4. Fitting of Data to Probability Distribution Functions

River discharge measurements can be described as a random process, in order to have a good description of the data is necessary to fit the data into a probability distribution (for the estimated data and the measurements) this is called distribution fitting.

The key of this procedure is to find the right distribution that will describe best the data, so the theoretical data is compared with the empirical data through the theoretical cumulative distribution function of several proposed probability distribution functions versus the cumulative distribution function (CDF) of the empirical data.

For this project, the following reasoning was considered for the distribution fitting:

Given the great variety of probability distribution functions, it is important to reduce the number of functions that will be used for comparison to finally achieve to one function that will fit the data. For this project the distributions considered were: The GEV, Gumbel, Weibull, Gamma, Lognormal.

Is important to remember that for this project the length of the data series of the discharge measurements is not equal to the length of the data series of the estimations and even tough this not affected the inference, it will influence the comparison of the data. To perform the validation, the timeline of the simulations has to be reduced to the length of the timeline of the measurements.

The procedure to compare the theoretical distributions to the empirical data followed some steps: First, the parameters of each of the previously mentioned distributions were computed using maximum likelihood estimation, then the cumulative distribution functions were plotted and finally a goodness of fit test was performed in order to attribute a theoretical distribution to the data,

Once a distribution has been selected then is possible to estimate the return periods of the maximum annual discharge for periods that exceed the timeline of the actual data, meaning that they were extrapolated to an arbitrary chosen value, for this project a return period of 100 years was chosen. Lastly, the cumulative distribution functions (corresponding to the fitted distributions) for the estimated data are computed and plotted.

4.5. Validation of the Model

In previous applications by Paprotny & Morales-Nápoles (2017), Couasnon (2017) and Nasr (2017) several measures were employed to assess the validation of the results (Pearson's coefficient of determination, Nash-Sutcliffe Efficiency, Standardized mean-square error, Mean Absolute Error and the Relative Error). For this project the first two are considered:

Pearson's coefficient of determination: (R^2)

$$R^2 = \frac{\sum_{i=1}^{n}\left(Q_i^{sim} - Q_{mean}^{obs}\right)^2}{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_{mean}^{obs}\right)^2} \qquad [0,1]$$

Where:

- Q_i^{sim} is the ith simulation value of the variable.
- Q_i^{obs} is the ith observation value of the variable.
- Q_{mean}^{obs} is the mean of the observations.

This indicator is used to measure the correlation between the observations and the simulated data, this can be explained as how many data points fall in the line formed by the regression equation. As the coefficient is higher, then higher is the percentage of points that falls within the line. So if the coefficient equals one, it means that all the data points falls in the regression line.

Nash-Sutcliffe efficiency (NSE)

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_i^{sim}\right)^2}{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_{mean}^{obs}\right)^2}\right] \qquad]-\infty,1]$$

NSE is used to measure if the model is biased. This indicator ranges between minus infinity to 1. If the NSE is equal to 1, it means that there's no bias between the measured and simulated values, if a negative number is obtained then it means that the measurements are better estimations than the simulations.

4.6. Methodology for Comparison of Results

The comparison of results will be carried out following the same methodology employed by Paprotny & Morales-Nápoles (2017). However, in this project some clear differences have to be highlighted.

Paprotny & Morales-Nápoles (2017) presented a catchment classification based on the area, in which a basin with an area less than 10000km^2 is considered small, but in this project all the catchments have areas smaller than such value, so a new classification will be considered using the following criteria:

Classification	Area Size [km2]
Small	0-3500
Medium	3500-7000
Big	7000-10000

Table 5 Classification of Watershed Areas.

Next, the values from the results of the model and the measurements are compared, this is done by applying the validation indicators presented in the previous section. From this results it will be determined if the model performance was positive or not.

Finally, a comparison of the model with other applications will be carried.

4.7. Inference

After validation of the model, the last step is to perform simulations in areas of the basin measurements are not available. In this project a total of 24 points were selected from the southern part of the basin which corresponds to the city of Guayaquil (see Figure 5), this location was preferred based on its importance of this city within the country and also to be able to provide values that could be used as boundary conditions for future studies related to flood hazard in the city. The next chapter contains the results of this procedure.

5. Results and Discussion

In this chapter the river discharges from the BN model are compared with the measurements from each station. The model is applied following the methodology presented in the previous chapter. A discussion of these results is also presented. Additionally, the results from data gathering and preparation are showed, then a comparison of the model performance for different catchment areas. Finally, the estimation of new points for inference around the city of Guayaquil is presented.

5.1. Results of Data Gathering and Processing

As mentioned in section 4.2, the information of each of the variables of the BN model (MaxEvent, Discharge, steepness, area, land cover-built-up, marshes, lakes, RunoffCoeff) were obtained from different sources and had different formats or units than the ones required by the model. In order to be able to use the information in the model, Matlab, Excel and Qgis tools were used. In this section, the results from this process are shown.

5.1.1. River Discharge Measurements:

As mentioned in section 4.2.1, river measurements were obtained from the GRDC. A total of 5 stations were considered for testing the model, together, this five stations do not cover completely the Guayas Basin, the remaining area is considered for inference in the next sections. The sub-basins are presented in the following figure:

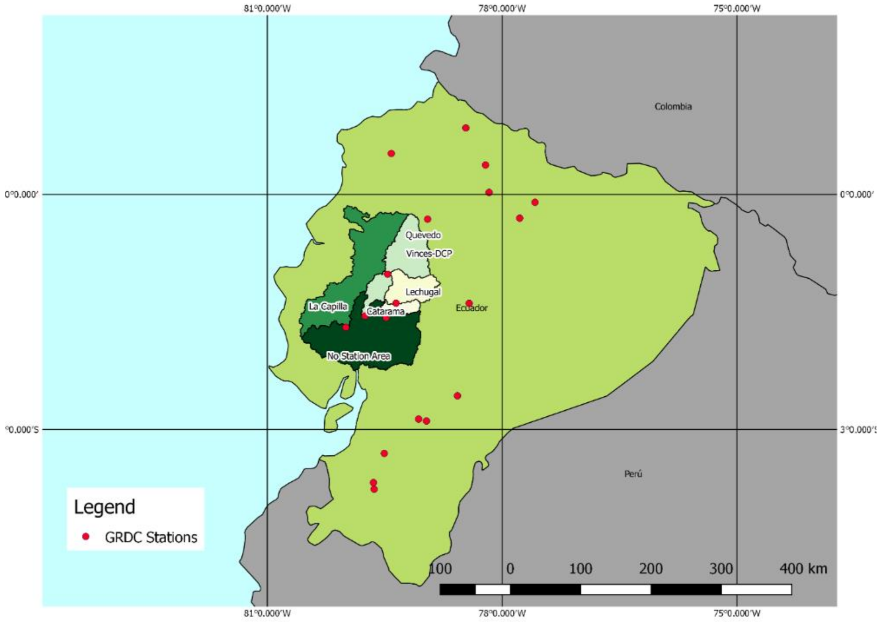


Figure 5 Configuration of the Guayas Sub-basins.

