Master Thesis
TVVR 18/5003

Regional Regression Models of Mean Annual Streamflow and Design Flow in a Tropical Region of Colombia.

A Study Performed in the Departments of Norte de Santander and Santander.

Anisa Zigaf
Amanda Eriksson

Division of Water Resources Engineering
Department of Building and Environmental Technology
Lund University

Regional Regression Models of Mean Annual Streamflow and Design Flow in a Tropical Region of Colombia.

A Study Performed in the Departments of Norte de Santander and Santander.

By:
Anisa Zigaf
Amanda Eriksson

Master Thesis

Division of Water Resources Engineering
Department of Building & Environmental Technology

Lund University
Box 118
221 00 Lund, Sweden

Water Resources Engineering
TVVR-18/5003
ISSN 1101-9824

Lund 2018
www.tvrl.lth.se

Master Thesis
Division of Water Resources Engineering
Department of Building & Environmental Technology
Lund University

English title:	Regional Regression Models of Mean Annual Streamflow and Design Flow in a Tropical Region of Colombia.
Authors:	Anisa Zigaf Amanda Eriksson
Supervisor:	Magnus Persson, Water Resources Engineering, LTH Edgar Villareal González, Civil and Agricultural Engineering, UNAL (Colombia)
Examiner:	Linus Zhang, Water Resources Engineering, LTH
Language:	English
Year:	2018
Keywords:	mean annual streamflow; extreme flow; design flow; regional regression model; ungauged watershed; hydrological homogeneous region; Gumbel distribution

Acknowledgements

We are very grateful for the opportunity to travel to Bogotá, Colombia with the purpose to execute this master thesis. We would like to thank our two supervisors professor Magnus Persson and professor Edgar Villarreal González for making this possible. Edgar, we especially want to express our appreciation for all the help, guidance and kindness you have shown before, during and after our stay in Bogotá.

Thanks to the kind and loving people we met in Bogotá, it is thanks to you that our visit became interesting and fun outside of the office hours. Thank you very much for opening your homes to us and introducing us to not only your culture but also to your families.

It should not be forgotten that this is the work of two people and together we have grown and we have supported each-other during difficult times and decisions. We are both very happy and grateful for the company and the work-ethics of the other person which ultimately made this master thesis possible.

Last but not least we are truly grateful for all the support we have received from our families during the making of this thesis and of course during all the years we have studied to reach this point. All the years of studying would have been even more difficult and dull if it weren't for all the amazing friends we have gotten to know over the years. Together with you we have created remarkable memories and a second family.

Abstract

A responsible and sustainable water resources management is important. It is also crucial to take the presence of watercourses into account during design of hydraulic structures such as small dams and bridges. Doing this requires knowledge of estimates of annual streamflow and extreme flows to be used for design purposes. This study is divided into two parts, both focusing on the departments of Santander and Norte de Santander in Colombia where a tropical climate dominates. Both studies treat the creation of regression models using statistical tests to secure their reliability. The first study creates models for estimation of mean annual streamflow by relating streamflow to geomorphic and climate characteristics. Models are created for each individual department as well as for smaller, hydrologically homogeneous regions. The majority of the results show remarkably good fits with adjusted R^2-values ranging from 86.1-99.3 % and VIF-values in the range of 1.00-3.20. The second study creates regression models for estimation of design flows, with return periods of 10, 20, 50 and 100 years, using geomorphic variables. It is found that watershed area, elevation and slope result in the best fit giving adjusted R^2-values in the range of 72.8-75.7 % and a VIF-value equal to 1.536.

Abbreviations

A	Area
GCS	Geographical Coordinate System
GIS	Geographical Information System
GXHHRX	Group X Hydrological Homogeneous Region X
HHR	Hydrologically Homogeneous Region
IDEAM	The Colombian Institute of Hydrology, Meteorology and Environmental Studies
MAF	Mean Annual Flow
MinAmbiente	Ministry of Environment, Housing and Territorial Development
OLS	Ordinary Least Squares
P	Precipitation
PCA	Principle Component Analysis
Te	Temperature
VIF	Variance Inflation Factor

Contents

1 Introduction

One of the world's most important and vital natural resources is water. The access to clean water and sanitation was recognized as a human right by the United Nations General Assembly on the 28 of July 2010 (un.org 2014). It is therefore important that water resources are managed in a way that allows for this right to be fulfilled by everyone, everywhere.

Water has a wider range of use than for drinking and sanitation; for example irrigation, hydropower, navigation and recreation. This use has to go hand in hand with watershed management. In order to obtain a responsible management, estimates of annual watershed runoff volumes are required (Vogel 1999). To enhance the management, information about extreme flows for design purposes is valuable and can be used in the construction process of hydraulic structures such as smaller dams and bridges and consequently diminish economical and social costs (Dingman & Palaia 1999). Runoff information can be collected through gauges placed in watersheds, although not all watersheds count with a gauge that can monitor the streamflow and according to Barbarossa et al. (2017) monitoring has declined since the 1980s. This calls for alternative methods to gain information about the annual watershed runoff volume and the design flows.

One alternative method that has been recognized is regional regression models (Barbarossa et al. 2017; Vogel et al. 1999). These models relate mean annual streamflow with geomorphic and climatic variables and have been widely used in the western and northern parts of the world where, in general, a similar climate regime can be found. The question is, however, if regional regression models would show as good results for other types of climate, specifically tropical ones.

This study has its focus on Colombia which has a total population reaching 50 million people (Nationalencyklopedin n.d.). The country has, since the 1960s suffered from a civil war which has implied deadly violence, sexual assaults and threats in the whole country. Due to this more than 5 million people have been displaced within the borders of the nation (Nationalencyklopedin n.d.). The urban population has thus increased making Colombia one of the most urbanized nations in Latin America (del Ama 2013) and the need for a more sustainable water management is needed, both in urban and rural areas. Therefore the alternative method of regional regression models has been analyzed in this study where the area of focus are two departments in Colombia, Norte de Santander and Santander, which can be acknowledged as tropical regions.

1.1 Previous Studies

According to statistics about water distribution around the world, water availability should not be an additional concern for Colombia because it is ranked as one of the countries in the world with most access to water per person and year. This however excludes that the country experiences problems with contamination, inadequate infrastructure and unreliable weather which damages and diminishes the amount of useful water (Stratfor 2016). According to an article in the Colombian newspaper El Espectador in 2014, Bogotá could suffer a water crisis within ten years if infrastructure is not improved and the city's inhabitants do not change their water consumption patterns (Gónima 2014). Due to these factors a sustainable water resources management is of greater importance than ever.

1.1.1 Previous Studies for Estimation of Mean Annual Streamflow

As mentioned above, monitoring of streamflow has decreased worldwide. Focusing on Colombia, the country has a history of information collection from various locations in a stream, although this has changed from the 1950s until now and direct monitoring has diminished and streamflow-registration is therefore far from abundant as well as poorly distributed in the country (Salazar 2016).

Seeing as how this is not particular for Colombia, other possibilities to estimate the streamflow in ungauged river basins have been developed throughout the years. Perhaps the most common being the use of runoff maps where runoff is estimated in relation to the watershed area (Vogel 1999). However in its simplest form streamflow can be predicted through the use of the continuity equation (Dooge 1992).

$$\mu_Q = \mu_P - \mu_E \tag{1}$$

where μ_Q, μ_P and μ_E represent mean streamflow, precipitation and evapotranspiration respectively.

Other, more complex options are the use of so called conceptual rainfall-runoff models such as the HBV model or a physically based model such as MIKE SHE. However both of these models require a relatively large amount of input data which has a high cost, both in time and money. Yet another possibility is to use empirical models which usually imply mathematical equations based on time series of input and output data (Gayathri 2015).

This study focuses on regional regression models which could be classified as statistically based empirical models. Advantages with regional regression models are that they are easily applicable and need a small amount of input data in comparison with conceptual models (Mahmoud and Paramar 2006). Other advantages presented by Vogel et al. (1999) are that they produce an objective equation that can easily be integrated and implemented in water resources management as well as in computer software such as geographic information systems. The accuracy and uncertainty of water yield can be assessed as well as the influence of climate. Finally, and the biggest advantage according to Vogel et al. (1999) is that the models can give information about both mean and variance of streamflow. This implies a wide range of information and guidance for planning and managing water resources.

Several studies exist where regression analysis has been executed and considerable results have been obtained. Vogel et al. (1999) performed a study for the whole United States of America and Barbarossa et al. attempted to formulate a global regression model in 2017. Vogel et al. formulated first a model for the entire country and then one where the country was divided into 18 regions according to the location of natural drainage divides. Both climatic and geomorphic variables were included and the article states that climate has an immense influence on the hydrological cycle, therefore climatic variables were obtained with care in order for them to be as accurate as possible. Other than performing different models for different areas, Vogel et al. present three types of model (1) one model with only area to explain the flow, (2) one with area, mean monthly precipitation and mean monthly temperature and (3) one with all basin characteristics in order to find the most suitable variables to explain streamflow. For model (3) the variables that were included were different for the 18 individual regions but overall the majority included area, precipitation and temperature, although for some regions the mean annual precipitation was included while for others the mean monthly precipitation for a specific month was included. Regarding the temperature, the maximum, minimum or mean for a specific month was included. The maximum amount of variables used was five. The study of Vogel et al. showed that the models for the smaller regions performed better along with using the most suitable variables, this gave a goodness of fit value, R^2, between 90.2% to 99.8%.

One could ask why a regional regression model should be performed when a global one already exists. Barbarossa et al. (2017) executed their analysis using 1885 catchments globally and unequally distributed. The majority of data was collected from North America and Europe. South America is mostly represented with data from Brazil and it can be assumed that some were collected from tropical areas. Again, variables included in the model were climatic; precipitation and temperature, and geomorphic; area, slope and altitude. Barbarossa et al. made two important statements that were considered in this analysis, (1) that mean annual precipitation represents the potential runoff of the catchment and (2) that 90% of the streamflow can be explained by the five variables with area as the biggest influencer.

Both studies mentioned above conclude that multiple regression improves the prediction of streamflow in ungauged areas.

A third study, performed in the USA, by Mohamoud and Parmar (2006) found that area, precipitation and temperature gave the best regression models. Different from the other studies Mohamoud and Paramar divided the area into hydrologically homogeneous regions as they stated that this would provide a model that has an enhanced predictive power. How this is incorporated into this study is explained in section 2.3.3.

The study performed in this thesis will be based on the work presented by Vogel et al. (1999) therefore the results from this study will be compared with the findings from Vogel et al. and hence the table presented below displays the overall results from their study.

Table 1: Overall findings from the study in USA by Vogel et al. 1999.

Model type	**R^2-range**	**VIF*-range**
Model 1 (A)	27.3 - 99.1	-
Model 2 (A, P, T**)	38.9 - 98.9	1.0 - 2.7
Model 3 (most suitable variables***)	90.2 - 99.8	1.1 - 7.4

*VIF = Variation Inflation Factor which is a statistical variable, it is explained in section 3.1.4 below.
**A = Area, P = Precipitaion, T = temperature
*** Most suitable variables in order to obtain the best fit for the model.

As mentioned before, the study area will be limited to the two departments of Norte de Santander and Santander They will be divided according to three different manners which will be analyzed and compared and therefore so called regional regression models are to be created. The first model will include all the basins that lie within the geographic limitations of Norte de Santander and Santander. The second one will have the same division although one model will be created for each department. Lastly, the third type will consist of hydrological homogeneous groups containing several watersheds, one model will be created for each group.

Additionally, the three types of models based on different area division will undergo three types of analysis where different variables are included. The most simple one includes only area, the second one includes area, mean annual precipitation and mean annual temperature. The third model includes both climatic and geomorphic variables in order to find the best possible prediction.

1.1.2 Previous Studies for Estimation of Design Flows

The most common method for estimation of floods at ungauged areas is through regression models of design flows with the use of drainage basin characteristics for explaining and prediction of extreme flow (Dingman and Palaia 1999).

Dingman and Palaia (1999) performed an investigation regarding the suitability of different variables to include in a regression model in order to explain extreme flows with a frequency of 10, 20, 50 and 100 years, also called design flows. They examined drainage basins characteristics as well as channel geometry as they claim that the last variable can produce a model with a predictability as well as or even preferable to a model based on drainage basin characteristics. The variables that explained drainage basins were area and elevation and for channel geometry only channel width was used. Their results exhibit in fact that a regression model including channel width performed better than the one with drainage basins. Nonetheless, the characteristics of the drainage basins are easier to obtain than information about channel geometry because it can often be recovered using GIS which implies that it is more economical as well as time efficient.

In this thesis regression models have been executed in Norte de Santander and Santander for drainage basin characteristics only. The analysis is based on the study of Dingman and Palaia therefore the characteristics that were included in the model are area and elevation. Further slope will be added to the analysis as it was also mentioned as a candidate in the previous study. In this analysis only one model was performed for the whole area size. Therefore it was not regional although four different model types according to input data were executed. One with only area, the second one with area and mean slope, the third one with are and mean elevation and finally the fourth one was a combination; area, mean slope and mean elevation.

1.1.3 A Previous Study in the Area of Norte de Santander and Santander

A previous study has been made by Salazar (2016) in Norte de Santander and Santader where flow duration curves were estimated in ungauged catchment areas using a method of regionalization. In this study Salazar assembled geographical, physiographic and climatic information about the watersheds in Norte de Santander and Santader in order to explain the flow duration curves in the watersheds in the two departments. Mean daily streamflow was provided by the IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales/The Colombian Institute of Hydrology, Meteorology and Environmental Studies) along with information about hydrological measurement stations in the area of study. These stations allowed for a delimitation of the watersheds in the departments which comprise the foundation of the study. Further he performed a tendency test to acquire information about any change in the hydrological cycle that could be due to anthropogenic impacts. He also grouped the watershed according to hydrological homogeneity utilizing statistical analyzing methods.

The study presented in this thesis is based on the findings of Salazar due to the time scope of this thesis and for the reason that both studies are executed in the same area. Because of the time consumption it would take to order information about streamflow from IDEAM it was considered more efficient to use the information that had been obtained by Salazar. The information used from his study is the one mentioned above and it is explained more in detail in section 2.3 below.

Regarding water resources management it can be said that this study is complementary to the one executed by Salazar. Information about mean annual flow (MAF) can be used for management at early stages whereas flow duration curves can be used for more precise analyses.

1.2 Objectives and Limitations

The aim of this thesis can be divided into two parts; the first is to obtain regression models that explain MAF for the study area using both hydroclimatological and geomorphological variables. This is based on the previous studies of Vogel et al. (1999) and Salazar (2016) which is stated in the paragraphs above. The second aim is to obtain a regression model that explains design flow using drainage basin characteristics. This is based on the previous study of Dingman and Palaia (1999) mentioned in the section above. The problem formulations behind the main objectives are:

1. Is it possible to formulate a reliable regression model that relates MAF to geomorphological and/or climatological characteristics of the watersheds of Norte de Santander and Santander in Colombia?
    (a) Would such a model perform better for larger areas i.e. the area of the individual departments, or would smaller areas, that are grouped according to their hydrological homogeneity perform better?

    (b) Which climatological and basin characteristics should be included in the regression equation to obtain the best possible fit of the model?

    (c) What criteria should be used to define the best possible fit of the model and thus if a model can be considered as reliable?
2. Is it possible to formulate a reliable regression model that relates design flows to drainage basin characteristics for the regions of Santander and Norte de Santander in Colombia?
    (a) Which basin characteristics should be included in the regression equation to obtain the best possible fit of the model?

    (b) What criteria should be used to define the best possible fit of the model and thus if a model can be considered as reliable?

The study has been limited to the Colombian departments of Norte de Santander and Santander, mainly due to the previous study by Salazar (2016) that was executed in these areas. This can be explained by the extension of this project, both when it comes to time limitations and access to data. The variables tested in the model will be chosen both due to their assumed influence on the flow as well as the simplicity to obtain information about them. This will hopefully facilitate a duplicate of this study in other areas.

2 Background

2.1 An Overview of the Colombian Water Resources Management

On a national level the institution responsible for the framework of laws and regulations surrounding water management in Colombia is the Ministerio de Ambiente y Desarrollo Sostenible (Ministry of Environment, Housing and Territorial Development - MinAmbiente). However it is the responsibility of each department to make sure that these laws are implemented. The regulations exist in order to assure the following objectives (Diez Diaz 2017):

- Assure that there is a sufficient supply of fresh water for the country by preserving the aquatic ecosystems and the hydrological cycle.
- Characterization, quantification and optimization of the country's demand for fresh water.
- Secure the quality of fresh water and to minimize the pollution of water bodies.
- Assure an integral water resources management and the risks associated with the supply and demand.
- Strengthen the institutions responsible for water resources management and optimize their working conditions.
- Fortify the governmental institutions responsible for water resources management.

IDEAM answers under MinAmbiente and is responsible for the collection and processing of meteorological- and hydrological data. The institution therefore holds the responsibility for the measurement stations spread out through Colombia.

The stations provide daily measurements of data such as temperature, precipitation and streamflow. These stations use measurement instruments of varying age and quality. Historically most instruments required physical readings from a person who visited the stations several times a day. However, today most instruments record the measurement values digitally. Due to the varying quality the measurements are not always reliable. Either the measurement instruments can fail to register a value for a certain day (or other time step) or the values can differ greatly from other values close in time due to natural differences.

An example of a measurement station that IDEAM uses for rainfall is a pluviometric station which can be viewed in figure 1 below.

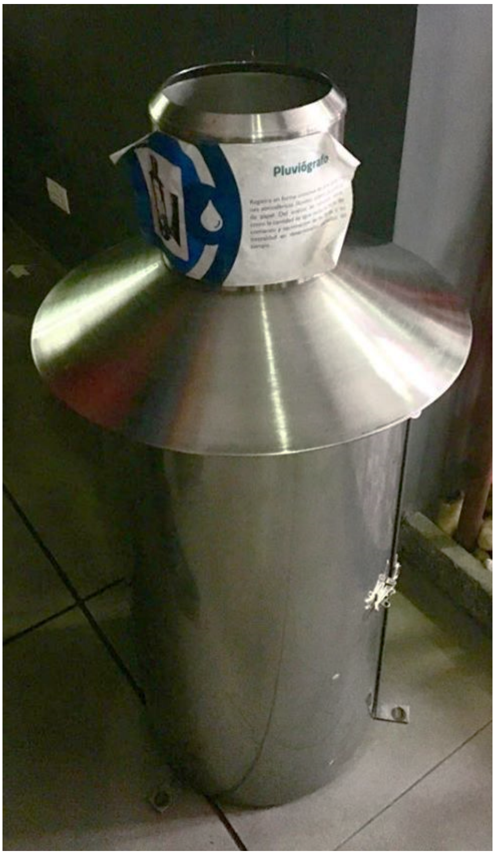


Figure 1: A pluviometric measurement station that IDEAM uses in order to gather information about rainfall.

Two phenomena that occur in Colombia and that can have great affect on the climate are El Niño - Southern Oscillation (ENSO) and La Niña. These are climate variations that are classified as natural phenomena. ENSO results in decreasing rainfall and an increase in temperature whereas La Niña is the opposite; rainfall increases and temperature drops. The phenomena are recurrent and take place with a certain number of years of frequency. Normally the ocean temperature is used as an indicator of when the phenomena are approaching (Siac n.d.; IDEAM n.d.) and the result can be seen from measurement stations.

IDEAM also works with regionalization of the country according to different methods. One such regionalization was performed in 2014 according to the precipitation regime of different parts of the country using the Principal Components Analysis (PCA) which will be mentioned further on in the text. This resulted in 17 homogeneous regions (Gúzman et al. 2014). These regions could for example be used in studies as the one described in this report.

2.2 Study Area description

Colombia is located in the Nothern part of South America with coasts both toward the Pacific and Carribean Ocean. It is known for its tropical climate as the equator passes through the southern parts of the country. However, it offers a varied climate as the high Andes run through the west, the large rivers Amazon and Orinoco leave marshlands in the east and south-east and long beaches can be found along both the Pacific and the Caribbean coast. This results in large differences in topography, temperature and precipitation throughout the country (Nationalencyklopedin n.d.).

As mentioned, this thesis will focus on two of the 32 departments of Colombia; Norte de Santander and Santander (Nationalencyklopedin 2017). They are located in the North-East part of the country bordering Venezuela. Figure 2 illustrates the area of study.

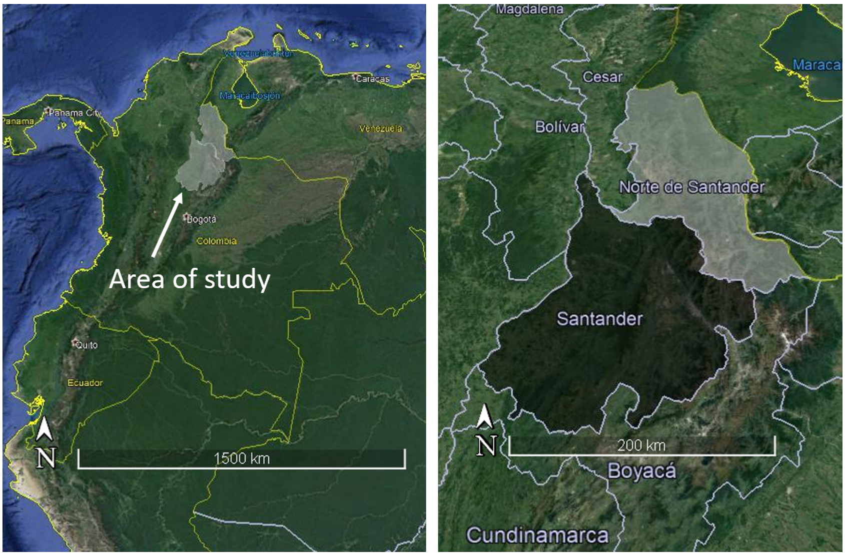


Figure 2: To the left, the location of the area of study in Colombia. To the right a zoomed-in picture of the two departments Norte de Santander and Santander (Google Earth Pro (2017), modified by the authors).

2.2.1 Norte de Santander

Norte de Santander has a common border with Santander in the west and south and with Venezuela in the north and east, see figure 2. The geography and main climate of the department can be divided into three different regions; the northern parts are characterized as wet and humid due to a rich hydrological system as a well as some remaining jungle that contains a lot of ravines. Average temperature in this region is 24°C and precipitation reaches 2,500 mm/year. In the east, high altitudes dominate reaching 3,300 m.a.s.l. and it is characterized by an altering climate from temperate to cold with temperatures lower than 12°C. The third area, south-west, is dry to very dry with temperatures between 18-24°C. As for precipitation the central parts of Norte de Santander has a small precipitation of 1000 mm/year whereas in the south the precipitation has an average of 3000 mm/year (Gobernacion de Norte de Santander 2017).

This department accounts for a rich hydrological system with three main river basins; Río Catatumbo in the North, Río Magdalena in the West and in the South-West Río Orinoco (Gobernacion de Norte de Santander 2017).

2.2.2 Santander

The water network in the department of Santander consists of bigger rivers such as the river Magdalena and Carare as well as smaller rivers, creeks and swamps. The climate is not homogeneous as the department is distinguished by two main physiographic sections. In the west lies the valley of Magdalena which is characterized by its heavy vegetation and low altitude. In this region the climate can be described by high temperatures, with an average of 29°C and heavy rain, which can reach 3,800 mm yearly. From the north-east to the south-east the Eastern Cordillera stretches and has, unlike the valley of Magdalena, a high variety in altitude with high slopes reaching 3000 m.a.s.l. In the Eastern Cordillera a dry environment dominates. In general the precipitation in this area is less intense varying from 500 mm to 2,000 mm yearly. Regarding the temperature it can reach up to 32°C in some parts of the region and lower than 7°C in others (Gobernación de Santander 2017).

2.2.3 Precipitation Regime

Amounts of precipitation for each department have been mentioned in the previous two sections. However, the yearly precipitation pattern is similar for both departments, showing a bimodal behavior with precipitation amounts reaching maximum values in March, April, May, September, October and November as can be seen in figure 3 (Gúzman et al. 2014). This pattern will be of special interest when deciding which climatic variables should be used in the regression models. It is assumed that the months with maximum amount of precipitation will have a greater influence on the mean annual streamflow than other months.

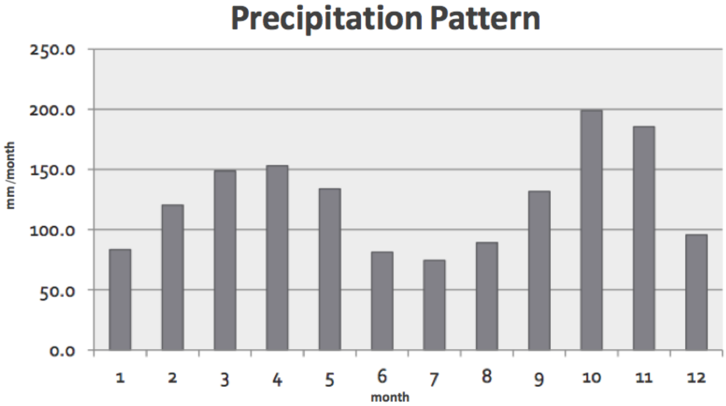


Figure 3: Bimodal precipitation pattern of the study area with peak amounts in March-April-May and September-October-November (Gúzman et al. 2014, modified by the authors).

2.3 Available Data

In this section the methods used in previous studies for obtaining data and the variables used in this thesis will be displayed. This mainly includes the methods executed by Salazar (2016).

2.3.1 Mean Annual Streamflow Data

In order to create regression models, long records of data are required. The World Meteorological Organization recommends that climatic analysis be executed with data for 30 years (Barbarossa et al. 2017). In this case that applies for annual streamflow, as this is the variable that will be analyzed, although for this analysis the requirement will be diminished to 20 years due to lack of data availability. Hydrological information can be requested from IDEAM which has a storage of data from hydrological and meteorological stations.

As mentioned above Salazar requested mean daily streamflow, this contained mean monthly streamflow as well as monthly absolute maximum values. He continued to select hydrological measurement stations which fulfilled the following criteria: (I) a minimum of 20 years of historical data, (II) a minimum of 90 % coverage per year, (III) a time range between 1981 - 2010. The time range was chosen between 1981 and 2010 because these years accounted for the most complete data. Within these years, the 20 years of flow measurements did not need to be continuous. The conditions were set by Salazar in his study and are followed in this study.

36 of the hydrological stations were identified as useful due to their fulfillment of the requirements. Another three stations did not account for sufficient flow record lengths although they were considered acceptable to be included in the analysis due to the limitation of data available. The chosen 39 stations are listed in table 2 below with their number, name and number of years of data. Observe that some of the stations are marked in bold, this will be explained below.

Table 2: Hydroloical stations included in the analysis with their number, name, department and number of years of data (Salazar 2016, p.34, modified by the author).

Number	Station code	Station name	Department	Nr of years with sufficient data
1	16017020	LA DONJUANA	N.STDER	30
2	16027060	PTO LEON	N. STDER	21
3	16027100	CORNEJO	N. STDER	27
4	**16037010**	**CAMPO TRES**	**N. STDER**	**24**
5	16037020	CAMPO DOS	N. STDER	29
6	16037030	PTE SARDINATA	N. STDER	29
7	16037040	CAMPO SEIS	N. STDER	19
8	16037050	PTE SAN MIGUEL	N. STDER	26
9	**16047010**	**PTE ABREGO**	**N. STDER**	**30**
10	**16057010**	**LAS VEGAS**	**N. STDER**	**24**
11	16057030	LA CABAÑA	N. STDER	23
12	**16057040**	**QUINCE LETRAS**	**N. STDER**	**26**
13	16067010	PTO BARCO-GABARRA	N. STDER	22
14	23127020	PTO ARAUJO	STDER	28
15	23127060	STA ROSA	STDER	29
16	23147020	PTE FERROCARRIL	STDER	28
17	**23197130**	**PTE SARDINAS**	**STDER**	**30**
18	**23197270**	**PTE PANEGA**	**STDER**	**17**
19	23197290	CAFE MADRID	STDER	28
20	23197370	SAN RAFAEL	STDER	28
21	**23197430**	**EL CONQUISTADOR**	**STDER**	**24**
22	24017570	SAN BENITO	STDER	26
23	24017580	JUSTO PASTOR GOMEZ	STDER	25
24	**24017590**	**PTE NACIONAL**	**STDER**	**27**
25	**24017640**	**LA CEIBA**	**STDER**	**18**
26	24027010	SAN GIL	STDER	27
27	24027030	NEMIZAQUE	STDER	30
28	24027040	PTE CABRA	STDER	30
29	24027050	PTE LLANO	STDER	27
30	24027060	PTE ARCO	STDER	29
31	24027070	MERIDA	STDER	27
32	24037370	MOMPA IZQUIERDO	STDER	29
33	24037390	CAPITANEJO	STDER	28
34	24047020	REMOLINO	STDER	24
35	24067010	EL TABLAZOS	STDER	20
36	24067030	PTE LA PAZ	STDER	26
37	37017040	PTE LOPEZ	N. STDER	26
38	37017050	VENAGA	N. STDER	29
39	37027010	PENA DE LOS MICOS	N. STDER	22

Additional to the importance of long record lengths the significance of the conservation of the hydrological conditions is emphasized (Salazar 2016; Vogel 1999). In other words the preservation of the land-use and infrastructure in the watersheds. As can be seen in table 2 nine stations are marked in bold which are the stations that show tendencies for change of hydrological conditions during the time span 1981-2010. Salazar has investigated this applying the Mann-Kendall test which is used to perform the null hypothesis. If the null hypothesis cannot be proven a hydrological linear trend can be assumed which indicates a change in the watershed management. The nine stations were still a part of the analysis although one regression model was performed for all measurement stations and another excluding the nine stations marked in bold in table 2.

2.3.2 Catchment Characteristics

Once the 39 hydrological stations were established, the delimitation of the watersheds was executed by Salazar, one for each station. This was performed with digital elevation models, DEM, over the departments Norte de Santander and Santander and with information about the river distribution in the departments. The resulting 39 watershed basins are displayed together with the stations in figure 4 below.

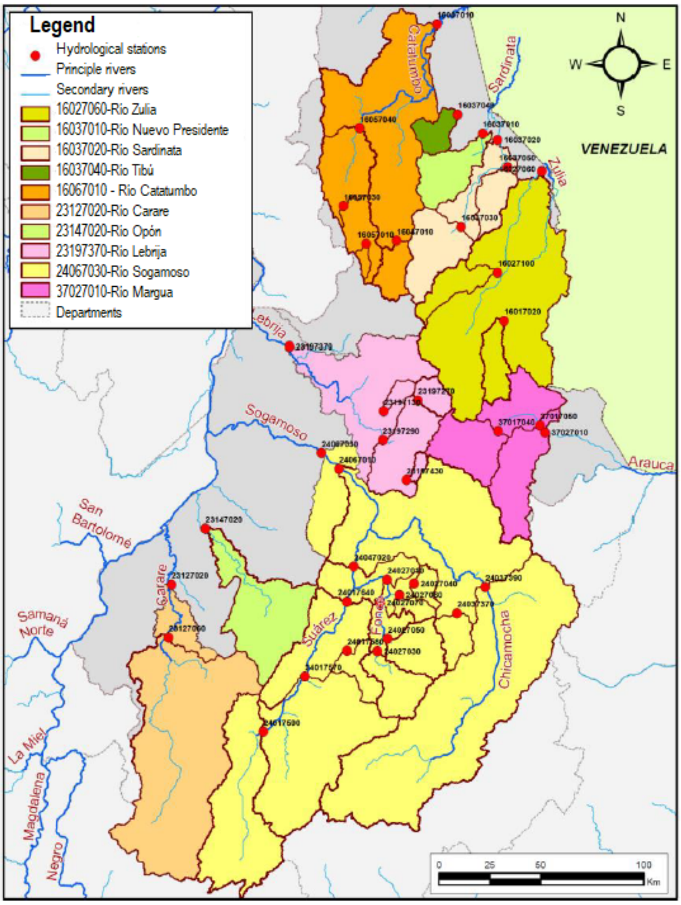


Figure 4: The 39 basins in the area of study (Salazar 2016, p.38, modified by the authors.)

The figure has been divided into ten main river basins which are distinguished by different colors. The ten main river basins contain smaller watersheds. The basin of Río Sogamoso, marked in yellow, is the largest area containing 15 of the 39 watershed basins. Two of the 15 basins, number 35 and 36 (see table 2), have a significantly larger area than the other basins, further they contain the 13 remaining basins. A first try on performing a model was done including these two catchments which showed that they could be considered as outliers according to the residual scatterplots. Therefore catchments number 35 and 36 were excluded from this study. Since the majority of the area of these two catchments was covered by the 13 smaller ones it was assumed that the loss of information by removing the catchments would be insignificant. The two concepts, outliers and residual scatterplots, are explained in section 3.1.4 further below.

It can also be observed in the figure that the watersheds do not cover the entire area of Norte de Santander and Santander which are the areas colored with a darker shade of gray. The reason for this is that basins, do not follow politically decided boundaries, in fact the south-east part of the Río Sogamoso basin is located outside of the department of Santander and belongs instead to the department of Boyacá. This is not optimal when creating models for the individual departments and might be misleading as not the whole department is included.

Other catchment characteristics that were obtained from the study of Salazar are the mean basin slope and mean basin elevation. They are chosen to be a part of the analysis as they may substitute the influence of other factors such as radiation, wind and vegetation that are more complicated to measure (Barbarossa et al. 2017).

This study tries to focus on variables that are easily accessible on a nationwide level for Colombia. The variables tested are listed among the other variables used in this analysis in table 4 below.

2.3.3 Hydrologically Homogeneous Regions

Salazar (2016) mentions regionalization as a globally recognized method to obtain hydrological information when observations are not available. This information can be used to identify hydrologically homogeneous regions (HHR). In this study models were performed for HHR in the study area and these regions were decided by Salazar. His method is briefly explained in this section.

A homogeneous region is a group of several regions that display similar climatological, physical or hydrological characteristics. When the hydrological pattern for two or more drainage basins resemble each other they are said to belong to the same HHR. An advantage with regression models that has not yet been mentioned is the possibility to incorporate different watersheds into one model that do not necessarily belong to the same watercourse, if they are classified as hydrologically homogeneous (Salazar 2016). According to Mohamoud and Parmar (2006) regression models for MAF show better predictions when the basins are divided into HHR.

Even though regionalization is frequently used, a common method for deciding which variables to incorporate in the regionalization analysis does not exist (Salazar 2016). Two regular methods for dividing the areas into HHR were tested; PCA and stepwise regression analysis. The PCA is a method that aims to diminish the number of explanatory variables by selecting variables that are linearly independent as well as a linear combination of the eliminated variables. It is executed by calculating the correlation (see section 3.1.4) and the variance between the variables. The variables are selected so that the variance decreases with 20 %. The step is repeated until a satisfactory amount of variables is obtained. The method of stepwise regression is a multiple linear regression method and it is based on how well an equation performs if a variable is added or removed. For each addition of variable the coefficient of determination, R^2 (see section 3.1.4) is examined. If a better fit can be observed, i.e. a higher number, according to the coefficient the remaining explanatory variables in the the equation are examined. Each variable is removed one by one and the R^2-value is inspected, if a lower value is obtained by the removal the specific variable is added again. This procedure is repeated for every variable added into the equation.

Once the most influential variables have been established the basins were divided into groups which was done in two different ways; k-means clustering and Andrews curves. These two methods will not be discussed in detail as it is not considered to be of importance for the understanding of this study, (for further reading, see Salazar 2016) however the basic differences will be mentioned. K-means clustering is an iterative method where the number of groups is determined at the beginning of the analysis. Andrews curves is a graphical method where one curve is drawn for each basin and the curves with least distance and difference will be grouped together.

By using the four methods mentioned above eight groups with different combinations of variables were created. Four groups were formed from all the 39 watersheds and the other four from the 30 watersheds that did not show any tendencies of having changed with time. In each group several HHR exist with multiple watersheds. In total 24 HHR were formed and the requirement for an HHR to be included in the study was that at least eight watersheds were represented. Twelve HHR fulfilled this requirement and they are listed in table 3 along with the watersheds in each group. The watersheds are numbered according to table 2. Observe the groups in bold which will be explained below.

Table 3: Hydrologically homogeneous regions divided into groups.

Groups	**Watersheds (nr)**											
G1HHR1	**11**	**12**	**23**	**26**	**27**	**28**	**31**					
G2HHR1	**2**	**4**	**5**	**7**	**8**	**14**	**15**	**16**	**20**			
G2HHR3	**9**	11	**23**	**26**	**27**	**28**	**30**	**31**				
G3HHR1	2	5	7	8	13	14	15	16	20			
G4HHR2	26	27	29	30	31	32	37	38	39			
G5HHR1	**1**	**9**	**10**	**11**	**12**	**19**	**21**	**28**				
G5HHR2	**3**	**4**	**5**	**7**	**8**	**16**	**17**	**29**				
G5HHR4	**2**	**13**	**14**	**15**	**20**	**22**	**25**	**33**	**34**			
G6HHR2	**3**	**4**	**5**	**6**	**8**	**16**	**23**	**27**	**29**	**30**	**31**	**32**
G7HHR1	2	13	14	15	20	22	33	34				
G7HHR2	3	5	6	16	23	27	28	31	32			
G8HHR1	3	5	6	8	23	26	29	30	31			

In table 3 some groups are marked in bold and others not. This depends on the number of watersheds that were included in the grouping. For the ones marked bold all the water basins were included i.e. also the ones which showed tendencies. The other groups are compiled only by 28 watersheds, see table 2. Four basins did not fit into any group; 18, 24, 35 and 36 (basins 35 and 36 were already excluded).

2.3.4 Extraction of Climate Characteristics

In order to create the regression models many different types of data were collected. Most climatological data was accessed through the Geovisor (IDEAM 2017a) and the Atlas Climatológico de Colombia (IDEAM 2017b) of IDEAM. These are free geographic information systems open to the public. Through these services maps can be downloaded as shape files to be used in geographical information system (GIS) softwares (such as ArcGIS in this case) or as map images. It is possible to download maps containing information of almost all climatological and meteorological data. Maps containing data for precipitation and temperature were downloaded from the Geovisor. Map images containing data on number of days of precipitation per month and maximum rainfall in 24 hours were downloaded from Atlas Climatológico.

Table 4 below shows a list of the different variables used in the regression analysis of this study and the source from where it was collected.

2.3.5 Selection of Variables

The variables were preliminarily selected similarly to Vogel et al. (1999) according to what was considered most likely to describe MAF. It is mentioned in section 1.1.1 that runoff maps are used to estimate the runoff in relation to the watershed area, which traditionally have been a common method for estimation of MAF. It is therefore reasonable to believe that area would have a major impact in the regression models as well. Other geomorphological variables were also studied, for example watershed perimeter and mean slope. Furthermore, several climatic variables were examined, such as precipitation and temperature, both on a yearly and monthly basis. For the monthly precipitation only six months were chosen which correspond to the six highest peaks in the bimodal precipitation pattern, see figure 3. These were selected since it was believed that they would have the greatest impact on runoff volumes, according to the continuity equation (equation 1) mentioned in section 1.1.1. Furthermore, number of days of precipitation in each month and the maximum amount of precipitation measured in 24 hours were thought to have a possible impact and were therefore tested in the analysis.

Other variables thought to possibly have an impact on the models were for example solar radiation and land use. These were analyzed by the use of maps. However, it was concluded that both of these variables showed very homogeneous values for the whole study area and it was therefore decided that they would not provide any additional information to the models.

Before performing the regression analysis further tests were performed to examine the variable's relevance, as described in section 4.3.

Table 4 shows a list of all the variables that were used in order to obtain reliable regression models.

Table 4: List of Geomorphic and Climatic Variables used as predictors in the models.

Variable name	*Definition*	*Unit*	*Source*	*Time Span*
Geomorphic Variables				
Area	Drainage Area	square kilometers	Salazar	
Perim	Perimiter of Basin	meters	Salazar	
S_{med}	Average basin slope	degrees	Salazar	
A/Perim	Drainage area divided by basin perimeter	meters	Salazar	
H_{med}	Mean basin elevation	meters	Salazar	
Climatic Variables				
μ_T	Mean annual temperature	degress Celsius	IDEAM	1981-2010
μ_P	Mean annual precipitation	mm/yr.	IDEAM	1981-2010
Evap	Mean annual evaporation	mm/yr.	IDEAM	1981-2010
P_{rs}	Mean precipitaion for the six most rainy months as described in section 3	mm	IDEAM	1981-2010
P_{amon}	Mean precipitaion for April, May, October, November	mm	IDEAM	1981-2010
P_{mam}	Mean precipitaion for March, April, May	mm	IDEAM	1981-2010
P_{son}	Mean precipitaion for September, October, November	mm	IDEAM	1981-2010
P_{jan} - P_{dec}	Mean precipitaion for January through December	mm/month	IDEAM	1981-2010
Max 24h P	Maximum precipitation in 24 hours in a certain month	mm	IDEAM	1981-2010
No. d. P	Number of days with precipitation in a certain month	days	IDEAM	1981-2010

2.3.6 Design flow in Norte de Santander and Santander

As mentioned in the introduction a regression analysis of extreme flow was performed which could be used for design purposes. For this analysis all data was not readily available. Extreme flow values i.e. the maximum value of streamflow for a specific month that has been registered by IDEAM, were extracted for each station and year (in the records called "Máxima absoluta") from the same data records as Salazar used. The same criteria for choosing measurement stations (minimum 20 years of historic data, minimum 90 % coverage per year, time range of 1981-2010) was used and thus data from the same stations and years was used also in the design flow analysis. This data was then analyzed further before the regression analysis was performed. How this was done is described further on.

Apart from the extreme flow data some readily available geomorphic characteristics were used, namely area, slope and elevation which can be found in table 4.

3 Theory

3.1 Regression analysis

Regression analysis is a widely used statistical tool in several areas, among others economics, chemistry, environmental science and psychology. It is a conceptual method used to examine the relationship between a response variable and one or more so called explanatory or predictor variables (Sambandsanalys 2012; Shewhart and Wilks 2006). In this thesis the response variable is the streamflow in the region of interest and the parameters that were examined as predictor variables are explained in table 4 in section 2.3.5. In other words the scientific question is how the response variable (the streamflow), is dependent of the predictor variables, or as expressed in chapter 1.2, how does the streamflow depend on climate and basin characteristics.

A regression analysis can be performed in different ways. In this thesis two types will be examined; simple and multiple linear regression. The difference between these two analysis is that the simple regression model only deals with one predictor variable while the multiple regression model incorporates various predictor variables. The models will be fitted by using Ordinary Least Squares method (OLS) to estimate the coefficients of the predictor variables which is explained in section 3.1.3.

In order to produce a reliable regression model with the OLS-method the following assumptions stated by Shewhart and Wilks (2006), have to be made and examined before any conclusions about the model can be drawn.