The red dots represent the GRDC discharge stations and the green-shaded areas are the sub basins of the Guayas river basin. The following table provides an overview of the data:

Name	Code #	River	Coordinates station		Catchment area (km2)	Units	Time series				# Years
			Latitude (dd)	Longitude (dd)			Initial		End		
							Month	Year	Month	Year	
Quevedo	3844300	Quevedo	1.0	79.5	3507	m^3/s	10	1962	12	2005	43
La Capilla	3844400	Daule	1.7	80.0	8690	m^3/s	12	1970	12	2005	34
Vinces-DCP	3844450	Vinces	1.6	79.8	4400	m^3/s	1	1964	12	2005	41
Catarama	3844460	Zapotal	1.6	79.5	3720	m^3/s	1	1982	12	2001	17
Lechugal	3844465	Zapotal	1.4	79.4	2980	m^3/s	1	1974	12	2005	27

Table 6 Summary of River Discharge information per station (GRDC).

5.1.2. Terrain Data:

The terrain of the Guayas basin has a maximum height of 4180 m.a.s.l and a minimum of zero, this information, together with catchment delineation were used to compute the area, steepness and percentage of land cover of the sub-basins (See Table 7).

Name	Code #	Lakes (%)	Mashes (%)	Build-up (%)	Steepness (m/km)
Quevedo	3844300	0	0	0.35	55.32
La Capilla	3844400	0	0	0.36	0.81
Vinces-DCP	3844450	0	0	0.60	48.06
Catarama	3844460	0.17	0	0.04	71.78
Lechugal	3844465	0	0	0.06	3.36

Table 7 Terrain data of the Guayas' sub-basins.

5.1.3. Climate Data

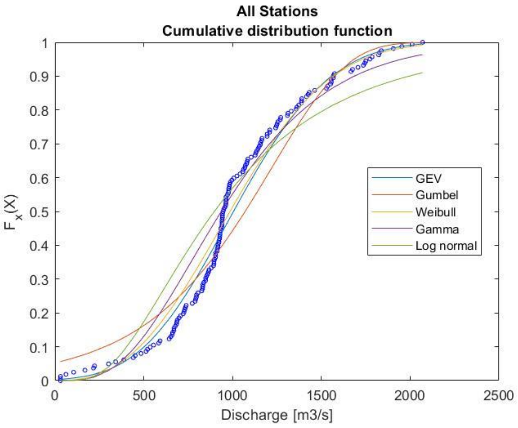
The procedure followed to obtain the data for MaxEvent (annual maxima of daily precipitation) and RunoffCoeff was explained in section 4.2.3. In the next table the range of the values for theses variables is shown.

Catchment	MaxEvent (mm/day)			RunoffCoeff (-)		
	Maximum	Minimum	Mean	Maximum	Minimum	Mean
Quevedo	116.35	79.14	93.39	1.01	0.94	0.99
La Capilla	85.86	42.76	65.93	0.95	0.85	0.90
Vinces-DCP	116.35	79.14	93.39	1.01	0.94	0.99
Catarama	145.97	95.51	113.44	1.01	0.99	1.00
Lechugal	145.97	95.51	113.44	1.01	0.99	1.00

Table 8 Statistics for Climate Data.

5.2. Results of The BN Model

As mentioned in section 4.3.1, the structure of the BN used in this project is the same as the European model presented by Paprotny & Morales-Napoles, (2017). In order to perform a proper assessment, the time scale of the simulations has to be the same as the time scale of the observations; from a total of 275 simulations, 162 were used.



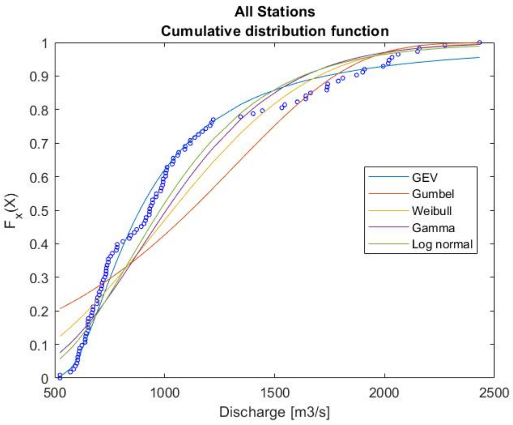


Figure 6 CDF for all the stations (Observed Data)

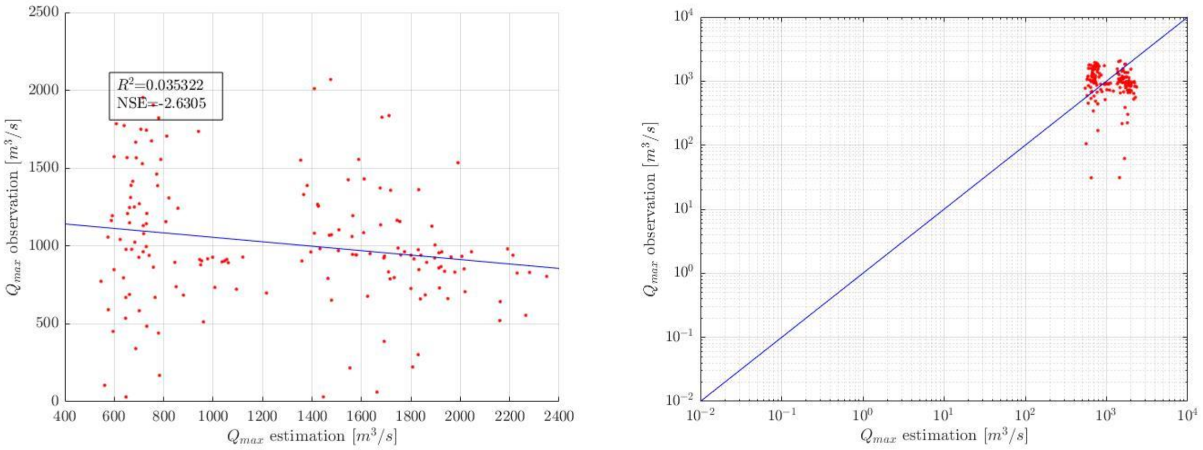
Figure 7 CDF for all the stations (Simulated Data)