1. Linearity assumption: A linear relation exists between the response and the predictor variables.
2. Assumption about the errors: The errors are assumed to be independent and identically distributed, this implies that:
   (a) the errors have a normal distribution
   (b) the mean is equal to zero
   (c) they all have the same variance that is unknown
   (d) they are linearly independent
3. Assumption about the predictor variables:
   (a) they are selected in advance
   (b) they are measured without errors
   (c) they are linearly independent from each-other
4. Assumption about the observations i.e. the dependent variable: they are all equally reliable and weigh equally in the regression model.

The first assumption can be examined by analyzing the correlation. The second can be validated if the residual plot looks reasonable i.e. the residuals are randomly distributed. The two first points (a and b) in the third assumption about predictor variables cannot be validated although the third point (c) is important and the correlation value is one way to control this. The last assumption cannot be validated either however the advantage with the OLS-method is that small violations of this assumption will not give misleading results (Shewhart and Wilks 2006). It is explained further on in this section (section 3.1.4) how these assumptions are analyzed and/or validated.

3.1.1 Simple Linear Regression

The simple linear regression is a model that results in an estimated line and its equation describes the estimated Y, i.e. the dependent variable. The line is a consequence of the equation since only one predictor variable is used to explain the dependent variable. The equation for simple linear regression is expressed as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

where Y is the matrix of the dependent variable, X the matrix of the predictor variables, β_0 is called the constant coefficient or the intercept, as it predicts the value of Y when X is equal to zero. The coefficient β_1 is termed the slope which indicates the change of Y when X changes with one unit and ε is the random error. The random error explains the variance in the approximation (Shewhart and Wilks, 2006).

3.1.2 Multiple Linear Regression

Multiple linear regression is a prolongation of simple linear regression, it can even be said that simple linear regression can be seen as a type of multiple linear regression. Different from the simple analysis is that it is harder to visualize the multiple linear regression because more dimensions are involved. When Y is explained by two predictor variables the multiple regression represents a plane instead of a line.

The relationship between the response and the predictor variables is approximated by the following equation (3).

$$Y = \beta_0 + \beta_j X_j + ... + \beta_p X_p + \varepsilon, \qquad j = 1, 2, 3...p \tag{3}$$

where β_0 is the constant coefficient and similar to simple linear regression, it gives the value of Y when X$_1$ = X$_2$ = ... = X$_p$ = 0. β_j is the regression coefficient, also called partial regression coefficient and it accounts for the influence of X$_j$ on the response variable Y after the coefficient and the variable have been adjusted to the other variables. j is the number of predictor variables and ε is the random error.

3.1.3 Model Fitting Using Ordinary Least Squares Method

In order to estimate the coefficients of the regression, or in other words, find the best fit of the observation points, several methods can be used, for example methods called the maximum likelihood method or the principal components method. However, the most commonly used method is called the least squares method which exists in several variants, for example ordinary and weighted least squares. In this study the ordinary least squares method was applied.

This method implies that the sum of squares of the errors are minimized. For simple linear regression this is expressed mathematically below in equations 4-7 (Shewhart and Wilks, 2006).

The regression equation (eq. 2) is first rewritten as:

$$\varepsilon = Y - \beta_0 - \beta_1 X \tag{4}$$

The sum of squares is then equal to:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon^2 \tag{5}$$

The least squares estimates $\hat{\beta_0}$ and $\hat{\beta_1}$ are calculated to minimize $S(\beta_0, \beta_1)$ through:

$$\hat{\beta_1} = \frac{\sum (y_i - \overline{y})(x_i - \overline{x})}{\sum (x_i - \overline{x})^2} \tag{6}$$

and

$$\hat{\beta_0} = \overline{y} - \hat{\beta_1}\overline{x} \tag{7}$$

where x_i and y_i are the i:th elements in X and Y and $\overline{x}$ and $\overline{y}$ are the mean values. Finally resulting in the least squares equation:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_1 \tag{8}$$

The case for multiple linear regression is analogous with that for simple regression with the only difference that the regression equation is rewritten as:

$$\varepsilon = Y - \beta_0 - \beta_j X_j - ... - \beta_p X_p \qquad j = 1, 2, ..., p \tag{9}$$

meaning that the estimates $\hat{\beta_1}, \hat{\beta_1}, ..., \hat{\beta_p}$ that minimize the sum of squares of the errors $S(\beta_0, \beta_1, ..., \beta_p)$ are calculated by solving a system of linear equations, resulting in the least squares regression equation:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X + ... + \hat{\beta_p} X_p \tag{10}$$

3.1.4 Statistical Analysis for Model Criticism and Validation

In order analyze and validate the resulting regression models and to decide which predictor variables give the best possible regression models several methods of statistical analysis were used as criteria.

Coefficient of determination, R^2, and adjusted coefficient of determination, $R^2{}_{adj}$

One of the important tests for examining the quality of the model is with the coefficient of determination also called goodness of fit index or as most commonly used R^2. This provides information about the accuracy of the regression line (simple regression) i.e. how much the dependent variable can by explained by the predictor variable(s). It is a coefficient that ranges from 0-1 where 1 means a perfect fit between the observed values and the fitted ones, in other words, that the predictor variable(s) perfectly explain the dependent variable, Y.

The R^2 is an analysis of the total variation in the model which can be explained for both simple and multiple regression as follows:

$$SST = SSR + SSE \tag{11}$$

where SST is the total variation i.e. the variation in the y-values. SSR means the part of variation in the y-values that can be explained by the regression, i.e. the predictor variable(s). SSE is the remaining variation which cannot be explained by the regression and is therefore the sum of the errors (residuals). It can be simpler put as follows:

$$''Total\ variation'' = ''variation\ explained\ by\ the\ line'' + ''unexplained\ variation''$$

It is described above that the R^2 is a measurement of how well X can explain Y and with the use of the variations explained in equation 11 the R^2 can be expressed as follows:

$$R^2 = \frac{SSR}{SST} \tag{12}$$

or

$$R^2 = \frac{Variation\ explained\ by\ the\ line}{Total\ variation}$$

Equation 12 gives the proportion of variance that can be explained by the explanatory variables (Shewhart and Wilks 2006).

The R^2-value can be misleading when used for several predictor variables as it will increase for each variable added to the equation without the necessity that the model fitness increases. This is solved by using the adjusted R^2 coefficient instead which also describes the correlation and only increases if the variables added actually explain the dependent variable. This means that the coefficient will increase only if the variable added improves the model (Investopedia n.d.). In other words the correlation between the predictor and each explanatory variable is calculated and for a higher correlation, which indicates that the explanatory variable can explain the predictor variable better the goodness of fit index will be allowed to reach a higher value.

In this study an (adjusted) R^2-value above 0.65 was seen as acceptable.

P-value

It belongs to the normality to perform a so called test of the null hypothesis when performing a regression analysis.

The null hypothesis is performed on the β-coefficient, see equation 2, because this coefficient explains how the response variable, Y, changes when X changes (Sambandsanalys 2012; Shewhart and Wilks 2006).

The hypothesis test wants to examine whether or not the β-coefficient could be equal to 0. If this cannot be denied then no linear correlation can be assumed between the response and predictor variable and the predictor variable should be removed from the model. The test is shown mathematically in equation 13 below (Sambandsanalys 2012; Shewhart and Wilks 2006).

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{13}$$

The null hypothesis can be performed with a so called t-test and H_0 can be rejected to a specific significance level which can be compared with the p-value. This is explained graphically in figure 5 where a t-distribution for a t-test is shown. The p-value is the sum of the two delimited corners (Shewhart and Wilk 2006).

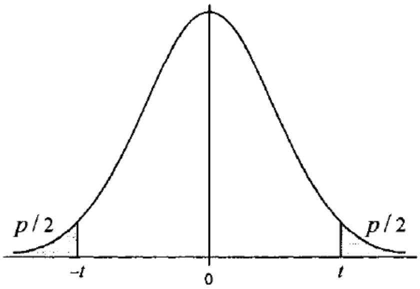


Figure 5: An example of a t-test with a t-distribution where the p-values are marked on the edges (Shewhart and Wilks, 2006, p. 34).

The p-value gives an insight of the significance of the predictor variable's (X) influence on the response variable (Y) i.e. the relationship between the two. It is often incorporated into statistical programs, for example R-statistics, where it is presented as in table 5 below. A lower value signifies a better relationship between the predictor and the response variable. In this study p-values lower than 0.1 were accepted.

Table 5: Range of p-value that are acceptable in this thesis.

0-0.001	0.001-0.01	0.01-0.05	0.05-0.1
***	**	*	.

Correlation Coefficient and Variance Inflation Factor, VIF

The correlation coefficient is a measurement of the linear relationship between the dependent (Y) and the predictor variable (X) or between different predictor variables. It can be observed graphically, by creating a scatter plot of Y vs. X and examine if they follow the same line, or mathematically where the coefficient ranges between -1 and 1, this is displayed in figure 11 in section 4.3. The closer the coefficient is to -1 or 1 the stronger the relationship is, the sign reveals if they are positively or negatively related.

The variance inflation factor, VIF, is an important aspect in regression analysis as it is a method for detecting multicollinearity among the predictor variables. Multicollinearity signifies that the estimated coefficients from the regression models are unstable due to a high probability of linear relationship between two or several predictor variables. Therefore identifying multicollinearity is highly important for the performance of the regression models. This can be done in different ways, for instance through a correlation test, although this method does not present the same precision as a VIF-test and some multicollinearity may not be detected.

The VIF-number is calculated for each predictor variable with the use of R^2 according to equation 14 below.

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, ..., p, \tag{14}$$

where j is the number of predictor variables and ${R_j}^2$ is the coefficient of determination for the predictor variable with number j when it is regressed against the remaining predictor variables. More simply put, the ${R_j}^2$-value measures linearity between the variables and the closer its value is to 1, the more linear they are. For the VIF-number this will result in a high number and if the number increases above 10, collinearity in the model can be assumed.

This is resolved by eliminating or changing one or more predictor variables in the model (Shewhart and Wilks 2006). In this analysis a maximum VIF-number of 5 will be accepted which is presumed to produce a more trustworthy model.

The two coefficients explained can often be included in statistical programs and do not need to be calculated manually. They can be used to confirm or deny the linearity assumption (nr 1) and the assumption about the predictor variables (3c).

Residuals

The definition of a residual is the vertical distance between the estimated line and the observed y-values. Residuals have a fundamental role in approving the assumption that the errors are normally distributed and independent as mentioned in section 3.1 assumption 2. By creating a scatterplot or a histogram of standardized residuals the assumption can be confirmed or not. The residuals are standardized when their mean is zero and the standard deviation is 1 (Shewhart and Wilks 2006).

The residuals can be plotted in a scatterplot against the fitted values where they should be uncorrelated and not show any tendencies. The histograms should visualize a normal distribution (Sambandsanalys 2012; Shewhart and Wilks 2006). In figure 6 and 7 a good (to the left) and a bad (to the right) example of a residual plots can be observed. In figure 6 the residuals are plotted against X which is identical to fitted values and in figure 7 histograms can be viewed, both will be used in this study (Shewhart and Wilks 2006).

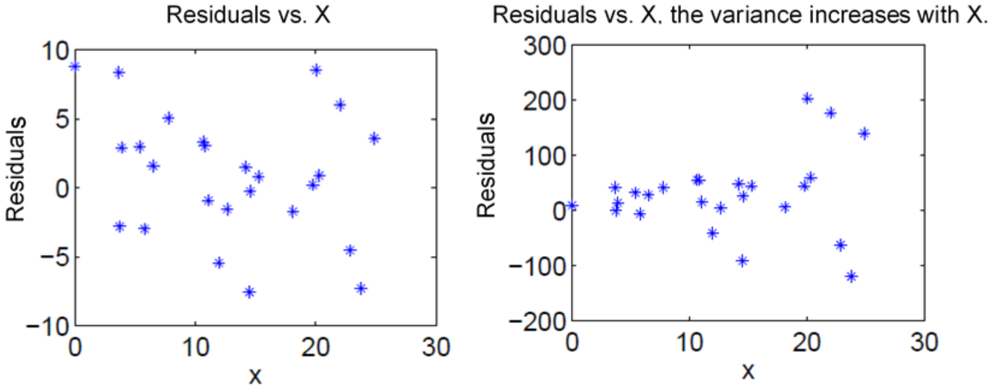


Figure 6: Two residual plots, the one to the left shows an example of a good residual plot and the one to the right shows a residual plot with trends (Sambandsanalys 2012 p. 13, modified by the authors).

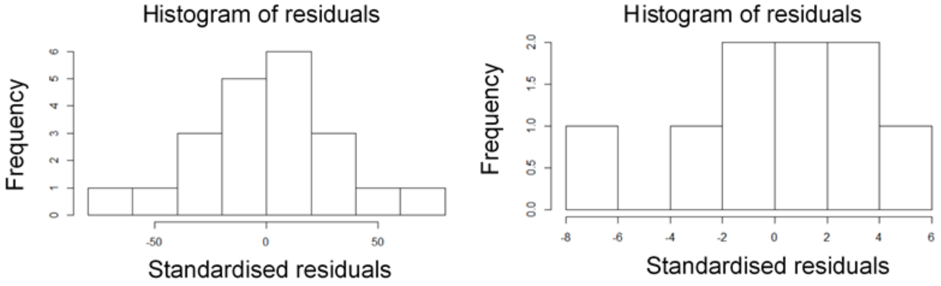


Figure 7: Two histograms of standardized residuals. The one to the left shows an example of a good histogram where the residuals have a normal distribution. The one to the right does not have a normal distribution and therefore the conditions are not fulfilled.

Outliers

Outliers are extreme values in the data with either an unusually high or low value in comparison to the rest of the data. The source of the outliers could be a measurement error or it could be a natural deviation in the data. In any of the cases the variance will have a notable effect on the resulting model. Accordingly one has to examine the outlier and decide if it should be erased from the data or if modification in the model will result in a better outcome (Sambandsanalys 2012).

For this project examples of common outliers were streamflow values which deviated considerably from other values measured close in time. The sources to these deviations was unknown and they were considered misrepresenting and were therefore removed from the data.

Summary of statistics coefficients

Table 6 gives a short summary of the coefficients explained in this section as well as the coefficient values that have been decided to be acceptable for this study. The values were selected by recommendations from literature.

Table 6: Summary of statistic coefficients

Coefficient	**Explanation**	**Acceptable value**
R^2	Evaluation of the model fit	> 0.65
P-value	Evaluation of the relationship between predictor and response variable	< 0.1
Correlation	Evaluation of the relationship between the different predictor variables	< 0.8
VIF	Evaluation of the relationship between predictor variables	< 5
Residuals	Deviation between the estimated line and the observed value	-
Outliers	Extreme values in data	-

3.2 The Area Weighted Method

The area weighted method creates isohyets or isotherms which are lines that enclose a rain-gauge and a specific rain-depth, or the equivalent for temperature. Each line represents the same precipitation or temperature. For simplicity's sake the method is described below for precipitation (isohyetal method) but it works in an analogous manner also for temperature.

The precipitation at the ungauged locations is interpolated within the isohyets. This method requires a dense gauge network and area information in order to create reliable isohyets (Goovaerts 1999). The method was used to calculate the mean of each parameter in each rainfall basin for the study using equation 15.

$$\overline{P} = \frac{\sum_{i=1}^{n}(A_i * \hat{P_i})}{\sum A_i} \tag{15}$$

where $\overline{P}$ is the mean precipitation for one runoff-basin, A_i is the area between two isohyets which is calculated in GIS as explained in section 4.2.1. $\hat{\mathrm{P}}$ is the mean precipitation within two isohyets, it is calculated according to equation 16.

$$\hat{P_i} = \frac{P_i + P_{(i+1)}}{2} \tag{16}$$

3.3 Distribution of Extreme Flows and Calculation of Design Flows

For the purpose of designing hydraulic structures such as small dams and ponds or bridges it is of interest to have knowledge of extreme flows for the watercourse in which the structure will be built. This can be done through statistical analysis of measured streamflow and the estimation of design flows for a reasonable return period.

Despite no official sources exist to confirm it is commonly accepted that extreme flows in Colombia are distributed according to the Gumbel distribution (Villareal González

2017). The Gumbel distribution is a type of extreme value distribution and can be used in many different areas, such as high temperatures, high wind speeds, large fluctuations in exchange rates and, as in this case, extreme flows. Other examples of extreme value distributions are the Fréchet and Weibull distributions (mathwave n.d.).

In this study design flow values for return periods such as 10-year, 50-year etc. were calculated by fitting the extreme flows to the Gumbel distribution function. The equations used for this were (U.S. Department of the Interior Geological Survey (USGS) 1981):

$$M = X_{bar} - 0.45005S \tag{17}$$

$$B = 0.7797S \tag{18}$$

$$X = M + B(-ln(lnP)) \tag{19}$$

where

M = *Mode* (depends on X$_{bar}$ and S)
B = *Slope*
X = *Magnitude*
X$_{bar}$ = *Mean*
P = *Exceedance probability*
S = *Standard deviation*

The magnitude (of the streamflow, in this case) could then be read for the exceedance probabilities of interest (USGS 1981).

4 Method

4.1 Software Used

In order to process data, calculate necessary variable values and perform the regression analysis some software were used. These are introduced below with a short explanation.

4.1.1 ArcGIS version 10.5.1

ArcGIS is a package of products such as ArcMap, ArcCatalog and ArcToolbox developed by ESRI (Environmental Systems Research Institute, Inc.). It is a software that allows for management of spatial data, GIS-data and map analysis and cartography (Larsson 2013). In this study version 10.5.1 was used.

4.1.2 R version 3.4.2

R is a language and an environment developed by the GNU project, which is a free operating system. R is a free software which can be used for statistical computations and graphics (R-project.org 2017). In this study version 3.4.2 was used. R-studio was the software used as development environment.

4.2 Calculation of Variable Values

4.2.1 Climatological Data Using Shape Files

Shape files with climatological data such as temperature and precipitation were downloaded from the Geovisor to be processed in ArcGIS together with area information in shape files obtained from Salazar (2016). The geographical coordinate system, GCS, for these shape files was WGS 1984 which is a threedimensional system used globally and the positions are defined by GPS-satellite images (Lantämteriet nd.). However, when incorporated in to ArcMap the GCS was converted to MAGNA which is a geographical coordinate system specified for Colombia. The area weighted method was used, as described in section 3.2, in order to obtain one single value of mean annual temperature and precipitation for each region. How this was executed in ArcGIS can be viewed in figure 8, 9 and 10 below where a temperature map from IDEAM is displayed. It is incorporated with the two departments of the study in the first figure, in order for the reader to gain a visual of the study area. In the second figure the 37 watersheds are incorporated into the map with the reason to acquire the temperature range within each region. Thereafter figure 10 displays one catchment and the temperature variation within this catchment. Each area, marked with a different color in the figure, represents one temperature range and as the area for each catchment was known the area for each division could be calculated when the two shape files were merged together.

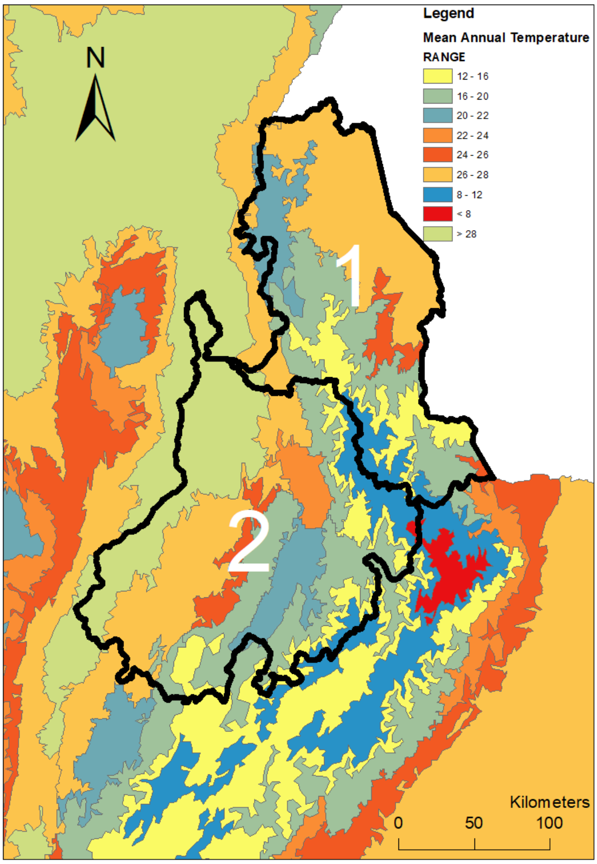


Figure 8: 1. Norte de Santander 2. Santander
A shape file with temperature information that was used in the study and the delimitation of the study area. Data source: Geovisor (IDEAM 2017a).

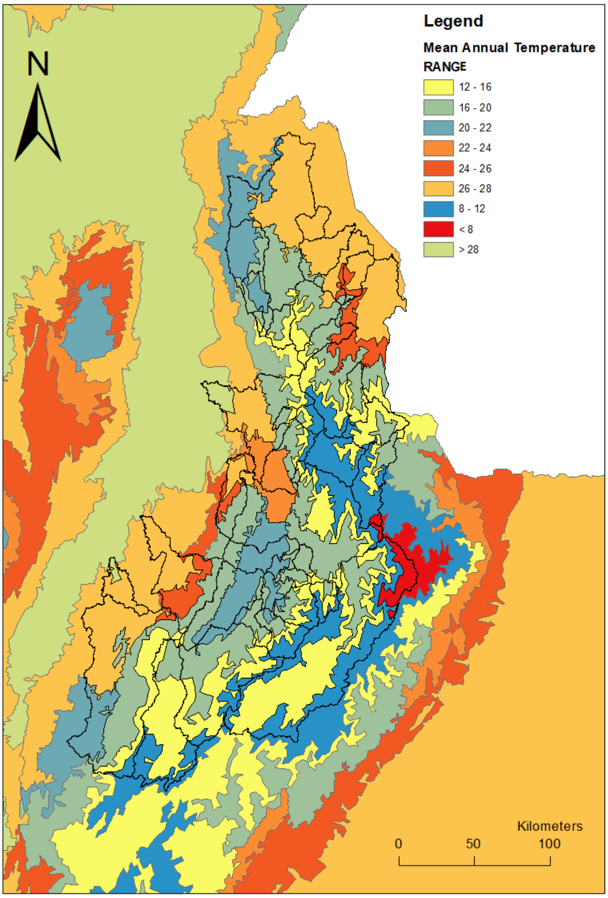


Figure 9: The map for temperature is divided into all the 37 watersheds in order to obtain the range within each region. Data source: Geovisor (IDEAM 2017a).

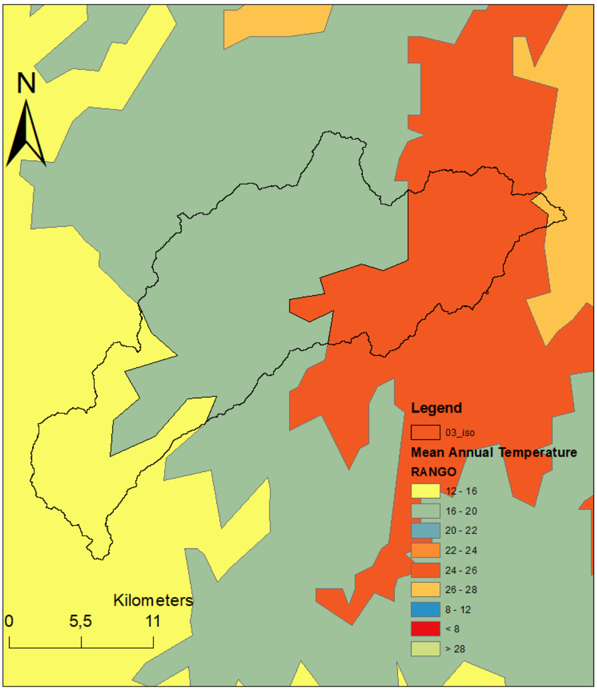


Figure 10: The isotherms within a catchment, each division/color has a proper temperature range. The area between each isotherm is calculated in GIS.

4.2.2 Monthly Values

Apart from calculating the mean annual values for precipitation and temperature as mentioned in section 4.2.1 mean monthly values for some months were also calculated for precipitation. These months were chosen according to the precipitation regime as explained in section 2.2.3. In addition, number of days of precipitation in each month and the maximum amount of precipitation measured in 24 hours during each month was also retrieved. These values were extracted from color graded maps by overlaying the contours of the area of the watersheds and reading the values.

4.2.3 Mean Annual Streamflow

As mentioned in section 2.3.1 data for MAF had already been collected and calculated previously by Salazar (2016). This data could therefore be used directly without further manipulation.

4.2.4 Design Flow

The design flow for each year in each basin was obtained from the same record of observations as Salazar (2016) had used in his study. How the design flows were calculated will be explained further on in section 4.4.

4.3 Performing the Regression Analysis

Once all values for the selected variables had been calculated the regression analysis were performed. This was done using R programming language in the software R-studio. An example of the R-code used can be found in appendix I. In order to reach the final regression models several steps were taken, according to the list below (Shewhart and Wilks, 2006):

1. **Test correlation between variables.** After selecting the potentially relevant variables the correlation between these was tested. This test was used both in order to define which predictor variables were most likely to give a good fit of the regression and to reject predictor variables which showed correlation between each other and could therefore not be approved as independent variables.

   Selecting the most suitable predictor variables was done by simply looking at the scatter plots of the response variable Q and each of the predictor variables. If there seemed to be a linear correlation between Q and any predictor variable this variable was selected for testing in the regression.

   In order to be accepted as an independent predictor variable the correlation coefficient between two variables was allowed to be maximum 0.8. If this was not the case one of the variables was rejected for that specific regression or the two variables were combined. An example of this is the two predictor variables area and perimeter, as they both explain size it is not surprising that they have a strong correlation. This is solved either by eliminating one of the variables or by combining them for example like this: area/perimeter.

   Figure 11 below illustrates an example of a correlation scatter plot. It can be seen that there is a clear linear correlation between Q (flow) and A (area) (correlation coefficient = 0.90) and also between Q and Per (perimeter) (correlation coefficient = 0.88). It is also clear that A and Per are closely correlated as is Te (temperature) and H$_{med}$ (mean elevation). This can be deduced both from observing the scatter plots relating these variables and also by their calculated correlation coefficients which are also presented in the figure and have values of 0.97 and 0.99, respectively. It was therefore clear that A and Per should not be used independently in the same regression, and neither should Te and H$_{med}$. A further explanation of what the correlation coefficient implies was offered in section 3.1.4 under the sub-heading correlation coefficient.

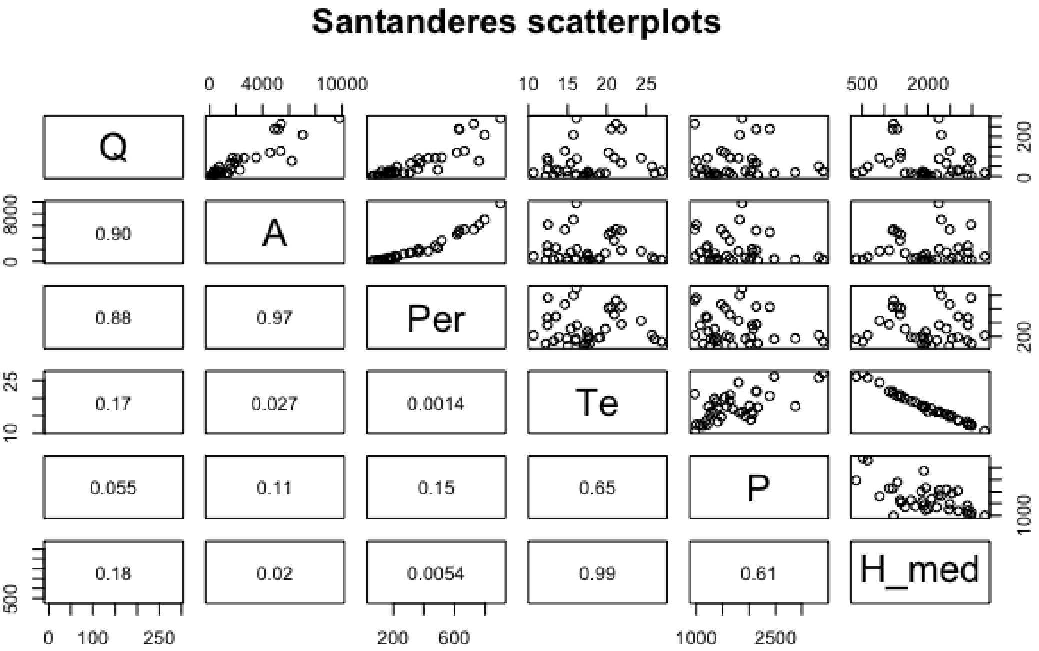


Figure 11: Example of scatter plots of some variables for the entire area of study. The variables presented in this figure are Q = MAF, A = watershed area, Per = watershed perimeter, Te = mean annual temperature, P = mean annual precipitation, H_ med = mean watershed elevation. The graphs display the correlation between the variables on each respective row (y-axis) and column (x-axis). Correaltion coefficients are displayed on the left side of the diagram in the same manner.

2. **Specification of model.** In this study it was assumed that the response variable could be related to the predictor variables by a linear model as was described in section 3.1. There is a possibility that the relationship will be non-linear instead and this was examined by observing the value of R^2.

3. **Estimation of model parameters.** Regression analysis was performed for each of the defined regions with the method stepwise regression which is explained in section 2.3.3 above. Several regressions were performed using different variables. They were performed in R using the function *lm* which fits a linear model to the data set using the OLS-method. These regressions are described below:

   (a) *Regression using only A as a predictor variable.* This was done in order to test the simplest possible regression model. Area is a variable which is easily obtained simply by having access to a map making this kind of regression model a useful tool in areas where data is scarce.

   (b) *Regression using A, Te and P as predictor variables.* Since it was believed that the use of climatic variables in the regression models would improve the model an analysis was also performed using the most easily accessible climatic variables, precipitation, P and temperature, Te.

(c) *Regression using most suitable variables.* Ultimately regression analysis were performed where the aim was to obtain the best possible model. This was done by testing combinations of the different variables until the best possible model was reached. More specifically the method of stepwise regression was utilized. This method takes advantage of the use of the determination coefficient, R^2, see section 3.1.4. As more explanatory variables were added to the equation the R^2 was compared with the previous value of R^2. If it was better the predictor variable was kept and the other variables in the equation were removed to investigate if the R^2 increased. This was repeated until there were no more predictor variables to add.

4. **Model Criticism and Validation.** In order to decide which model should be seen as the best possible, a number of statistical tests were performed as explained in section 3.1.4. Many different tests exist but for this study the below stated were found most convenient. Both due to the fact that some of them were used in the studies this thesis was based on and because they were easily accessible for use in the software employed.

   It is impossible to obtain a model which is perfect in all aspects. Therefore, it was necessary to choose models which were good enough, fulfilling the criteria stated below. If several models fulfilled the criteria the model with the R^2-value closest to 1 was chosen. The coefficients and their acceptable values are mentioned in table 6 and are repeated below.

   (a) *P-value.* It was considered desirable for the p-value to have a value below 0.1 for each predictor variable. If this was not the case the variable with a p-value higher than 0.1 was disregarded or changed to another variable.

   (b) *Residual histogram* should look normally distributed. Figure 12 illustrates an example of a histogram obtained in R where the residuals appear to be normally distributed. Models with residual histograms similar to this were seen as acceptable and trustworthy models.

   (c) *VIF* should have a value between 1-5 to be accepted as a trustworthy model.

   (d) R^2 should reach a value as close to 1 as possible.

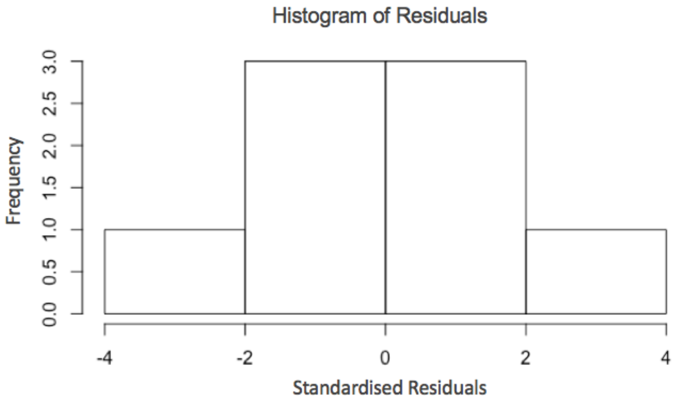


Figure 12: A typical residual histogram where the residuals appear to be normally distributed.

(e) *Residual scatter plot.* This was examined in order to discover correlation between the residuals. A regression model was only seen as acceptable if the residuals were randomly distributed. Figure 13 below illustrates a scatter plot of residuals obtained in R where the residuals appear to be randomly distributed. Models with scatterplots similar to this were seen as acceptable and trustworthy.

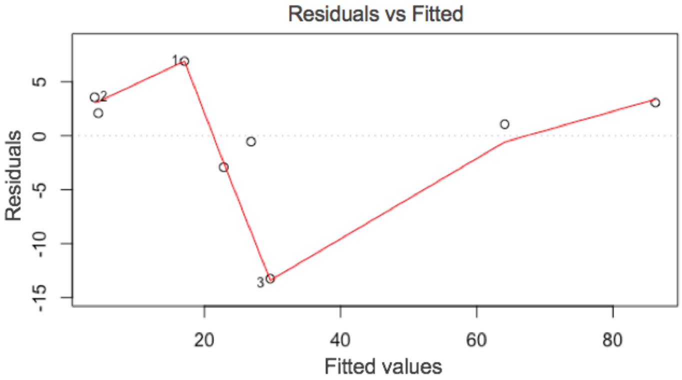


Figure 13: A typical residual scatter plot where the residuals appear to be randomly distributed.

In order to perform a final validation of the regression models the observed streamflow values were plotted against the estimated values for streamflow calculated using the obtained models. A regression line was fitted to this plot and the R^2 was observed.

4.4 Regression Models of the Design Flows

Once the extreme values for each watershed and each year were obtained the values were plotted in the Gumbel plot to make sure that this was a reasonable distribution to use. When this had been confirmed the equations in section 3.3 could be used in order to calculate values for the design flows with return periods of 10, 20, 50 and 100 years. Figure 14 below shows an example of a Gumbel plot used to obtain the design flows.

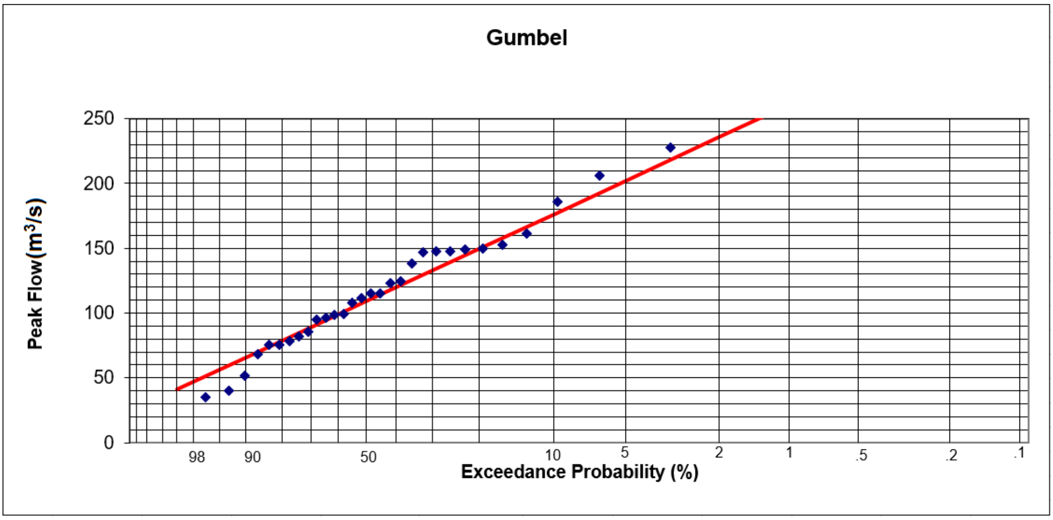


Figure 14: An example of a Gumbel plot used to extract the design flow values. The blue dots represent the measured extreme flow values while the red line is used to read the design flows.

Once these values had been calculated a regression analysis was performed similar to the one described in the previous section. However, this time the focus was on relating the above mentioned design flows to geomorphic variables such as area, mean basin slope and mean basin elevation according to the study of Dingman and Palaia (1999). Four different regression models were performed for each design flow:

1. Regression using only A as a predictor variable.
2. Regression using A and H_{med} as a predictor variables.
3. Regression using A and S_{med} as predictor variables.
4. Regression using A, H_{med} and S_{med} as predictor variables.

The same criteria were followed again in deciding whether a model was acceptable or not.

Similar to the model above a final validation test was performed on the regression models where the design flows calculated with the Gumbel distribution were plotted against the estimated values for design flow. The estimated values were calculated with the use of the models produced. A regression line was fitted to this plot and the R^2 was observed.

5 Results

5.1 Regression Analysis of Mean Annual Streamflow

As mentioned in section 4.3 the regression analysis was performed in three steps; using only A as predictor variable, using A, P and Te as predictor variables and finally using the seemingly most suitable variables to create regression models for MAF.

The results from these regression analysis will be presented in the following sections in three parts; first the coefficients of the predictor variables along with R^2 and maximum VIF will be displayed in a table. Thereafter the histograms of standardized residuals which were considered to be normally distributed will be exhibited, the remaining histograms can be found in appendix II. Finally graphs where observed mean streamflow is plotted against estimated mean streamflow are presented for the four big regions in this study; the entire area of study, 28 river basins, Norte de Santander and Santander. These regions were selected since essentially they are of highest importance in the study. The remaining graphs can be found in appendix III.

5.1.1 Regression Analysis using Area as Predictor Variable

The equation used to formulate this model is as follows:

$$\mu_Q = a + bA \tag{20}$$

where μ_Q is the MAF, a is the intercept, b is a constant and A the first predictor variable, i.e. the area.

Table 7 below shows the estimated coefficients for each region.

Table 7: Regression Models for Mean Annual Streamflow using area as predictor variable.

Region	Intercept (a)	Area (b)	Multiple R^2
Entire Area of study	3.0234	0.0307	0.812
28 River Basins	5.577	0.0308	0.785
Norte de Santander	-5.281	0.0372	0.874
Santander	5.838	0.0294	0.791
G1HHR1	-4.129	0.0380	0.726
G2HHR1	4.227	0.0395	0.820
G2HHR3	-2.730	0.0436	0.958
G3HHR1	-3.771	0.0424	0.842
G4HHR2	-1.988	0.0376	0.904
G5HHR1	2.645	0.0106	0.931
G5HHR2	4.283	0.0419	0.850
G5HHR4	47.376	0.0237	0.285
G6HHR2	0.434	0.0425	0.849
G7HHR1	-2.630	0.0436	0.892
G7HHR2	-123.530	(A/Per) 37.650	0.292
G8HHR1	0.378	0.0391	0.950

Note that for G7HHR2 A/Per has been used as the predictor variable instead of only area. The reason for this is that it showed a slightly better result than using only area. This was tested for all regions but it only showed a better result in this one case. The reason this was tested was due to the strong correlation between these two variables as explained in section 4.3.

As can be deduced from the table (7) most regions present an acceptable R^2-value (higher than 0.65) making the models appear trustworthy according to the criteria stated in table 6. However the models for two regions, G5HHR4 and G7HHR2, have much lower R^2-values, 0.285 and 0.292 respectively, which means that the response variable Q is not well explained by using area (or area/perimeter) as a predictor variable.

Figure 15 below demonstrates the residual histograms for the regions which were viewed as acceptable. It cannot be said that they are completely normally distributed although they have a shape similar to the one in figure 6. Therefore they are acceptable histograms which indicate that the normality assumption for the errors, 2.1. mentioned in section 3.1 can be validated. This is required to obtain acceptable and trustworthy models. The histograms for HHR G2HHR3 and G6HHR2 do not show a perfect normal distribution but together with their high R^2-values it was still decided that they could be seen as acceptable. The residual histograms which were viewed as non-acceptable can be found in appendix II.

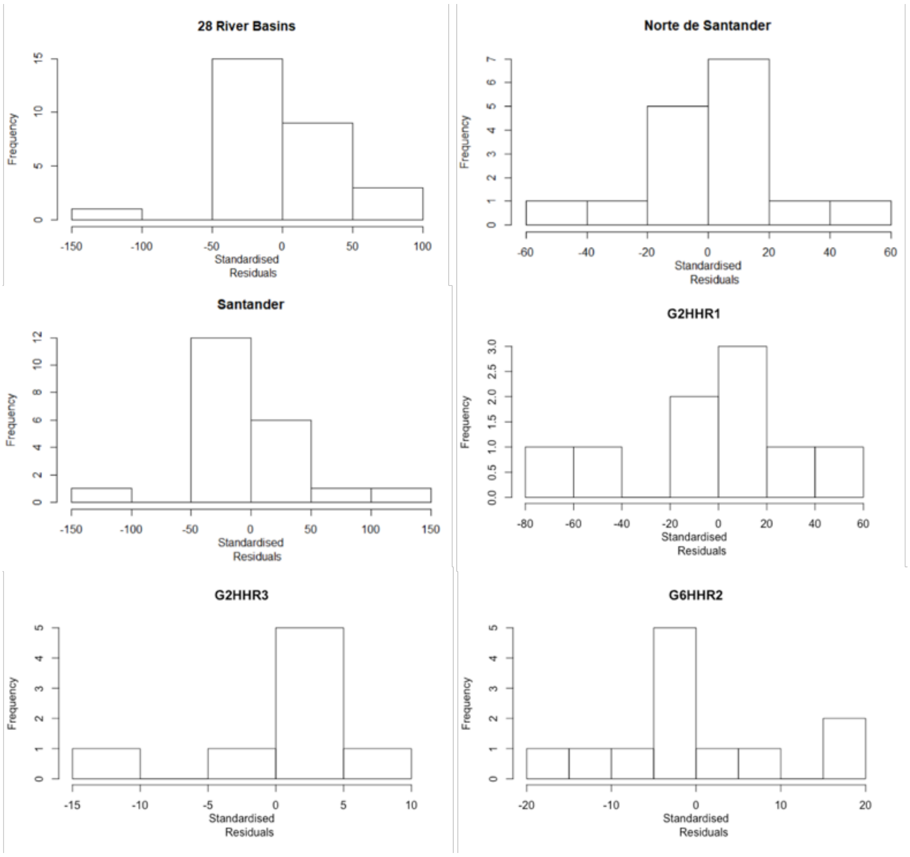


Figure 15: Residual histograms which are seemingly normally distributed. Obtained from regression models with only area as predictor variable.