The previous figures show the plot of the empirical cumulative distribution function (CDF) of the data of all the stations (observed and simulated) together with the theoretical CDFs. For both cases it was found that the data sets adjusted better to a GEV (Generalized Extreme Value) distribution function, furthermore the shape parameter was analyzed to determine if the GEV corresponded to a type I, II or III distribution (Gumbel, Fréchet and Weibull respectively). Both data sets fitted to a Weibull distribution.

This analysis was also carried for each station, given the different characteristics in each one of them, however, their fitting, as in the entire data, is subjected to uncertainty due to limited amount of information available, nevertheless all of them fitted to a Weibull distribution. The plots for each station are available at Appendix A.

Validation of the model

The validation of the model was carried for the data of all the stations as a whole and also for each station.



(1) Normal scale.

(2) Logarithmic scale

Figure 8 Plot of the observed vs simulated annual maximum discharge.

Figure 8 shows the result of the estimation of annual maxima discharge of all five stations plotted against the observed annual maxima discharge. Figure 8 (a) - Normal scale- also shows the values for the coefficient of determination R^2 and the NSE, the former is a number that is close to zero (0.035) which means that a small amount of points of the data falls in the regression, therefore the model represents a poor fit of the data, while the NSE coefficient shows a negative value (-2.63) which indicates a poor agreement between return periods computed from the model and observations.

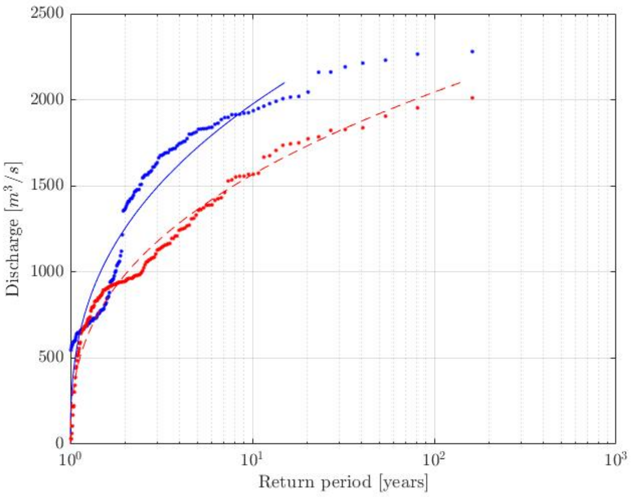
The slope of the regression line is another important factor to take into account in the analysis, looking at the trend of the line it can be said that estimation values increase significantly for small (decreasing) variations in the observations, given as result a negative slope, this means that at places where the discharges are the smallest, the model is predicting the highest values; the model is overestimating the simulations.

Several reasons can be the cause of such odd behavior, for example, overall, all the stations have limited data that cannot represent the basins accurately or the information can be doubtful or have errors. Such aspects could affect the quality of the results.

Return period

In this section an assessment on the model's prediction based on values for return periods will be carried out.

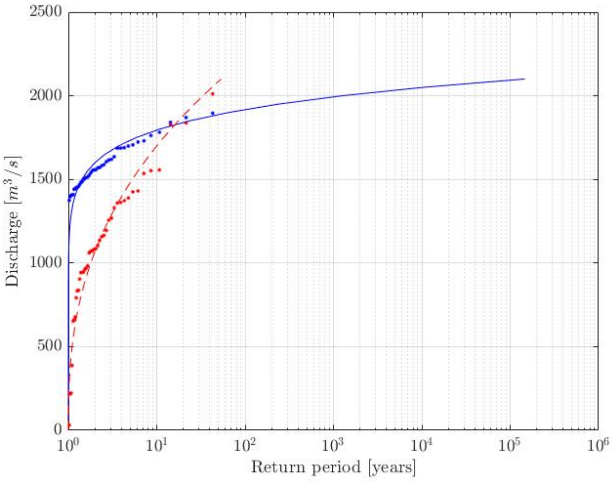
In order to obtain the discharge for a given return period, a frequency analysis is performed, the parametric distributed used for this procedure is the same as the one resulting from the fitting of the data, the Weibull distribution. The following graph shows a plot of the return periods for the data of all the stations:



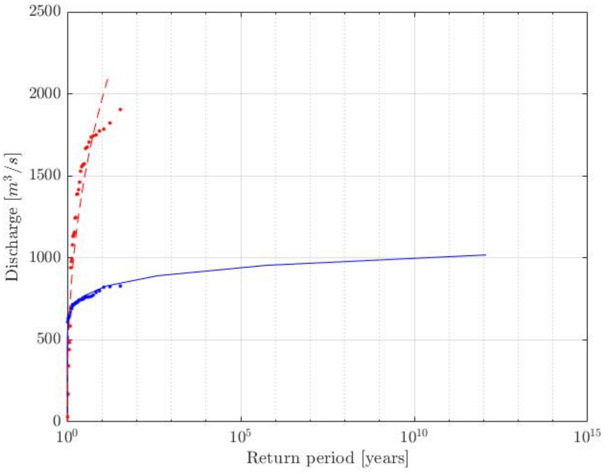
Blue dots: Estimated Annual Maxima Discharge, Red dots: Observed Annual Maximum Discharge Blue line: Weibull distribution from simulated data, Red Line: Weibull distribution from simulated data.

Figure 9 Plot of Return Period (years) versus Discharge (m^3/s) for data of all the stations

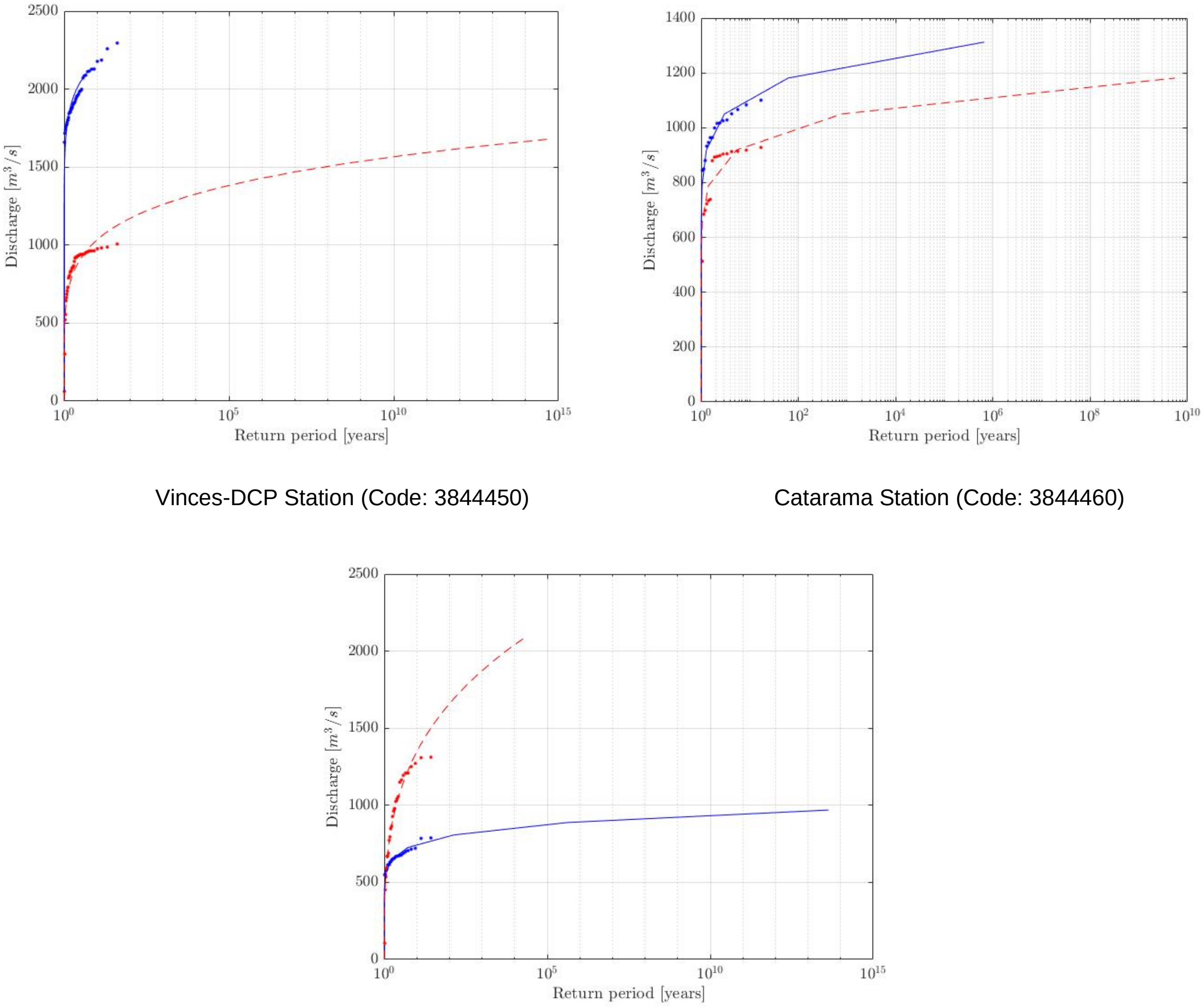
The same plot was built for each station considering a Weibull distribution:



Quevedo Station (Code 3844300)



La Capilla Station (Code: 3844400)



Vinces-DCP Station (Code: 3844450)

Catarama Station (Code: 3844460)

Lechugal Station (Code: 3844465)

Figure 10 Plot of Return Period (years) versus Discharge (m^3/s) for each station.

From the previous figure it can be concluded that the best fit correspond to the La Capilla Station (Code: 3844400), which is also the one of the stations with the longest dataset from all the stations. Additionally, this basin has a bigger area compared to the others, lying in the classification of big areas relative to the new classification presented in this project. However, it is still important to consider calibration of the model for catchments smaller than 10000km2.

Comparison with other applications:

Previously, the BN model was used for Europe and for the USA with positive results, as shown in Table 9. The results of calculating the same measures for the present study show that the model does not perform good in the Guayas Basin. One reason for this difference could be the size of the basin areas, as mentioned previously, the biggest area for the Guayas Basin is smaller than the smallest area considered for the European model. As presented by Paprotny & Morales-Napoles (2017) the area can be considered as the most important factor in the BN model, for which the smaller the area the less accurate the results, this can be confirmed by looking at Figure 4, where the curve between the discharge and the area is the line with the brightest magenta indicating a high positive correlation between those variables.

Indicator	Europe	U.S.A	Ecuador (Guayas River Basin)
R^2	0.92	0.858	0.035
NSE	0.92	0.757	-2.63

Table 9 Indicators for previous applications and the Guayas Basin.

Moreover, other factor is the amount of information considered for the applications, that for the case of Europe and USA the data was abundant and for the Guayas Basin, all of the stations have less than 50 years of measurements.

Number of simulations in the model

Additionally, the results from the BN were computed considering different sampling sizes, for this project 1000,5000,10000 and 50000 samples were considered. The following table shows the indicator values for each case:

Stations	R2				NSE			
	1000	5000	10000	50000	1000	5000	10000	50000
Quevedo	0.004	0.013	0.012	0.001	-1.337	-1.387	-1.380	-1.370
La Capilla	0.004	0.031	0.025	0.017	-1.128	-1.060	-1.083	-1.085
Vinces-DCP	0.002	0.002	0.005	0.001	-32.815	-33.477	-33.924	-33.770
Catarama	0.002	0.011	0.034	0.015	-2.716	-2.692	-2.572	-2.528
Lechugal	0.122	0.019	0.018	0.009	-0.756	-0.752	-0.076	-0.075

Table 10 Indicators value for different sample sizes at each station.