Six histograms can be viewed which imply that 10 out of 16 regions show a residual behavior that is not normally distributed. These models should therefore not be seen as trustworthy.

In figure 16 below the fitted line for observed values plotted against values estimated with the obtained regression models can be observed.

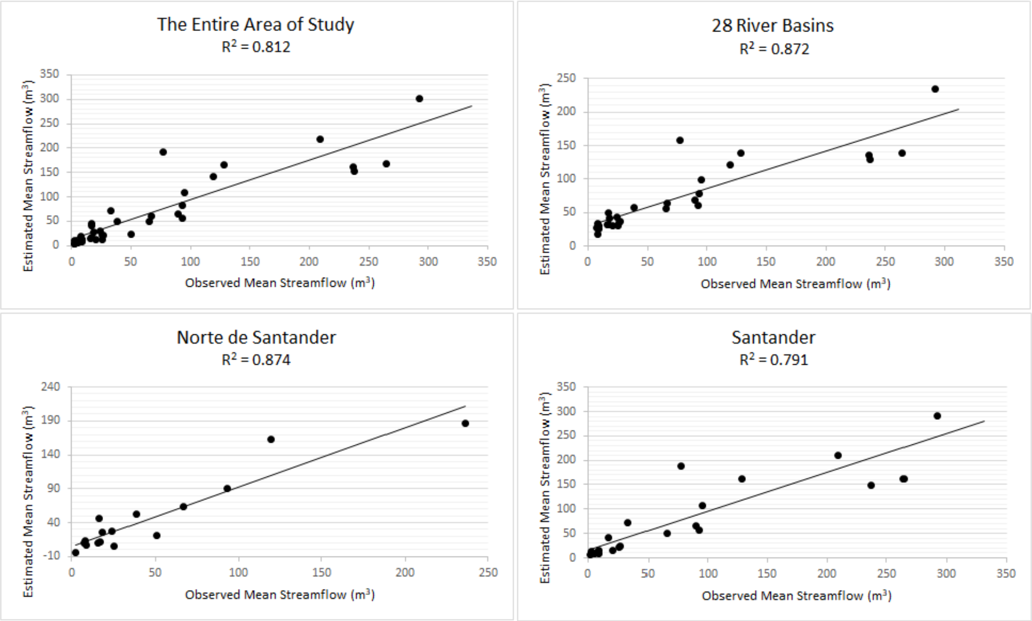


Figure 16: Comparison of estimated and observed MAF using regression with only area as predictor variable.

As can be understood from the above plots the fit is relatively good for the four larger regions. This can also be evaluated looking at the R^2-values which are quite high in comparison with the required value stated in table 6. The points in the graphs are spread out along the line although this is no indication of outliers as the streamflow is estimated/observed for different catchments. Provided that the points are situated along the line it can be said that there is a small difference between the estimated and the observed points. This is true for all four graphs above therefore no outliers are present.

Considering all the results six models end up fulfilling all the criteria. These models correspond to the regions named 28 River Basins, Norte de Santander, Santander, G2HHR1, G2HHR3 and G6HHR2.

5.1.2 Regression Analysis using Area, Precipitation and Temperature as Predictor Variables

The results from the second regression model including climatic variables are displayed in table 8. Different from table 7 is that the adjusted R^2 is presented, which considers multiple variables, and the maximum VIF which informs about the relationship between the predictor variables, see section 3.1.4.

The equation used to formulate this model is as follows:

$$\mu_Q = a + bA + c\mu_P + d\mu_T \tag{21}$$

where μ_Q is the MAF, a is the intercept, b, c and d are constants, A the first predictor variable, area, μ_P is the second predictor variable mean annual precipitation. Lastly μ_T is the third predictor variable mean annual temperature.

Table 8: Regression Models for Mean Annual Streamflow using area, precipitation and temperature as predictor variables.

Region	**Intercept (a)**	**Area (b)**	μ_P **(c)**	μ_T **(d)**	**Adjusted R^2**	**Maximum VIF**
Entire Area of study	-65.315	0.0310	0.00739	3.116	0.837	1.773
28 River Basins	-69.668	0.0312	0.00659	3.505	0.805	1.604
Norte de Santander	-12.0272	0.0388	0.0363	-3.176	0.914	5.0413
Santander	-150.882	0.0289	0.0145	8.0412	0.860	1.0944
G1HHR1	25.635	0.0389	0.0181	-3.810	0.896	1.352
G2HHR1	-564.300	0.0671	0.00793	20.750	0.831	6.441
G2HHR3	-4.592	0.0423	0.0119	-1.254	0.992	1.166
G3HHR1	-507.000	0.0635	0.00290	19.060	0.851	4.791
G4HHR2	-29.531	0.0407	0.0255	-1.299	0.979	6.344
G5HHR1	8.361	0.0133	0.00476	-0.812	0.919	3.941
G5HHR2	-21.591	0.0415	0.00354	0.825	0.806	2.678
G5HHR4	-421.692	0.0426	0.0118	18.709	0.682	1.780
G6HHR2	-26.424	0.0426	0.00867	0.446	0.861	1.606
G7HHR1	-31.144	0.0400	0.00584	1.135	0.861	1.539
G7HHR2	-427.400	0.0423	0.00831	19.310	0.658	1.759
G8HHR1	-4.564	0.0396	0.0115	-1.0667	0.968	1.121

Table 8 shows that all models have an acceptable R^2-value above 0.65, see table 6, the lowest values are from models G5HHR4 and G7HHR2 with the values 0.682 and 0.658 respectively. This implies that the area, precipitation and temperature from the two models only explains 68.2% and 65.8% of the variance of MAF in the regions. Concerning the maximum VIF three models present values above the acceptable value of 5, see table 6; Norte de Santander, G5HHR1 and G4HHR2. Despite the valid R^2-values for these models it is questionable whether or not they can be considered as trustworthy as the high VIF-number indicates collinearity between the predictor variables which violates assumption 3c in section 3.1 about the predictors being linearly independent. Therefore they were not considered as acceptable.

Figure 17 below displays the six models with acceptable residual histograms. The histogram for HHR G6HHR2 is more questionable than the other ones, but along with its high R^2-value it was still decided that it could be seen as acceptable.

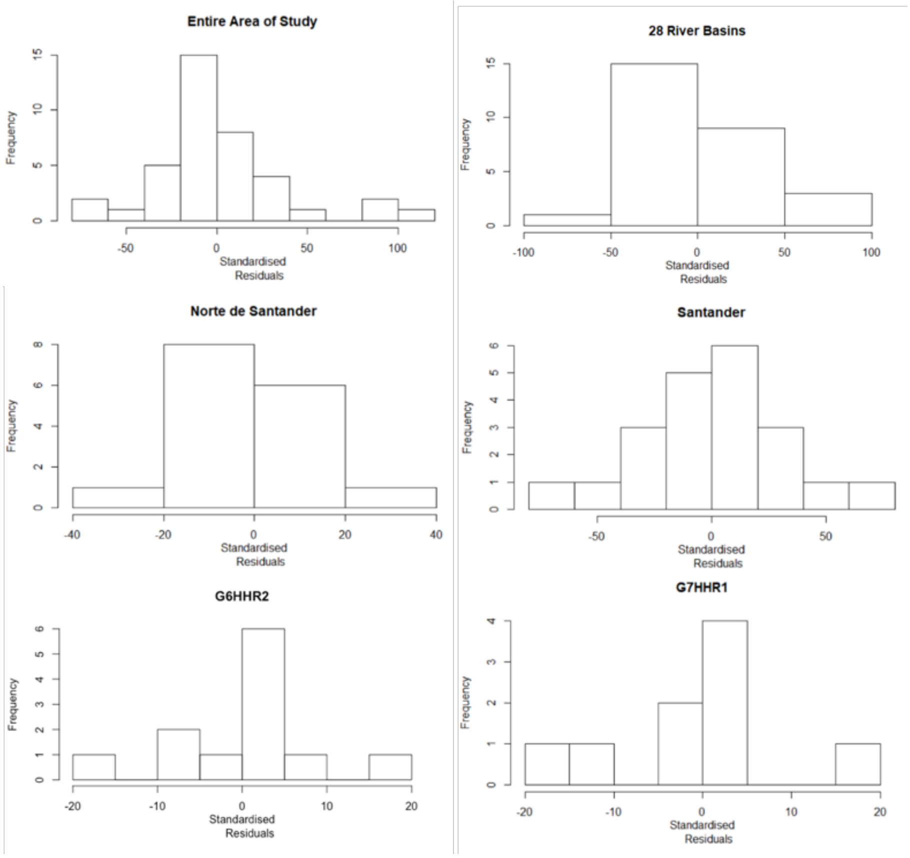


Figure 17: Residual histograms which are seemingly normally distributed. Obtained from the regression models with area, precipitation and temperature as predictor variables.

This result signifies that for 10 out of the 16 models tested, the normality assumption for the errors cannot be validated, see graphs in appendix II, and therefore those models are not trustworthy.

The fit of the observed against the estimated MAF for the four big regions can be observed in figure 18 below.

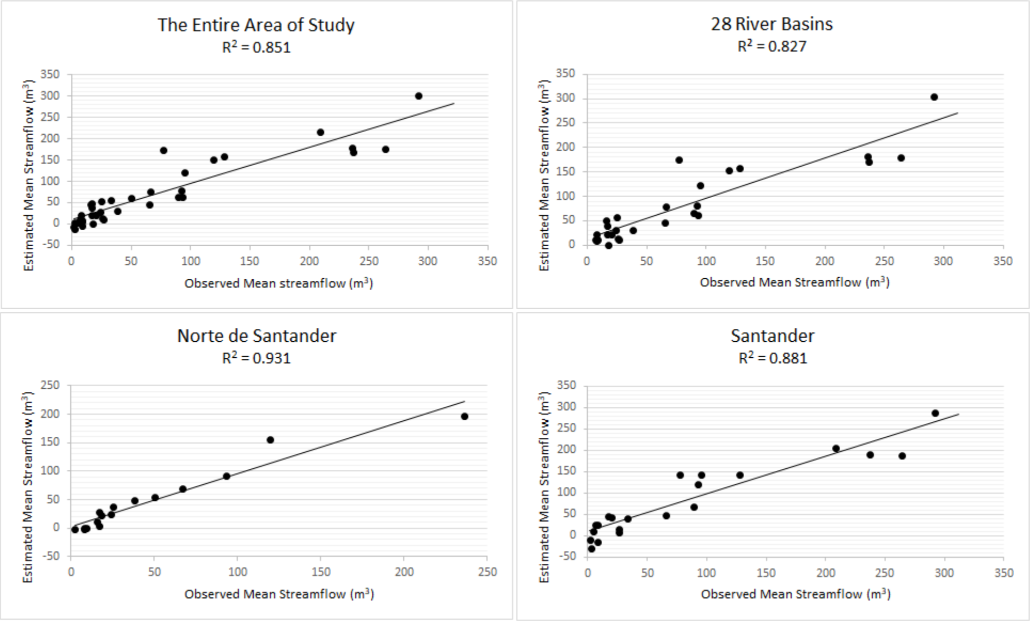


Figure 18: Comparison of estimated and observed mean annual streamflow using regression with area, precipitation and temperature as predictor variables.

All four graphs count with a high R^2-value. The points in the graphs are spread out along the line which means that there is a small difference between the estimated and the observed points and no outliers are present. Even though the model for Norte de Santander has the best R^2 it did acquire a high VIF and therefore it cannot be considered as a trustworthy model.

Considering all the results for regression analysis with area, precipitation and temperature only five models can be considered trustworthy as they account for good R^2 and VIF-value as well as good histograms. These models represent the regions named the entire area, the 28 river basins, Santander, G6HHR2 and G7HHR1.

5.1.3 Regression Analysis using the Most Suitable Variables as Predictor Variables

The results fro m the third and last model with most suitable variables is presented in table 9. The following equation was used to formulate this model:

$$\mu_Q = a + b(1st\ predictor) + c(1st\ predictor) + d(1st\ predictor) \tag{22}$$

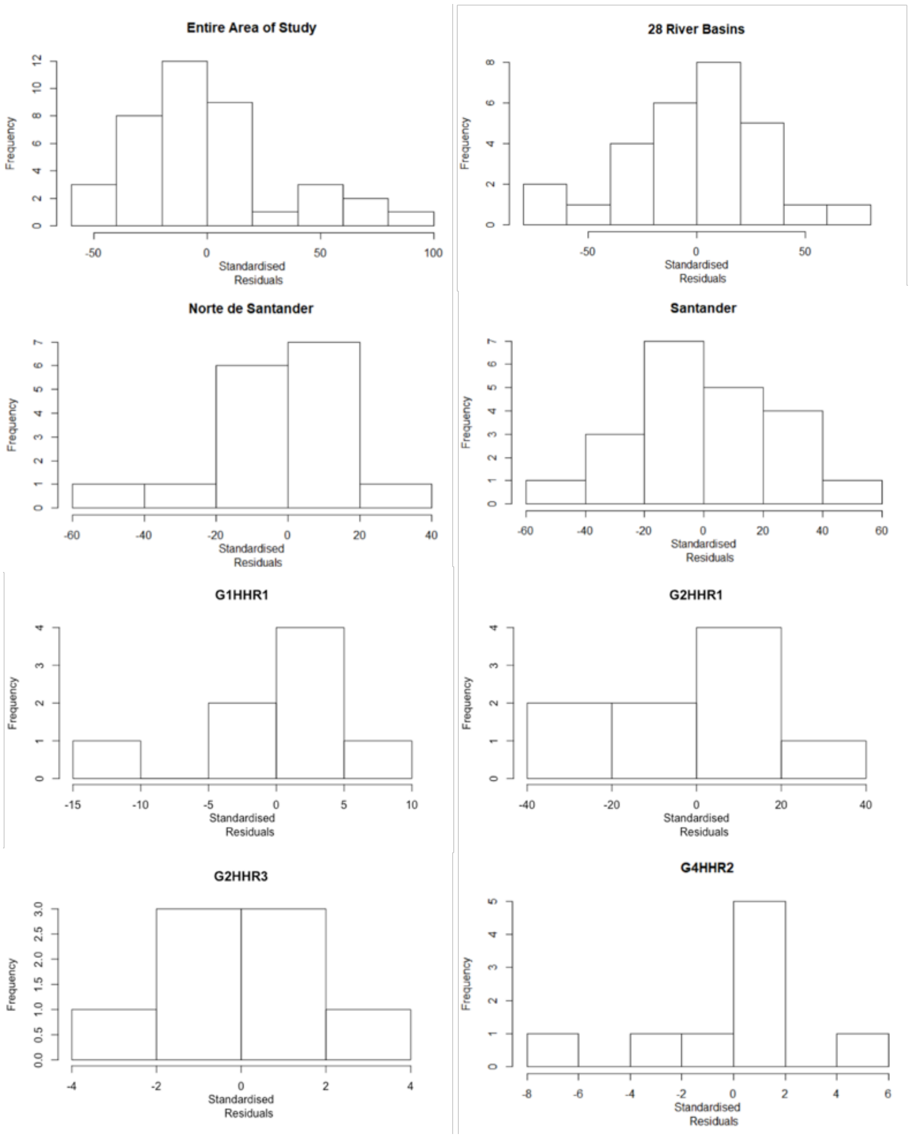
where μ_Q is MAF, a is the intercept, b, c and d are constants. The predictor variables were decided according to table 9.

Table 9: Regression Models for Mean Annual Streamflow using most suitable variables as predictor variables.

Region	**Intercept (a)**	**b**	**c**	**d**	**Adjusted R^2**	**Maximum VIF**
Entire Area of Study	-78.0292	A 0.0296	P_{mar} 0.288	μ_T 2.396	0.88	1.159
28 River Basins	-138.157	A/Per 29.255	P_{may} 0.307	-	0.861	1.00341
Norte de Santander	-44.894	A 0.0361	P_{mam} 0.247	-	0.913	1.0160
Santander	-219.680	A 0.0334	P_{oct} 0.516	S_{med} 1.654	0.921	1.106
G1HHR1	-118.456	A/Per 16.640	No. d. P_{oct} 4.573	-	0.941	1.000
G2HHR1	-208.554	A/Per 33.867	P_{mar} 0.811	-	0.911	1.000
G2HHR3	-28.008	A 0.0413	P_{mam} 0.0713	No. d. P_{may} 0.599	0.993	3.202
G3HHR1	-103.565	A 0.0386	P_{mar} 0.629	-	0.959	1.0551
G4HHR2	-61.903	A 0.0409	No. d. P_{amon} 3.0416	-	0.987	1.0780
G5HHR1	-11.644	A/Per 4.236	Max 24h P_{may} 0.0107	-	0.991	1.0185
G5HHR2	-27.554	A 0.0352	P_{mam} 0.159	-	0.918	1.239
G5HHR4	-368.600	A/Per 39.720	P_{apr} 1.0400	-	0.944	1.00137
G6HHR2	-30.184	A 0.0399	P_{mam} 0.140	-	0.922	1.0377
G7HHR1	-36.486	A 0.0418	P_{may} 0.0656	No. d. P_{ave} 1.114	0.982	1.531
G7HHR2	-198.555	A 0.0299	P_{rs} 1.0989	Max 24h P_{apr} -0.338	0.993	1.412
G8HHR1	43.833	A 0.0406	Max 24h P_{apr} -0.240	S_{med} -0.340	0.991	1.412

The majority of the models need only two predictor variables to show the best possible fit, one that explains the size of the region and one that explains the precipitation regime. Both the adjusted R^2 and the maximum VIF show highly acceptable values which signifies that the predictor variables can be used to explain most of the variance in MAF.

The acceptable residual histograms for the models are displayed in figure 19. The histogram for HHR G4HHR2 is more questionable than the other ones, but along with its high R^2-value it was still decided that it could be seen as acceptable. Out of the 16 models only two were estimated not acceptable and can be seen in appendix II.


Entire Area of Study
Frequency
0
2
4
6
8
10
12
-50
0
50
100
Standardised Residuals
28 River Basins
Frequency
0
2
4
6
8
-50
0
50
Standardised Residuals
Norte de Santander
Frequency
0
1
2
3
4
5
6
7
-60
-40
-20
0
20
40
Standardised Residuals
Santander
Frequency
0
1
2
3
4
5
6
7
-60
-40
-20
0
20
40
60
Standardised Residuals
G1HHR1
Frequency
0
1
2
3
4
-15
-10
-5
0
5
10
Standardised Residuals
G2HHR1
Frequency
0
1
2
3
4
-40
-20
0
20
40
Standardised Residuals
G2HHR3
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-4
-2
0
2
4
Standardised Residuals
G4HHR2
Frequency
0
1
2
3
4
5
-8
-6
-4
-2
0
2
4
6
Standardised Residuals

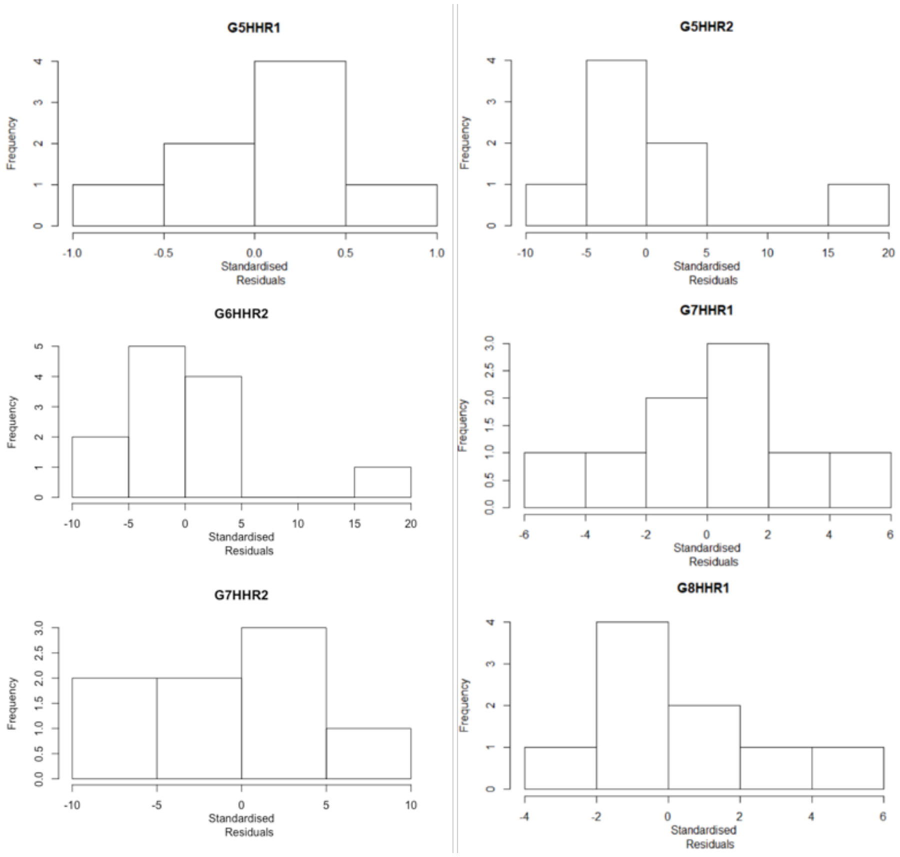


Figure 19: Residual histograms which are seemingly normally distributed. Obtained from the regression models with most suitable variables as predictor variables.

Below the fit between observed and estimated MAF is displayed in figure 20.