From the previous table is shown that for different sample sizes there is a variation between the values of the indicators, however, such difference is really small. Overall, the number of samples does not improve the results of the model.

5.2.1. Prediction of Annual Maximum Daily River Discharges

After the validation of the model, the next step is to predict annual maxima discharges in a basin with no measurements. In this project the model did not had a good performance in estimating the annual maxima discharges, however, inference in an ungauged basin will still be carried out.

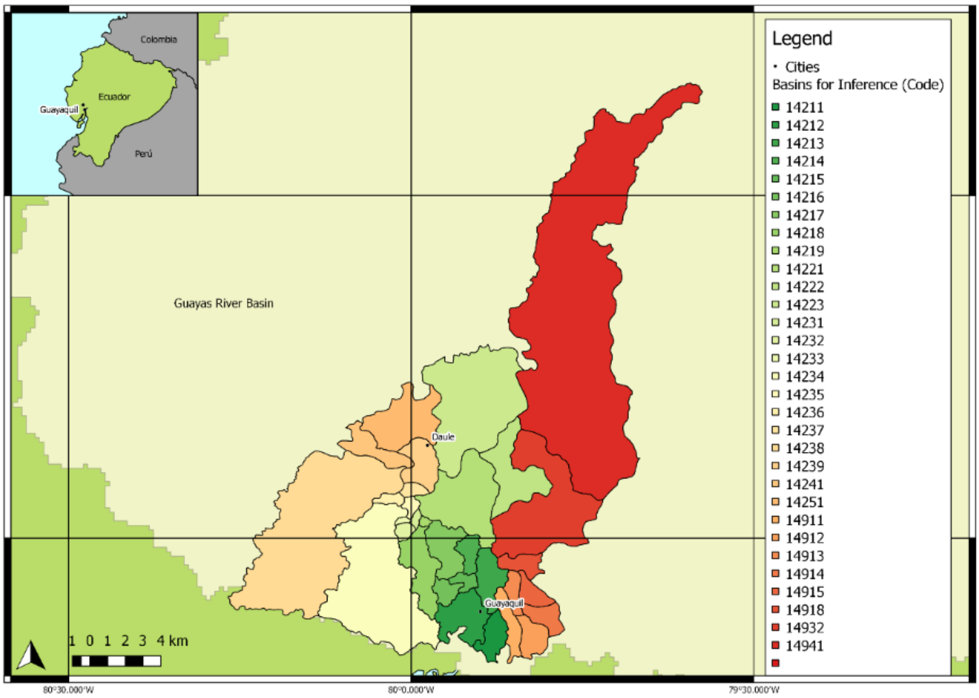


Figure 11 Basins considered for inference in an ungauged region.

The selected area consists of 24 small basins around the city of Guayaquil, these points were chosen given the importance of the city and that big part of its population is threatened by floods (Figure 11).

Inference in this region will help to set the first flood hazard maps in the region based in a BN model. Figure 12 shows the map for 100 year return period of the discharge per unit area.

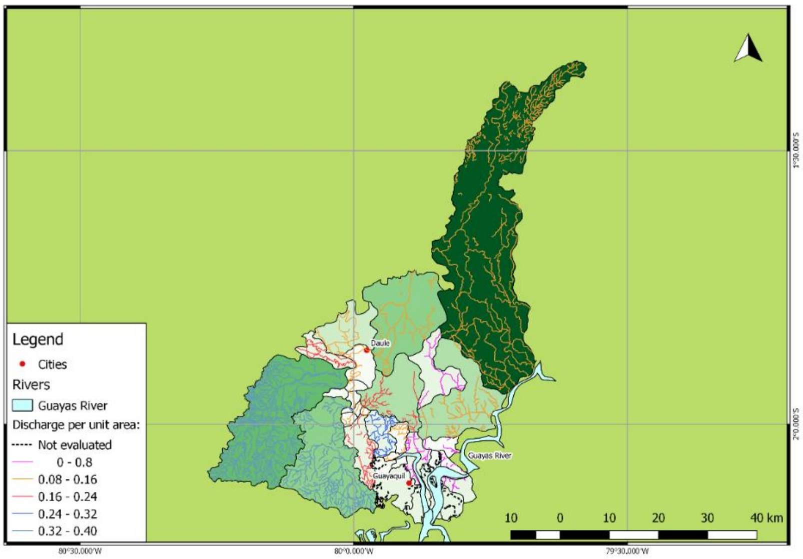


Figure 12 Discharge per river catchment area for the basins surrounding the city of Guayaquil.

From the previous figure it can be appreciated that around the city the basins have different values of discharge per km^2, ranging from 0-0.08 m^3/s per km^2 at the north, to much higher values ranging from 0.32-0.40 m^3/s per km^2 at the west, the basins at the west are also the basins with the flatter terrain of all which makes them more vulnerable under a flood scenario, therefore this type of maps can be used as a preliminary and rough assessment for flood risk in the region. However, the area of each basin (Figure 13) corresponds with the “small” category of the classification of the areas presented for this project, so the results can be greatly affected by this fact.

The catchment delineation was obtained from the National Information System (SNI from its acronym in Spanish) of the government of Ecuador and the rest of the terrain and climate data were obtained from the same sources presented in the methodology section.

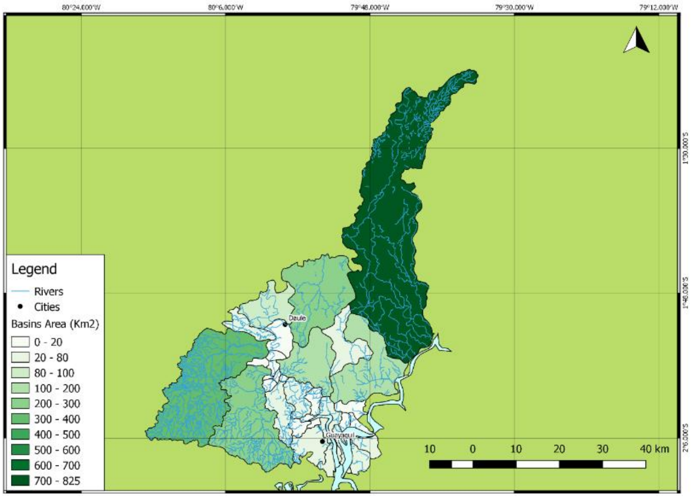


Figure 13 Area classification of the ungauged zone.

6. Conclusions and Recommendations

Conclusions

The BN method is a strong tool to estimate data for a variable of interest based on the knowledge of other variables. This model allows to visualize the data in a simple way while treating with complicated relationships between the random variables.

The model has proven to be effective in Europe and in the United States as shown in the previous chapter. For the case of the Guayas river basin the method shows low performance as for the entire data and in a station by station analysis. The validation over the entire data resulted in a R^2 of 0.035 and a NSE of -2.63. When the data was analyzed per station, it was also found that the R^2 and the NSE coefficient had similar results as the analysis of the entire data set, such values could be caused for the lack of data in the region (each station has less than 50 years of measurements), the influence of area size (less than the minimum suggested by other applications) in the model or perhaps poor quality of the data. In conclusion, the application of the model in the Guayas River basin resulted in low performance, the model overestimates the annual maxima discharge.

The area of the catchment is a decisive variable within the model which in turn depends on the quality of the terrain data. Catchment delineation depends on the quality of the digital elevation models; for this project national data bases were not considered to keep accordance with previous applications of the model, however this data bases could have better resolution DEMs and therefor improve the quality of the catchment delineation. The same can be said about information about built-up, lakes and marshes. Nevertheless, the maps resulting from this project are good example of how flood hazard maps help quantify flood risk in a region.

After testing the model for different sample sizes (1000,5000,10000,50000) there were no significant differences between the results. In conclusion, the increase of sampling does not improve the performance of the model.

Recommendations

As the area of a basin is the most influential variable in the model, it is recommended to use areas that resemble the size suggested by previous studies, however in Ecuador very few catchments have such size, therefore the model needs further calibration for small catchments.

Land data (percentage of built-up, lakes and marshes) is data that does not change over time within the model, however, in reality land cover is constantly changing and such changes should be included in the model to improve its performance. Additionally, other types of land cover (agriculture land, tropical forests, paramo, Andean forests among others) should be included for the case of Ecuador, that is a country in which the use of the land changes dramatically in relatively short distances. Over the next decades a great number of new hydroelectric projects are planned in the country which need to create reservoirs, this hydraulic structures affects the behavior of the rivers and moreover the discharges, such aspects should also be included in the model. National data bases are good sources for this type of information.

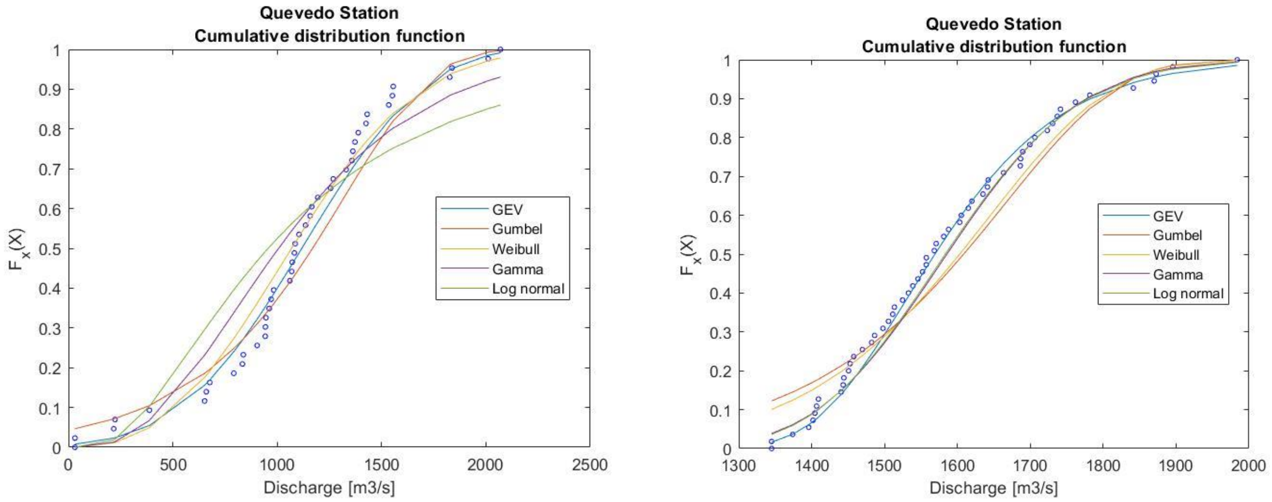
Ecuador is severely affected by el Niño phenomenon which is not included in the model, but the maps presented in this project will help to visualize the more vulnerable zones and therefore influence the ways these areas can be protected during such extreme events.

Fitting of the data resulted in a Weibull distribution by analyzing the shape coefficient of the GEV distribution, but given the big uncertainty surrounding the data, the GEV distribution should be considered for further studies in the region.