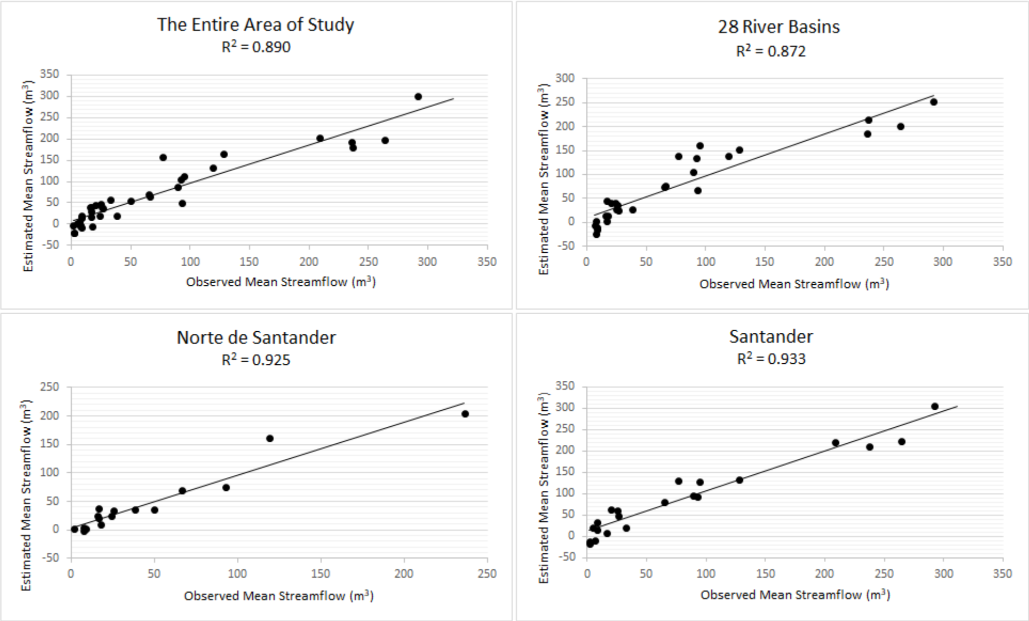


Figure 20: Comparison of estimated and observed mean annual streamflow using regression with the most suitable predictor variables.

The comparison indicates that the models can be used to explain the observed values well. The points in the graphs are spread out along the line therefore no outliers are present.

The results for the third model-type implies that 14 regression analysis form trustworthy models considering the high R^2, the low VIF-value and the normal distribution of the residual histograms. The two models not considered satisfactory due to their histograms are G3HHR1 and G5HHR4.

In general the results show that models using variables which have not been chosen by trial and error, to reach the best possible fit, perform better for larger areas. For smaller areas it proved difficult to reach satisfactory histograms. Models for larger areas also perform well when choosing the most suitable variables through trial and error but for this case it is easier to reach a high R^2-value for the smaller hydrologically homogeneous regions. Something that works well on a large scale does not necessarily work as well on a smaller scale. This is important to keep in mind when working with models of this kind.

Further it can be said that the third model type with most suitable variables result in the best fit, although both the first and the second model types produce fairly well estimated values when compared to the observed ones. Regarding the catchments that showed tendencies for changes, no difference could be observed in the model performance between these and the catchment with no tendencies.

5.2 Regression Analysis of Design Flow

In this section the obtained results from regression analysis for design flow in Norte de Santander and Santander will be presented. Adjusted R^2 and maximum VIF will be presented in table 10 along with the coefficients for the predictor variables area, slope and elevation. Thereafter the residual histograms will be presented for the models that showed best results according to table 10. Finally the goodness of fit index of the model will be analyzed by presenting graphs where calculated design flow, from the Gumbel distribution plot, has been plotted against estimated design flow from the model.

In general it can be said that the results in table 10 shows that all of the four different regression models can be of use since all R^2-values are above 0.65 and all the VIF-values are well below five. Observe the identical VIF-values for the models with same predictor variables but different design flow. This is not a coincidence for the reason that the VIF-value evaluates the relationship between the variables which of course is equal in all four models.

It can also be observed that the most creditable models, according to R^2, are the ones including all three of the predictor variables. Additionally the models perform better for lesser extreme flows i.e. shorter return period, although the difference is no more than an increase in approximately three percentage units from Q_{10} to Q_{100}.

Table 10: Regression models for design flow using area, slope and elevation as predictor variables.

Variables	**Intercept (a)**	**Area (b)**	**Slope (c)**	**Elevation (d)**	**Multiple R^2**	**Adjusted R^2**	**Maximum VIF**
Q_{10}							
Area	363.836	0.129	-	-	0.707	-	-
Area and Slope	761.980	0.106	-87.750	-	-	0.745	1.434
Area and Elevation	718.875	0.134	-	-0.189	-	0.725	1.026
Area, Slope and Elevation	974.387	0.113	-75.753	-0.142	-	0.757	1.536
Q_{20}							
Area	418.931	0.141	-	-	0.696	-	-
Area and Slope	864.2303	0.115	-98.143	-	-	0.735	1.434
Area and Elevation	811.777	0.146	-0.209	-0.209	-	0.713	1.026
Area, slope and Elevation	1098.2301	0.122	-84.926	-0.157	-	0.747	1.536
Q_{50}							
Area	490.459	0.156	-	-	0.683	-	-
Area and Slope	996.816	0.127	-111.600	-	-	0.724	1.434
Area and Elevation	931.978	0.162	-	-0.235	-	0.7003	1.026
Area, Slope and Elevation	1258.540	0.135	-96.817	-0.175	-	0.735	1.536
Q_{100}							
Area	550.6302	0.167	-	-	0.675	-	-
Area and Slope	1104.380	0.134	-122.0449	-	-	0.718	1.434
Area and Elevation	1027.847	0.173	-	-0.254	-	0.692	1.026
Area, Slope and Elevation	1385.873	0.143	-106.145	-0.188	-	0.728	1.536

It was found that all histograms for all four return periods were seemingly acceptable. The residual histograms using all three variables are displayed below, the rest can be found in Appendix II.

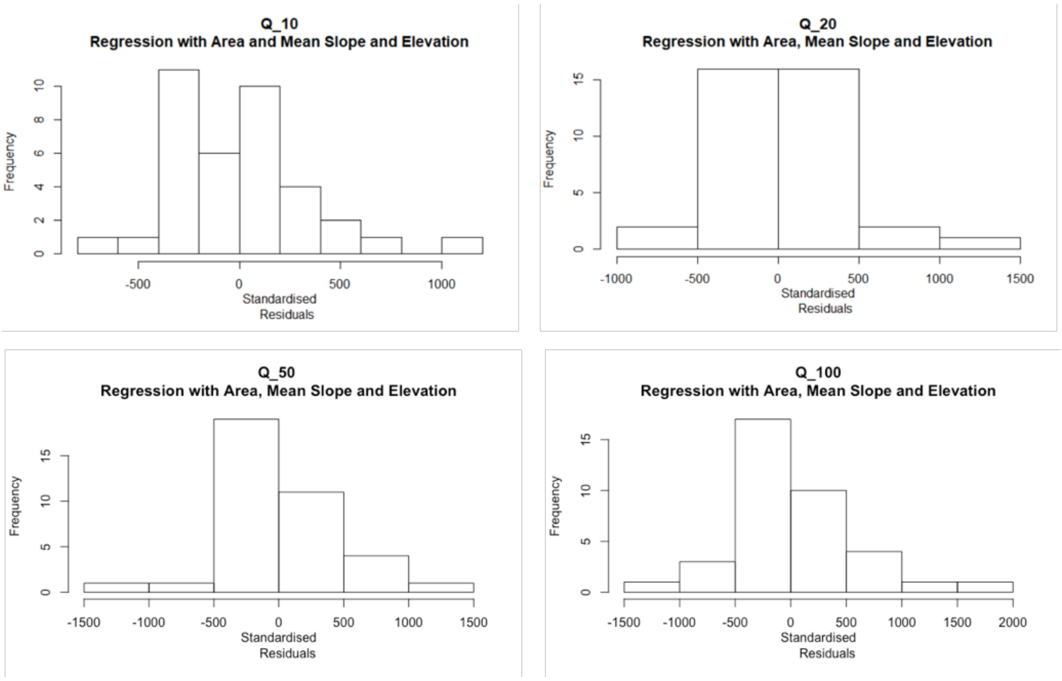


Figure 21: Residual histograms using all three predictor variables watershed area, mean basin elevation and slope. The residual histograms seem normally distributed for all four return periods.

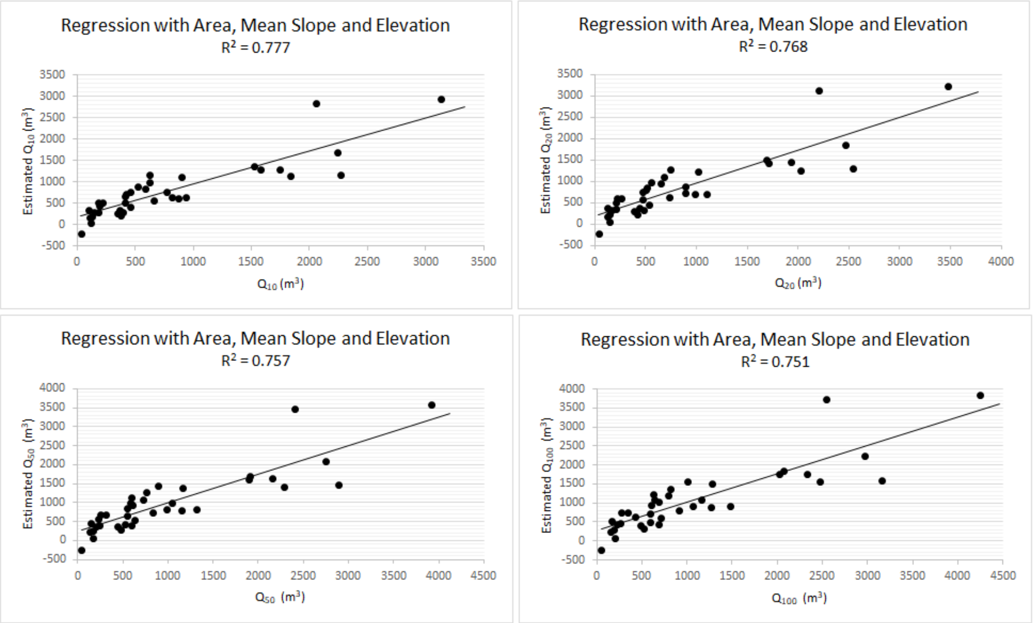


Figure 22: Comparison of calculated and estimated design flow using regression with area, mean slope and mean elevation. The graph at the upper left is for 10-year flows, the upper right is for 20-year flows and the the bottom left and bottom right are 50-year and 100-year flows respectively.

Figure 22 shows results that are consistent with the ones in table 10 above. The R^2-value decreases with higher return period although all are above 0.7 which indicates a good fit. The points in the graphs are spread out along the line and therefore no outliers are present.

6 Further Discussion

The global climate is changing and the need for a sustainable water resources management is therefore of great importance. However, as mentioned in the introduction monitoring of watersheds has decreased since the 1980s. These two factors combined imply that the results of this study should be of great use in water resources management and planning. Looking at Colombia specifically the objectives of MinAmbiente require knowledge about water supply, which could be acquired using the results in this study. Further these models could be used as preliminary case studies in ungauged watersheds as they disclose the behavior of the streamflow. This could be beneficial for longterm planning and general knowledge. A more precise example for a feasibility study would be to examine monthly water availability in a watershed which could be helpful in managing water supply and irrigation for a region.

This study has resulted in regression models which can be used to estimate MAF and design flow in an area where a tropical climate dominates. According to the statistical criteria decided on in this study, the models can be evaluated as reliable. This could be seen as an indication that it is possible to formulate regression models for other areas with tropical climates, which had not been done previous to this study in Colombia. As mentioned in the text many possible statistical tests could be applied in order to decide whether a model should be seen as reliable or not. Several tests should be used in combination as no decisions can be made by evaluating each statistical test by itself.

The results for MAF depict the importance of using several statistical tests to analyze the reliability of a model. If one would only rely on the R^2- and VIF- variables, which are the ones presented in the tables above in the results, one can be deceived to believe that the models are better than they actually are. This can be seen by examining the residual histograms, also presented in the result, which diminish the number of acceptable models as they do not fulfill the requirement of a normal distribution. This implies that the assumption about the independence and the normal distribution for the error of the model cannot be proven and therefore a regression model cannot be accepted as is explained in section 3.1.

6.1 Uncertainties

One of the principal uncertainties when working with hydrological data is measurement errors. As mentioned in section 2.1 measurement stations use varying techniques and are of varying quality. A requirement for including the absolute maximum values in the analysis was that 90 % of daily streamflow was registered. Despite this it was noticed that data for several years seemed questionable. Some months had many days where the same measurement value was repeated many times in a row, which seems very unlikely for something random like streamflow. However, this was data obtained from IDEAM which is a Colombian governmental institution. It was necessary to assume that data obtained from them is reliable.

As mentioned previously, this study relies heavily on data obtained from a previous study of Salazar (2016). He also obtained data sets from IDEAM which he further analyzed and performed calculations on. Many of the resulting calculations were used in this study and it was assumed that they were correct.

Yet another uncertainty was the fact that the values for maximum precipitation in 24 hours and number of days with precipitation were extracted from maps. No shapefiles existed for these maps and they could therefore not be analyzed using ArcGIS. The map images were used as pdf-files and the shape of the study area with the watersheds outlined was overlayed and the values simply estimated by looking at it. This can, of course, not be seen as a very precise method but it was the best possible considering the access of data.

6.2 Analysis of Mean Annual Streamflow

The results from the obtained regression models in section 5.1 show a similar trend as the results from the study performed by Vogel et al. (1999) in the United States of America. This analysis was similar to the one executed in this study and is explained briefly in section 1.1.1 alongside with some of its results. In both studies the models are improved when more variables are added in order to explain the streamflow. Both clearly show that multiple regression with most suitable variables formulate the most reliable models.

The results from the two studies can be compared mainly by examining the goodness of fit index (R^2) and the VIF-value. When comparing the goodness of fit index it is found that the study made in the USA in general shows a better value. Another observation that can be made is the number of variables used in order to obtain the best fit; in this study a maximum amount of three variables were used and for the study in the USA this number reached five. This can explain the better outcome in the adjusted R^2 - value in the study by Vogel et al. seeing that the coefficient has a tendency to increase when more parameters are included into the model. Another aspect is the VIF-value which in the study by Vogel et al. in general is higher than for this study, although all are under ten and the majority are under five. What can be said about this is that Vogel et al. probably allowed a higher VIF-value in order to obtain a higher R^2-value. In this study the focus was rather on retrieving a model that would have the best fit according to all of the statistical variables.

A big difference between the two studies is the size of the study area and the amount of data used. This analysis includes only two departments in Colombia which is a small area in comparison to the entire USA which is included in the study by Vogel et al. Further the two departments are divided into smaller HHR containing a small amount of measurement stations. At first it was believed that the smaller areas would lead to better models due to their homogeneity and small size. The results from the first and third model, i.e. with only area and with most suitable variables, does in general produce improved results for the homogeneous groups than for the bigger areas according to the R^2-value. This can be explained by considering that the smaller groups are grouped according to homogeneity instead of only according to the watershed divides, as has been done for the whole departments.

It is worth remembering that the departments are characterized by very different geomorphological and climatological differences within themselves which could explain why the models for most suitable variables perform best for the HHR. These models take into consideration the precipitation regime which was not equally distributed in a department.

As mentioned above the model that seemingly performed the best is the one that includes the most suitable variables. In the majority of the models the number of variables are two; one that explains the size of the region and one that explains the precipitation. Information about the area can be considered as easily accessible, although it is questionable if the same can be said about the precipitation due to the fact that this variable is not used as mean annual precipitation for all regions, instead it is unique for all of them. It could be precipitation for a specific month or the average for only some months that are included in the model which indicates that more work would be necessary to obtain the correct information.

If the above would not be possible the other two models, the first where only area is explanatory variable and the second where area, precipitation and temperature are explanatory variables could be of use. These may not be as reliable as the results show a lower R^2-value, higher VIF and weaker residual histograms although for the bigger regions the results from these statistical test's were in general considered acceptable. Continuously the comparison between estimated and observed mean annual streamflow showed an acceptable fit. Therefore it can be assumed that these models can be used with caution. Regarding the smaller groups it is doubtful whether or not the same can be said, owing to the weak residual histograms for the majority of the groups.

What can be said about these findings is that area and precipitation can explain the larger part of variance in the streamflow therefore it would be interesting to evaluate how well a model would perform if only area and precipitation were included as explanatory variables. This would probably not produce a better model than for the one with most suitable variables. Nevertheless, it could increase the model sufficiently for it to be considered useful which would imply less workload acquiring correct data.

Ultimately, in this study two different models were made where the whole area was considered, more specifically one that took linear tendencies into consideration resulting in nine regions being removed, and another where all 37 regions were included. The two models did not differentiate considerably in their outcome, subsequently this indicates that the anthropogenic influences in the nine regions did not affect the model performance. This can be strengthen by the fact that no difference was observed between the models performed for the HHR where some contained catchments with tendencies and other not. Nonetheless this study cannot be considered as big enough to draw any definite conclusion about this.

6.3 Analysis of Design Flow

When comparing the results for the model outcome for design flow with the analysis made by Dingman and Palaia in 1999 it can be seen that they follow the same trend. The ability to predict design flow diminishes with higher return period which can probably be explained with the reason that a ten-year flow is more frequent and more similar to the ordinary flows than a 100-year flow. It can also be seen that the models for Colombia yield a lower R^2-value than the one for the USA. Some possible explanations for this could be that regression models are not as suitable in the tropical areas of Colombia or that the heterogeneity in geomorphology is unsuitable for these models. Another explanation could be that the maximum streamflow would have been better described with another distribution than the Gumbel distribution. A fourth possible reason could be that the choice of variables to include in the analysis were not the most appropriate.

On the other hand all of the models are considered to be of use being that the R^2- and the VIF-values as well as the graph of calculated vs. observed design flow all can be accepted according to the requirements of this study. Although the models should be used with caution and with the reason of guidance and not for decision making.

Even though this is an analysis about design flow and therefore extreme flow, extreme events such as la niña and ENSO, that often result in an increase or decrease in rainfall which naturally would increase or decrease the streamflow, are not consciously accounted for in this study. Outliers that are detected in the analysis could be a result of this phenomenon, however this requires further research which goes beyond the objectives of the present study. As one of the purpose of the models in this study is to manage hydraulic infrastructure such as small dams or bridges it can be considered that they would not just benefit but in fact require information about the extreme flow resulting from a phenomenon such as la niña in order for it to be able to withstand such floods. Of course this is very difficult due to the fact that the event does not occur regularly nor with a constant force. This complicates the possibility to include la niña in a model which is why it is not incorporated into this study.

6.4 Possible Improvements and Future Studies

As mentioned earlier, according to The World Meteorological Organization (Barbarossa et al. 2017) climatic analysis should be executed with data sets of 30 years. This is something that could be improved in this study where 20 years of data was used. Due to time limitations, requesting additional records apart from those available from the study of Salazar was not an option. Longer data records that contained information about the more recent years could increase the reliability of the model.

Regarding the analysis of design flow it is commonly assumed that the statistical distribution of extreme values in Colombia can be described according to the Gumbel distribution. However, this has not been proven. A possible improvement for this study is therefore to further study this distribution to assure that the Gumbel distribution is in fact the best choice.

Another improvement in regard to the design flow models is, as mentioned in section 6.3, to obtain channel width. This would allow for better possibilities of comparing with the study of Dingman and Palaia (1999). Other drainage basin and channel characteristics would also be interesting to evaluate in order to produce a better model.

As was discussed in section 6.2 differences in the resulting regression models for MAF could be seen depending on the size of the regions analyzed. However, if this was actually due to size or if it was a result of the smaller regions being divided according to their hydrological homogeneity cannot be concluded from the results. What can be said is that the watersheds used when creating the HHRs are quite small on a national level (small river systems). It would be interesting, as a future study, to create HHRs using larger watersheds (larger river systems) for example for the 17 catchments that IDEAM has divided the country into. This would also increase the ability to compare the results with the once obtained by Vogel et al. Further the influence of the homogeneity could be compared with areas of the same size. Nonetheless, the size of the watersheds used should always be chosen in agreement with the aim of the study. For example there is a difference in analyzing a large river basin such as the Amazon or a smaller basin of a watercourse at the source of the Amazon which would imply a difference in model size. It could be beneficial for the country to acquire a national model for an overview in the national water resources management. However as mentioned in section 2.1 it is the responsibility of each department to fulfill the requirements of MinAmbiente therefore regression models on a more regional scale would be beneficial.

Considering the study of Barbarossa (2017) it could be interesting to perform a similar, global study merged with the methods used in this thesis. Instead of creating just one model for the whole world, HHRs could be created on a global scale and regression models created for each of these regions. This should result in a somewhat more reliable model as it would take climatic differences into account. To increase reliability even more creating models on a continental level could be a good idea. Concerning Colombia it could be beneficial for the country to acquire a national model for an overview in the national water resources management. However as mentioned in section 2.1 it is the responsibility of each department to fulfill the requirements of MinAmbiente therefore regression models on a more regional scale would be beneficial.

7 Conclusions

In this study the possibility to formulate reliable regression models for a tropical region in Colombia was examined for mean annual streamflow and design flows. Different characteristics were investigated as predictor variables, different sizes and compositions of area were put together for streamflow and different return periods were analyzed for design flow in order to evaluate what type of model would be appropriate for this tropical region. The reliability of the models was evaluated by the use of different statistical tests.