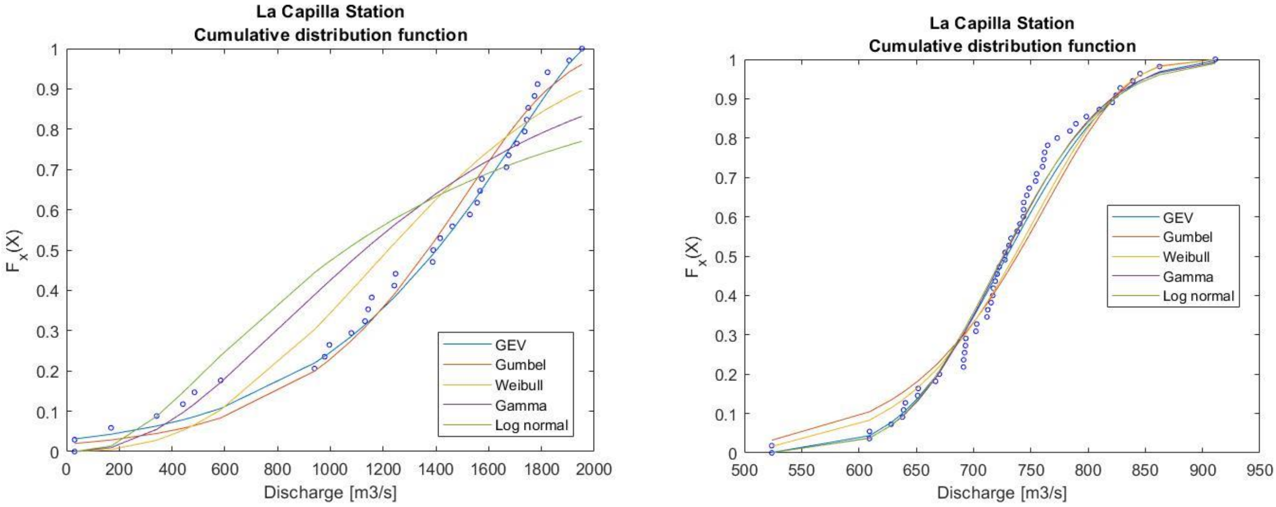
To improve the performance of the model, data from neighbor countries could be a good option, for example, the application in Colombian rivers presented by Nasr (2017) could be applied in Ecuador and the first step into creating a model for South America could set in motion.

Appendix A

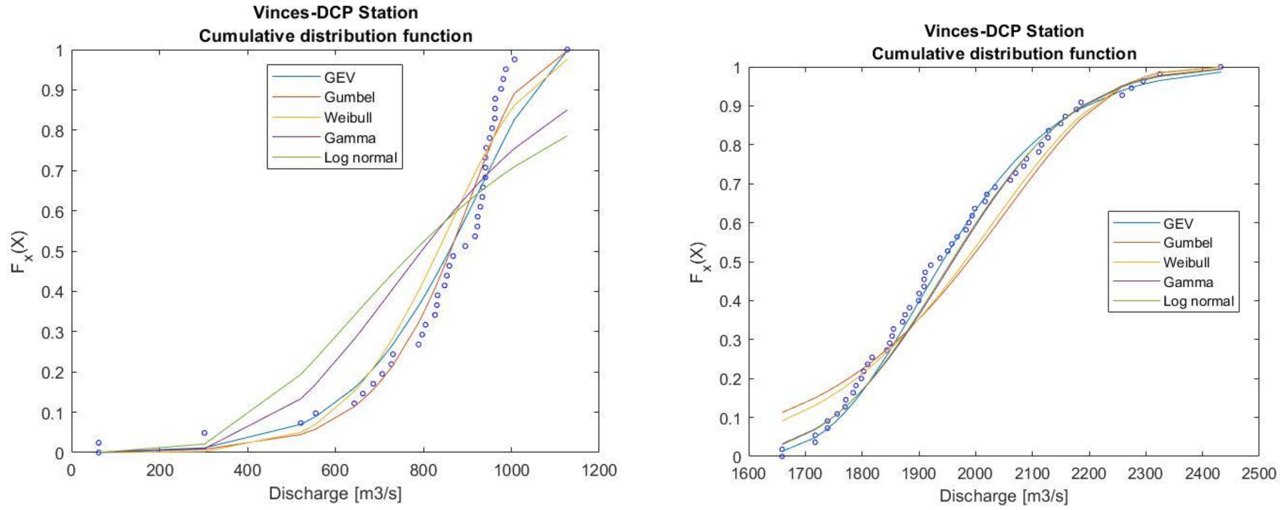
Fitting Distributions for River Discharge Data (Observations and Simulations)



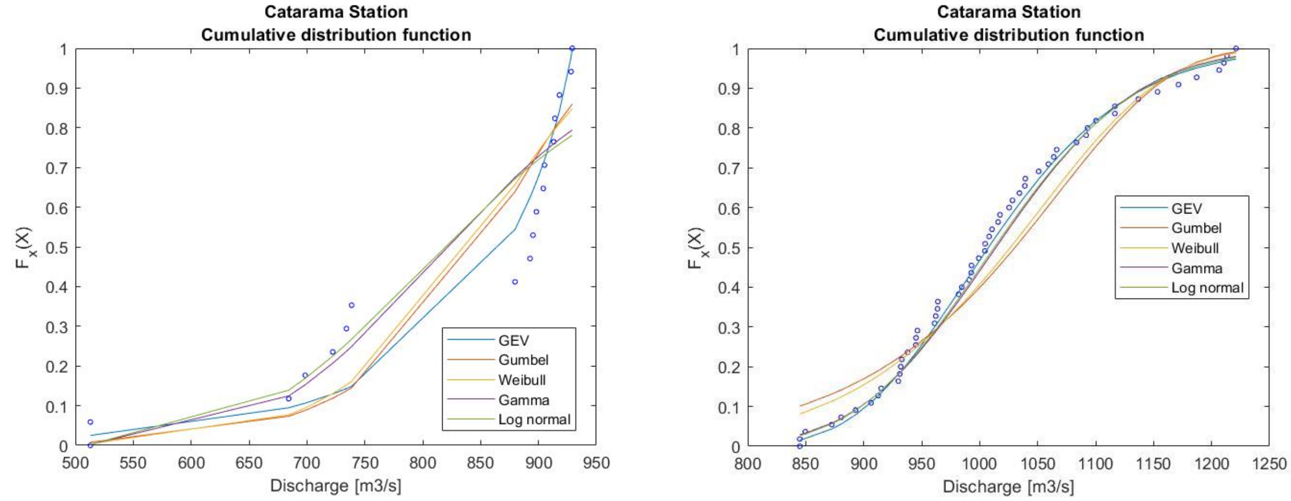
Quevedo Station- (Left) Observations (Right) Simulations.