To conclude this study in the most clear way possible the objectives will be answered one by one in the order they are stated in section 1.2.

1. According to the criteria set up in this study it is possible to formulate a reliable regression model relating mean annual streamflow to geomorphic and climatic characteristics. It can be concluded that testing different variables to achieve the best possible fit results in the most reliable models. The conclusions to the sub-objectives stated below will therefore be based on these results.

   (a) According to the results presented in section 5.1 it seems like the models for the smaller HHRs perform better than those including the whole study area or the individual departments. However, it is not possible to draw any clear conclusions from this. It could be that analyzing smaller areas result in more reliable models. On the other hand the reason could be that dividing the study area into HHRs gives better results, or it could be a combination of the two; small areas grouped into HHRs. Further studies are necessary to evaluate this.

   (b) Looking at the models created with the aim to create the best possible model the characteristics included in order to give the best possible fit of the model vary between different regions, as could be expected. In almost all regions studied, including the area as one of the predictor variables is important. Apart from this including one or two climatic variables, most often some variety of precipitation, gives the best possible fit.

   (c) As mentioned earlier many different statistical tests could be employed in order to evaluate the reliability of a model. It is simply a matter of convenience and habits of the user which tests are chosen. However, it is important to apply several tests to assure reliability.

2. According to the criteria set up in this study it is possible to formulate a reliable regression model relating design flow to drainage-basin characteristics.

   (a) Examining the results presented in section 5.2 one can conclude that the models including all three basin characteristics area, elevation and slope produce the best models.

   (b) As the same statistical tests were used to evaluate the models of both parts of this thesis the same conclusion as stated above in 1. (c) also answers this objective.

Some additional conclusions that can be drawn from this analysis concerning mean annual streamflow is that bigger areas than the HHR are necessary in order to obtain a reliable model using only area as predictor variable. The same can be said for when area, precipitation and temperature are used as predictor variables.

Concerning design flow an acceptable model was obtained although the R^2-values leaves an indication that there is room for improvements.

8 References

Barbarossa, V., Huijbregts, M.A.J., Hendriks, A.J., Beusen, A.H.W., Clavreul, J.,King, H., d, Schipper, A.M. (2016) Developing and testing a global-scale regression model to quantify mean annual streamflow. *Journal of Hydrology*, 544 (2017), pp. 479–487.

del Ama, B. (2013) *Colombia: Urban Past, Rural Future?.* [Online] Available at: http://www.cnbc.com/id/100876430 [Accessed 23 October 2017].

Diez Diaz, J. (2017) *Solicitud de cita. Radicado e1-2017-028656.* [e-mail] (Personal communication 14 November 2017).

Dingman S.L. and Palaia K.J. (1999). Comparison of Models for Estimating Flood Quantiles in New Hampshire and Vermont. *Journal of the American Water Resources Association.*, 35 (5), pp. 1233-1243

Dooge, J. C. I. (1992). Sensitivity of runoff to climate change: A Hortonian approach. *Bull. Am. Meteorological Soc.*, 73(12), pp. 2013-2024.

Gayathri, K. D., Ganasri B. P., Dwarakish G. S. (2015). A Review on Hydrological Models. *Aquatic Procedia.* 4 (2015), pp. 1001-1007

Gobernacion de Norte de Santander. (2017) *Información General Norte de Santander.* [online] Available at:
http://www.nortedesantander.gov.co /Gobernaci%C3%B3n/NuestroDepartamento /Informaci%C3%B3n-Gen eral-Norte-de-Santander [Accessed: 03-October 2017]

Gónima, N. (2014). *Bogotá podría enfrentar crisis de agua en diez años.* [online] Available at: https://www.elespectador.com/noticias/bogota/ bogota-podria -enfrentar-crisis-de-agua-diez-anos-articulo-510892 [Accessed 24 Oct. 2017].

Google Earth Pro 7.3.0.3832. (2017). Colombia 7^{o}35'51.78"N, 73^{o}0143.17"W elevation 1899m [Accessed 15 October 2017]

Goovearts, P. (1999). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology.* 228 (2000), pp. 113-129

Guzmán D., Ruíz, J. F., Cadena M. (2014). *Regionalización de Colombia según la estacionalidad de la precipitación media mensual, a través análisis de componentes principales (ACP).* Grupo de Modelamiento de Tiempo, Clima y Escenarios de Cambio Climático Subdirección de Meteorología, IDEAM.

IDEAM. (2017a). *Geovisor* [online] Available at: visor.ideam.gov.co:8530/geovisor [Accessed 20 =CT. 2017]

IDEAM. (2017b). *Atlas Climatológico de Colombia* [online] Available at: http://atlas.ideam.gov.co/visorAtlasClimatologico.html [Accessed 23 Oct. 2017]

IDEAM. n.d. *Fénomeno El Niño* [online] Available at: http://www.ideam.gov.co/ web/tiempo-y-clima/clima/fenomenos-el-nino-y-la-nina [Accessed 02 January 2018]

Investopedia. n.d. *What's the difference between r-squared and adjusted r-squared?* [online] Available at: https://www.investopedia.com/ask/ answers/ 012615/whats-difference-between-rsquared-and-adjusted-rsqu ared.asp [Accessed 10 December 2017]

Lantmäteriet. n.d. *WGS 84.* [online] Available at: https://www.lantmateriet. se/Kartor-och-geografisk-information/GPS-och-geodetisk-matning/ Referenssystem/Tredimensionella-system/WGS-84/ [Accessed 18 October 2017]

Larsson, K. (2013).*How to... in ArcGIS.* Department of Physical Geography and Ecosystem Analysis & GIS Centre, Lund University. [pdf]

Mathwave. (n.d.). *Extreme Value Distributions.* [online] Available at: http://www.mathwave.com/articles/extreme-value-distributions.html [Accessed 05 February 2018]

Mohamoud, Y.M., & Parmar, R.S. (2006) Estimating Streamflow and Associated Hydraulic Geometry, the Mid-Atlantic Region, USA. *Journal of the American water resources association*, pp. 755-768.

Nationalencyklopedin. n.d. *Colombia.* [online] Available at: http://www.ne.se /uppslagsverk/encyklopedi/lång/colombia [Accessed 03 October 2017]

R-project. (2017). *R* (version 3.4.2) [software]. Available at: https://cran.r-project.org/mirrors.html

R-project.org. (2017). *R: The R Project for Statistical Computing.* [online] Available at: https://www.r-project.org [Accessed 17 October 2017]

Salazar Oliveros, J. C. (2016). *Una Metodología para la Estimación de Curvas de Duración de Caudales (CDC) en Cuencas No Instrumentadas. Caso de Aplicación en Colombia en los Departamentos de Santander y Norte de Santander*, Master Thesis, Universidad Nacional de Colombia. [pdf] Available at: http://www.bdigital.unal.edu.co/56437/20 /JuanC.SalazarOliveros.2016.pdf [Accessed 16 October 2017]

Shewhart, W.A., & WILKS, S.S. (ed).(2006). *Regression Analysis by Example.* Volume 4. New Jersey: John Wiley & Sons, Inc.

Siac. n.d. *Fenómenos del Niño y la Niña* [online] Available at: http://www.siac.gov.co/ninoynina [Accessed 02 January 2018]

Stratfor, (2016). *In Colombia, Abundant Water Brings No Security.* [online] Available at: https://worldview.stratfor.com/article/colombia-abundant-water-brings-no-security [Accessed 11 September 2017]

Un.org. (2014). Human right to water and sanitation | International Decade for Action 'Water for Life' 2005-2015. [online] Available at: *http : //www.un.org/waterforlifedecade/human_right_to_water.shtml* [Accessed 23 Oct. 2017].

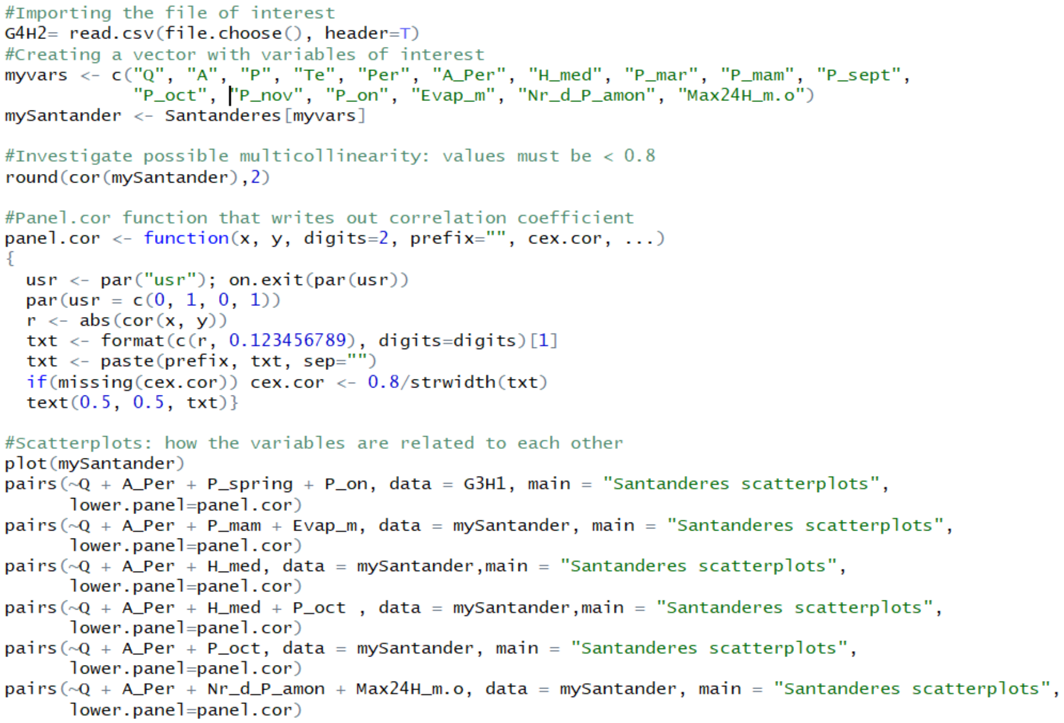
U.S. Department of the Interior Geological Survey (USGS). (1981). *Guidelines for Determining Flood Flow Frequency* (Bulletin #17B). Available at: *https : //water.usgs.gov/osw/bulletin17b/dl_flow.pdf* [Accessed 12 December 2017]

Villareal González, E. Personal meeting at Universidad Nacional de Colombia, 5 November 2017.

Vogel, R. M., Wilson, I. and Daly, C., (1999). Regional Regression Models of Annual Streamflow for the United States. *Journal of Irrigation and Drainage Engineering*, 125 (3), pp. 148-157.

Appendix I - R-code

The code used in order to perform the statistical tests in R is displayed below.



```
#Importing the file of interest
G4H2= read.csv(file.choose(), header=T)
#Creating a vector with variables of interest
myvars <- c("Q", "A", "P", "Te", "Per", "A_Per", "H_med", "P_mar", "P_mam", "P_sept",
            "P_oct", "P_nov", "P_on", "Evap_m", "Nr_d_P_amon", "Max24H_m.o")
mySantander <- Santanderes[myvars]

#Investigate possible multicollinearity: values must be < 0.8
round(cor(mySantander),2)

#Panel.cor function that writes out correlation coefficient
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt)}

#Scatterplots: how the variables are related to each other
plot(mySantander)
pairs(~Q + A_Per + P_spring + P_on, data = G3H1, main = "Santanderes scatterplots",
      lower.panel=panel.cor)
pairs(~Q + A_Per + P_mam + Evap_m, data = mySantander, main = "Santanderes scatterplots",
      lower.panel=panel.cor)
pairs(~Q + A_Per + H_med, data = mySantander,main = "Santanderes scatterplots",
      lower.panel=panel.cor)
pairs(~Q + A_Per + H_med + P_oct , data = mySantander,main = "Santanderes scatterplots",
      lower.panel=panel.cor)
pairs(~Q + A_Per + P_oct, data = mySantander, main = "Santanderes scatterplots",
      lower.panel=panel.cor)
pairs(~Q + A_Per + Nr_d_P_amon + Max24H_m.o, data = mySantander, main = "Santanderes scatterplots",
      lower.panel=panel.cor)
```

```
#Regression model with A
results1 = lm(Q ~ A, data=G4H2)
results1
summary(results1)

#Cheking that residuals are normally distributed.
hist(resid(results1),main='Histogram of residuals',xlab='Standardised
     Residuals',ylab='Frequency')
plot(results1, which = 1)

#Plotting the confidence bounds on a scatterplot
plot(G4H2$A, G4H2$Q, pch = 18)
resultsP = lm(Q ~ A, data=G4H2)
abline(resultsP, col = "red")
newX = seq(min(G4H2$A), max(G4H2$A), 1)
prd.CI = predict(resultsP, newdata = data.frame(A = newX),
                 interval = "confidence", level = 0.95)
lines(newX, prd.CI[, 2], col = "blue", lty = 2)
lines(newX, prd.CI[, 3], col = "blue", lty = 2)
prd.PI = predict(resultsP, newdata = data.frame(A = newX),
                 interval = "prediction", level = 0.95)
lines(newX, prd.PI[, 2], col = "green", lty = 3)
lines(newX, prd.PI[, 3], col = "green", lty = 3)

#Model with A, P, Te
results2 = lm(Q ~ A + P + Te, data=G4H2)
results2
summary(results2)

#checking collinearity with VIF. Values should be close to 1 but under 5.
#values >10 indicate that the variable is not needed and can be removed
library(car)
vif(results2)

#Cheking that residuals are normally distributed.
hist(resid(results2),main='Histogram of residuals',xlab='Standardised
     Residuals',ylab='Frequency')
plot(results2, which = 1)
```

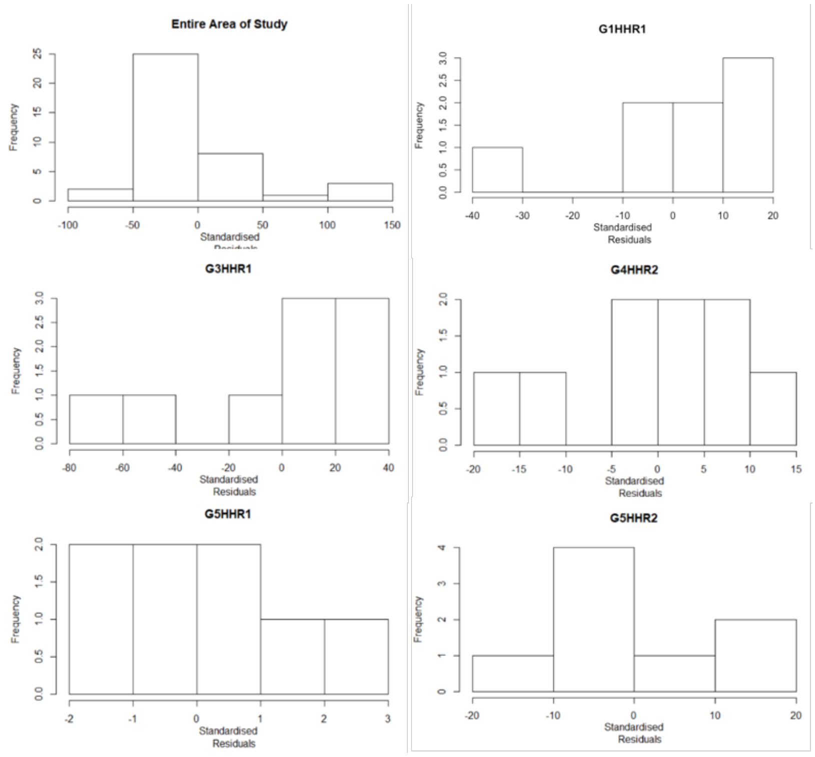
```
#Model with other variables
results3 = lm(Q ~ A + Nr_d_P_amon, data=G4H2)
results3
summary(results3)

#checking collinearity with VIF. Values should be close to 1 but under 5.
#values >10 indicate that the variable is not needed and can be removed
library(car)
vif(results3)

#Cheking that residuals are normally distributed.
hist(resid(results3),main='Histogram of residuals',xlab='Standardised
     Residuals',ylab='Frequency')
plot(results3, which = 1)
```

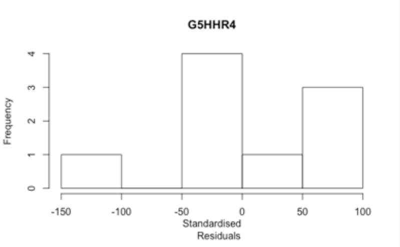
Appendix II - Residual Histograms

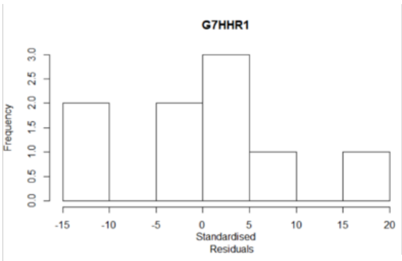
This appendix presents the residual histograms for mean annual streamflow which were considered to not fulfill the requirements of being normally distributed. It was therefore decided that the models for these regions could not be counted on as trustworthy.

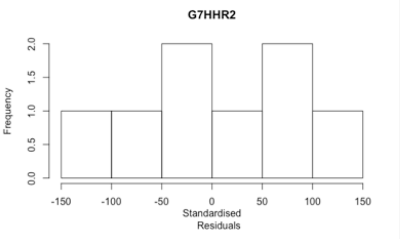
Area only

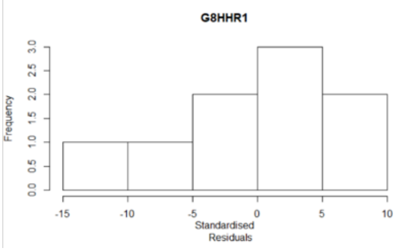
The histograms for the following regions were considered to not meet the criteria when area only was used as predictor variable.





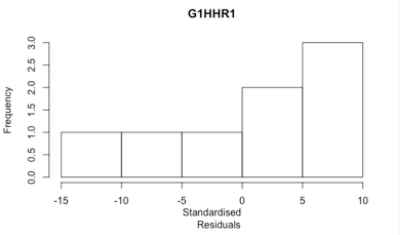


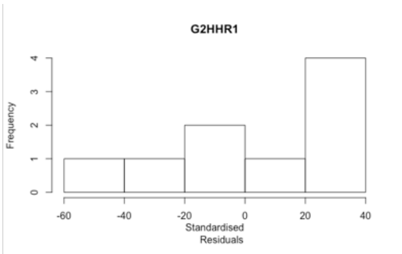


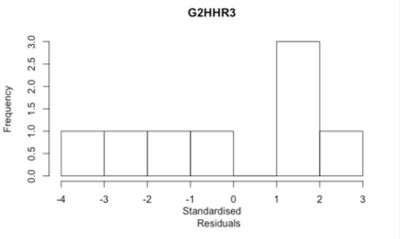


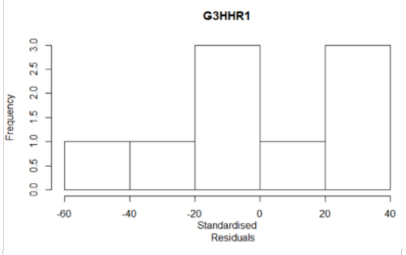
Area, Precipitation and Temperature

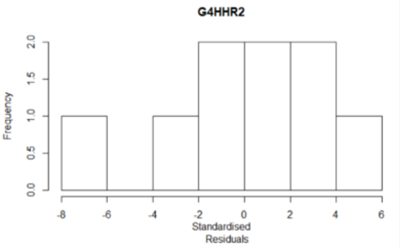
The histograms for the following regions were considered to not meet the criteria when area, precipitation and temperature were used as predictor variables.

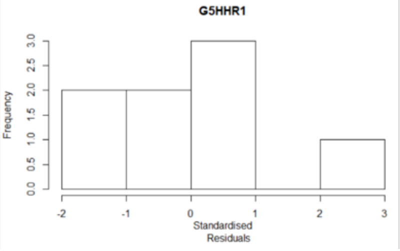


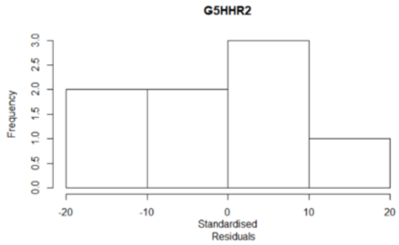


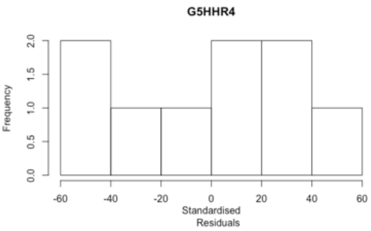


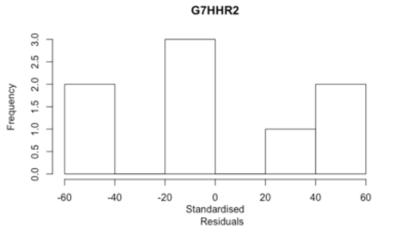


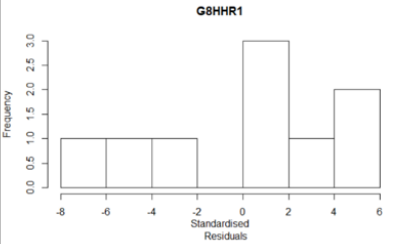






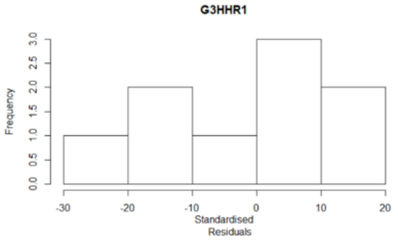


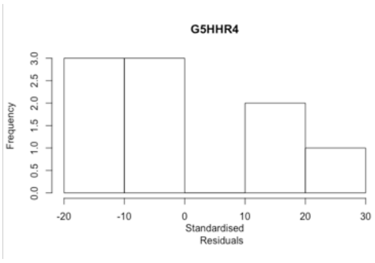




Most Suitable Variables

Only the histograms for the following two regions were considered to not meet the criteria when models were created using the most suitable predictor variables.