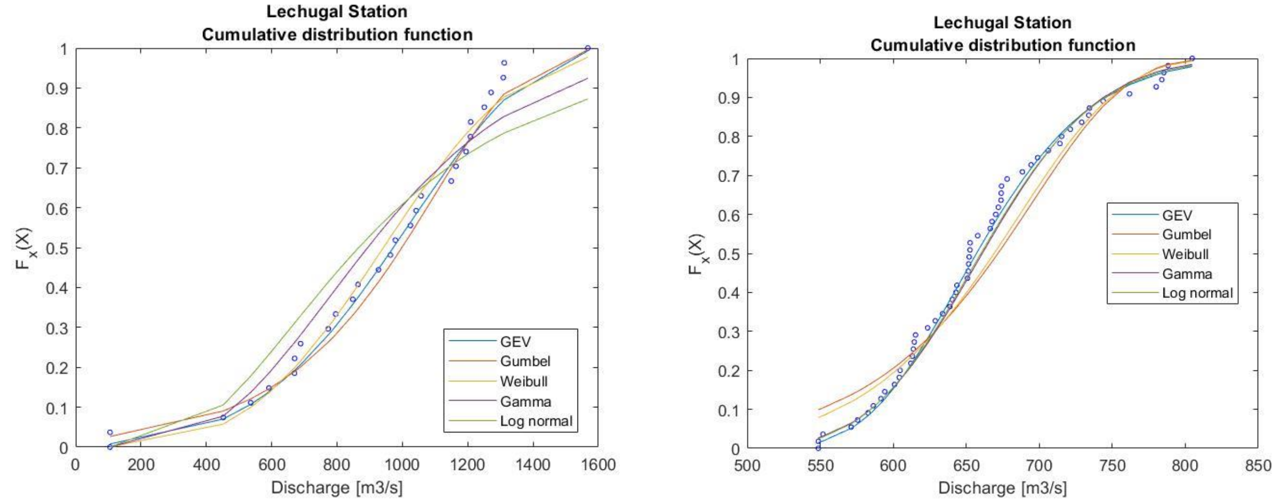
La Capilla Station- (Left) Observations (Right) Simulations.



Vinces-DCP Station- (Left) Observations (Right) Simulations.



Catarama Station- (Left) Observations (Right) Simulations.



Lechugal Station- (Left) Observations (Right) Simulations.

Bibliography

Couasnon, A. (2017). *Characterizing Flood Hazard at Two Spatial Scales with the Use of Stochastic Models-An Application to the Contiguous United States of America and the Houston Ship Channel.* Delft, South Holland, The Netherlands: Delft University of Technology.

FAO. (2010, December). *Organización de las Naciones Unidas para la Alimentación y la Agricultura.* Retrieved September 18, 2017, from http://www.fao.org/nr/tenure/infores/ltmdocs/es/

FLOODsite-Consortium. (2005). *Language of Flood Risk*. Retrieved from www.floodsite.net.

Genest, C., & Favre, A.-C. (2007). Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, 22.

Glossary of Excalibur Program. (2013). Delft, Zuid-Holland, The Netherland: Delft University of Technology.

Hanea, A., Kurowicka, D., & Cooke, R. (2006, August 24). Quality and Reliability Enginnering International. *Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets*.

Hanea, A., Morales Napoles, O., & Ababei, D. (2015, May 22). Reliability Engineering and System Safety. *Non-Parametric Bayesian networks: Improving theory and reviewing applications.*

INEC. (2010). *Resultados del Censo 2010 de población y vivienda en el Ecuador. Fasc{iculo Nacional.*

Joe, H. (1997). *Multivariate models and dependence concepts.* London: Chapman & Hall.

Kron, W. (2005). Flood Risk = Hazard • Values • Vulnerability. *Water International*, 58-68.

Kurowicka, D., & Cooke M., R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling.* Chichester, UK: John Wiley & Sons, Ltd.

Lehner, B. a. (2004). Development and Validation of a Global Database of Lakes, Reservoirs and Wetlands. *Journal of Hydrology*, 296/1-4: 1-22.

Lehner, B., Verdin, K., & Jarvis, A. (2006). *HydroSHEDS Technical Documentation.* World Wild life Fund US, Washington, DC. Retrieved from HydroSHEDS Technical Documentation.: https://hydrosheds.cr.usgs.gov/

Ministerio de Desarrollo Urbano y Vivienda del Gobierno de la Republica del Ecuador-MIDUVI. (2015). *Informe Nacional del Ecuador para la Tercera Conferencia de las Naciones Unidas sobre Vivienda y el Desarrollo Urbano Sostenible HABITAT III.* Quito.

Montaño Armijos, M., & Sanfeliu Montolío, T. (2008, Octubre). Ecosisteme Guayas (Ecuador), Medio Ambiente y Sostenibilidad. *Reviste Tecnológico ESPOL, 21*(1), 1-6.

Morales Napoles, O., & al, e. (2013, May 2). Reader for course: Introduction to Bayesian Networks. Delft, The Netherlands.

Munich Reinsurance Company. (1997). *Flooding and insurance.* Munich: Münchener Rückversicherungs-Gesellschaft.

Nasr, A. (2017). *Bayesian Networks and Data Driven Models for Estimating Extreme River Discharges. Case Study: Magdalena-Cauca Basin, Colombia.* Delft, The Netherlands: UNESCO-IHE Institute for Water Education.

Oratlas. (2017). *Libro Mundial de Hechos*. Retrieved Diciembre 15, 2017, from http://www.oratlas.com/libro-mundial/ecuador/geografia

Paprotny, D., & Morales-Napoles, O. (2017, June 2). Estimating extreme river discharges in Europe through a Bayesian Network. Delft, South Holland, The Netherlands.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plasuible Inference.* San Francisco: Morgan Kaufmann.

Sarewitz, D. P. (2003). Vulnerability and Risk: Some Thoughts from a Political and Policy Perspective. *Risk Analysis, 23*, 805–810.

Schanze, J. a. (2007). *Flood Risk Management: Hazards, Vulnerability and Mitigation Measures.* Springer Netherlands.

The Global Runoff Data Centre. (n.d.). 56068 Koblenz, Germany.

V. Choulakian, R. L. (1994). Crámer-von Mises Statistics for discrete distributions. *The Canadian Journal of Statistics, Vol. 22 No. 1*, 125-137.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Stamentent on p-Values: Context, Process and Purpose. *The American Statistian*, 5.

WBGU. (1998). *World in Transition-Strategies for Managing Global Environmental Risks.* Berling and others: Springer.

Wilby, R. L., & Keenan, R. (2012). Adapting to flood risk under climate change. *Progress in Physical Geography*, 348-378.