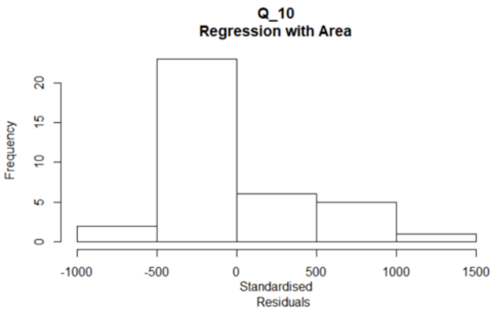


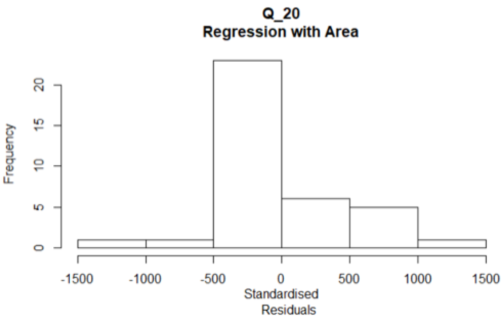
Design Flow

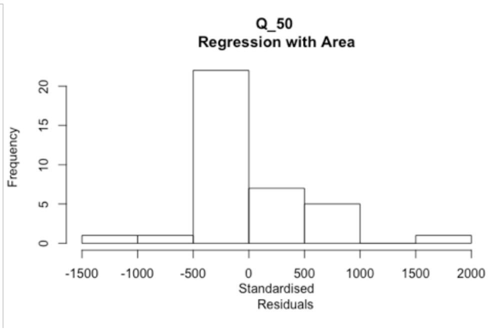
Regarding design flow all models were considered to meet the criteria. The histograms below belong to the models with one or two predictor variables.

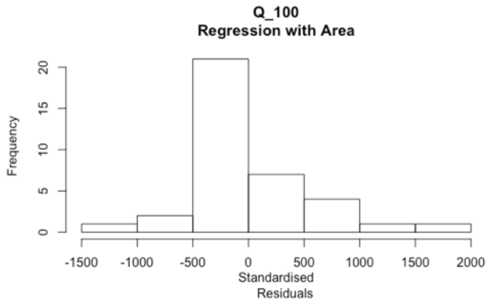
Area Only

The histograms below were obtained from performing a regression analysis using only watershed area. The histograms seem normally distributed for all four return periods.


Q_10
Regression with Area

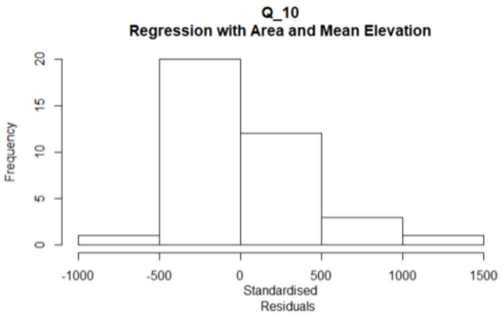

Q_20
Regression with Area

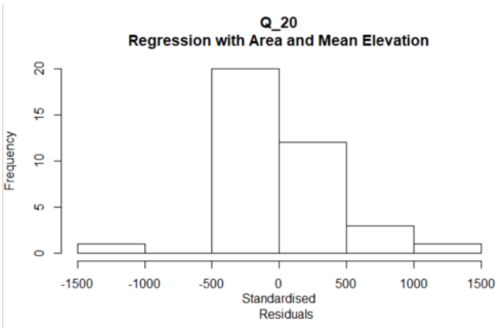

Q_50
Regression with Area

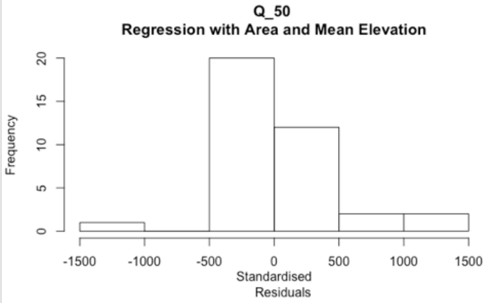

Q_100
Regression with Area

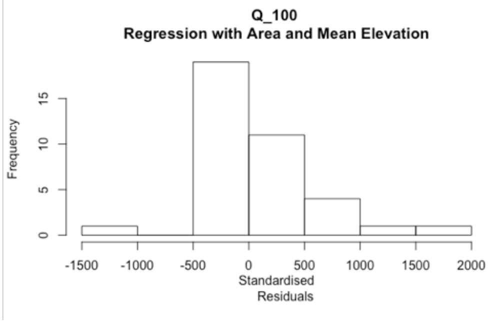
Area and Mean Elevation

The histograms below were obtained from performing a regression analysis using watershed area and mean basin elevation. The histograms seem normally distributed for all four return periods.


Q_10
Regression with Area and Mean Elevation

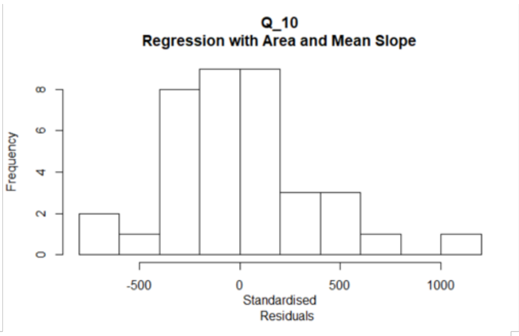

Q_20
Regression with Area and Mean Elevation

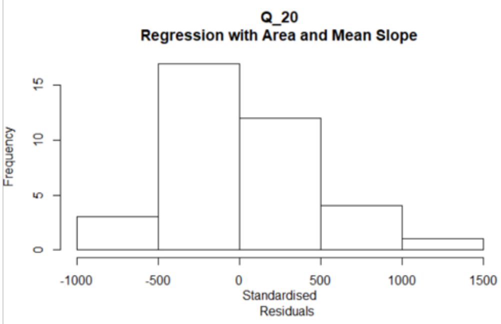

Q_50
Regression with Area and Mean Elevation

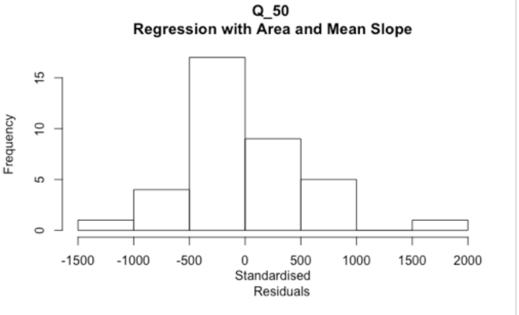

Q_100
Regression with Area and Mean Elevation

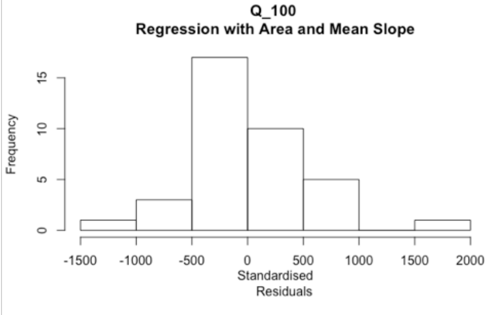
Area and Mean Slope

The histograms below were obtained from performing a regression analysis using watershed area and mean basin slope. The histograms seem normally distributed for all four return periods.







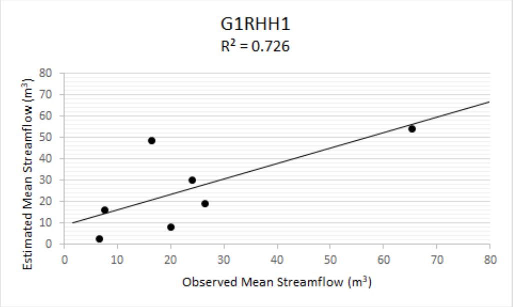


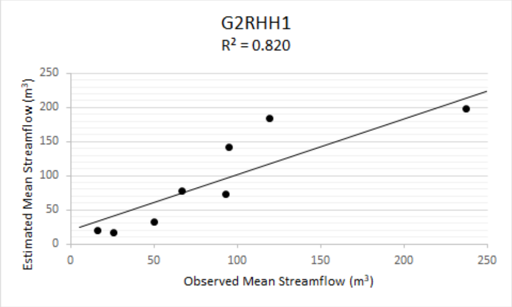
Appendix III - Comparison of Estimated and Observed Mean Annual Streamflow

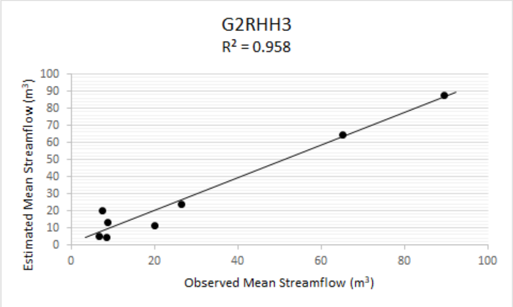
In this appendix the comparison between estimated and observed mean annual streamflow is presented for the hydrologically homogeneous regions.

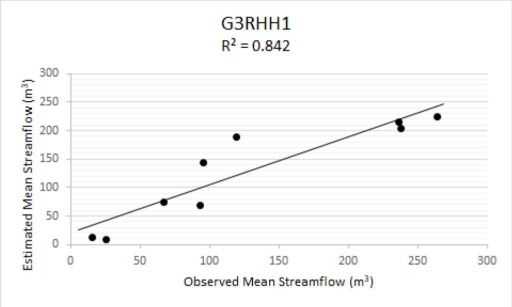
Area Only

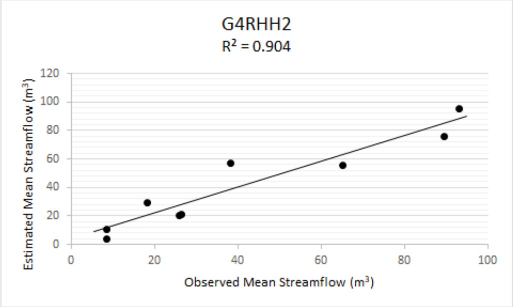
Observed mean annual streamflow compared to estimated mean annual streamflow when area only was used as predictor variable.

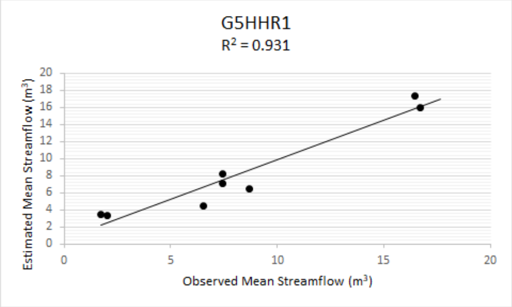


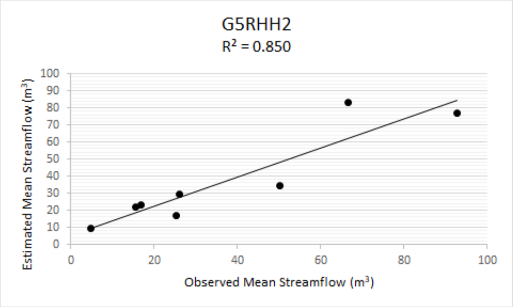


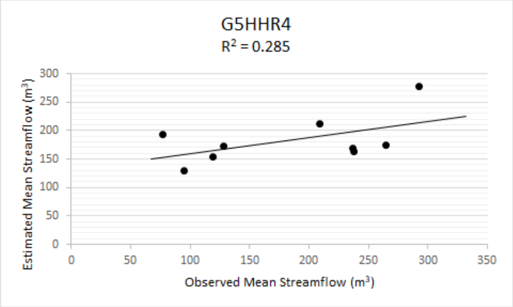


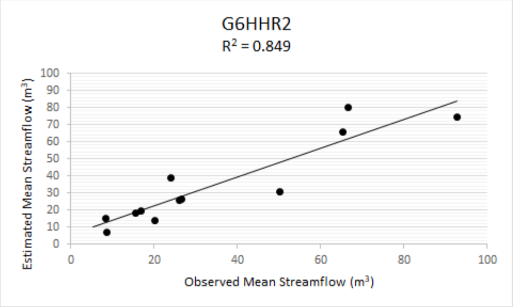


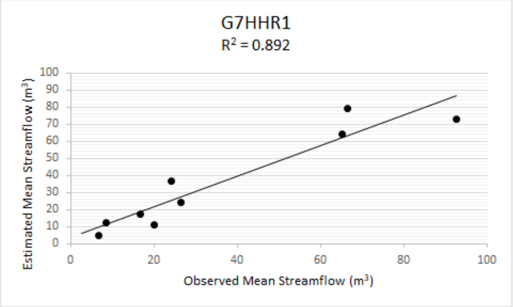


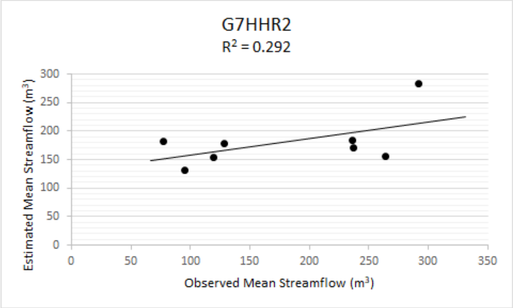


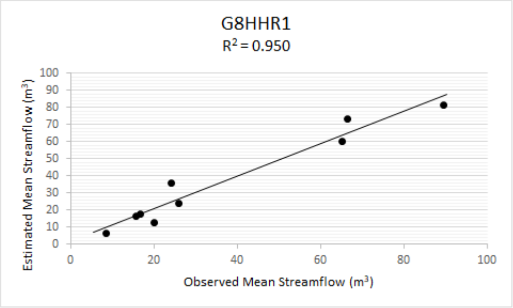

G5RHH2
R² = 0.850
Estimated Mean Streamflow (m³)
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
Observed Mean Streamflow (m³)


G5HHR4
R² = 0.285
Estimated Mean Streamflow (m³)
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
Observed Mean Streamflow (m³)


G6HHR2
R² = 0.849
Estimated Mean Streamflow (m³)
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
Observed Mean Streamflow (m³)

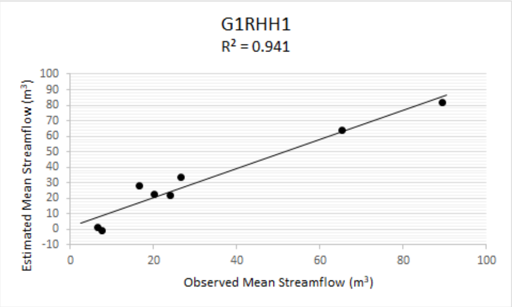

G7HHR1
R² = 0.892
Estimated Mean Streamflow (m³)
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
Observed Mean Streamflow (m³)

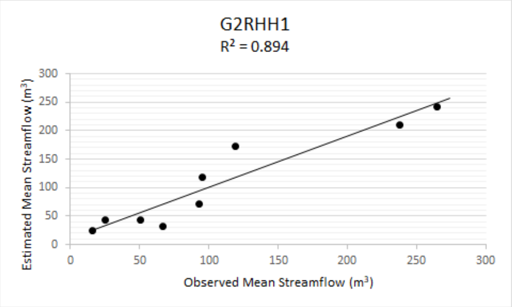

G7HHR2
R² = 0.292
Estimated Mean Streamflow (m³)
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
Observed Mean Streamflow (m³)

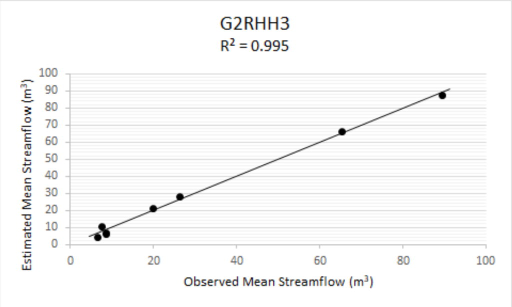

G8HHR1
R² = 0.950
Estimated Mean Streamflow (m³)
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
Observed Mean Streamflow (m³)

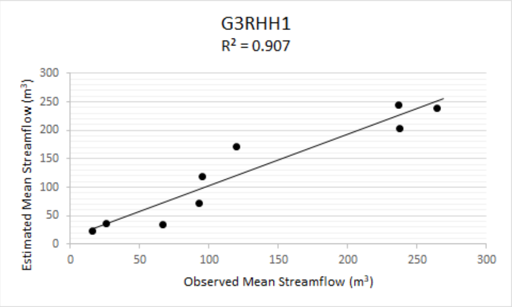
Area, Precipitation and Temperature

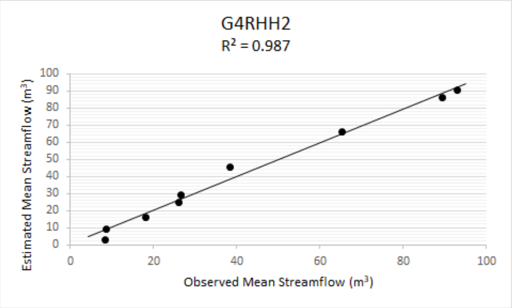
Observed mean annual streamflow compared to estimated mean annual streamflow when area, precipitation and temperature were used as predictor variables.

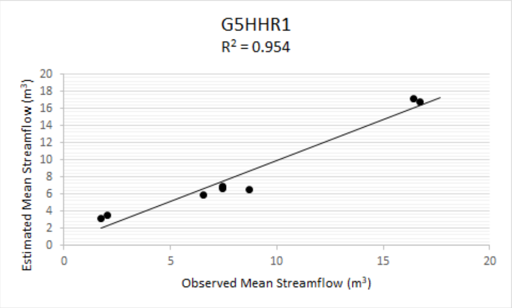


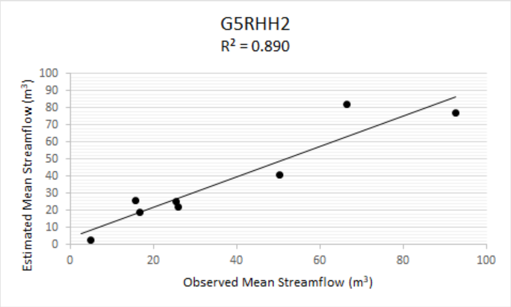






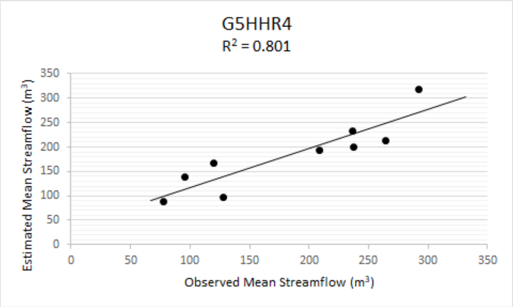






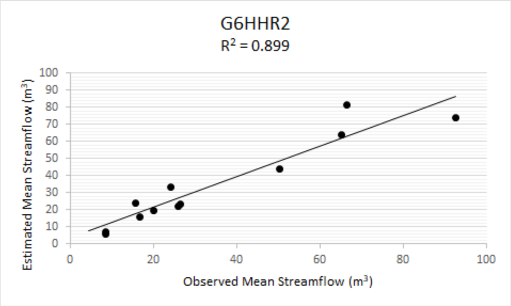
G5RHH2

$R^2 = 0.890$



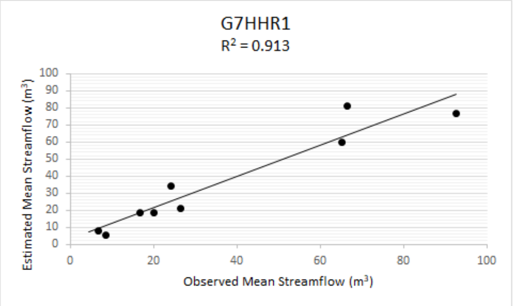
G5HHR4

$R^2 = 0.801$



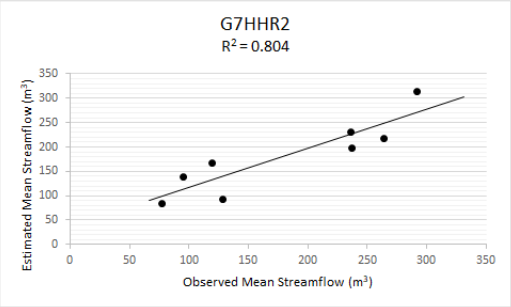
G6HHR2

$R^2 = 0.899$



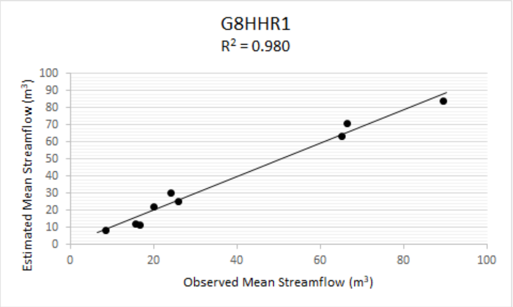
G7HHR1

$R^2 = 0.913$



G7HHR2

$R^2 = 0.804$

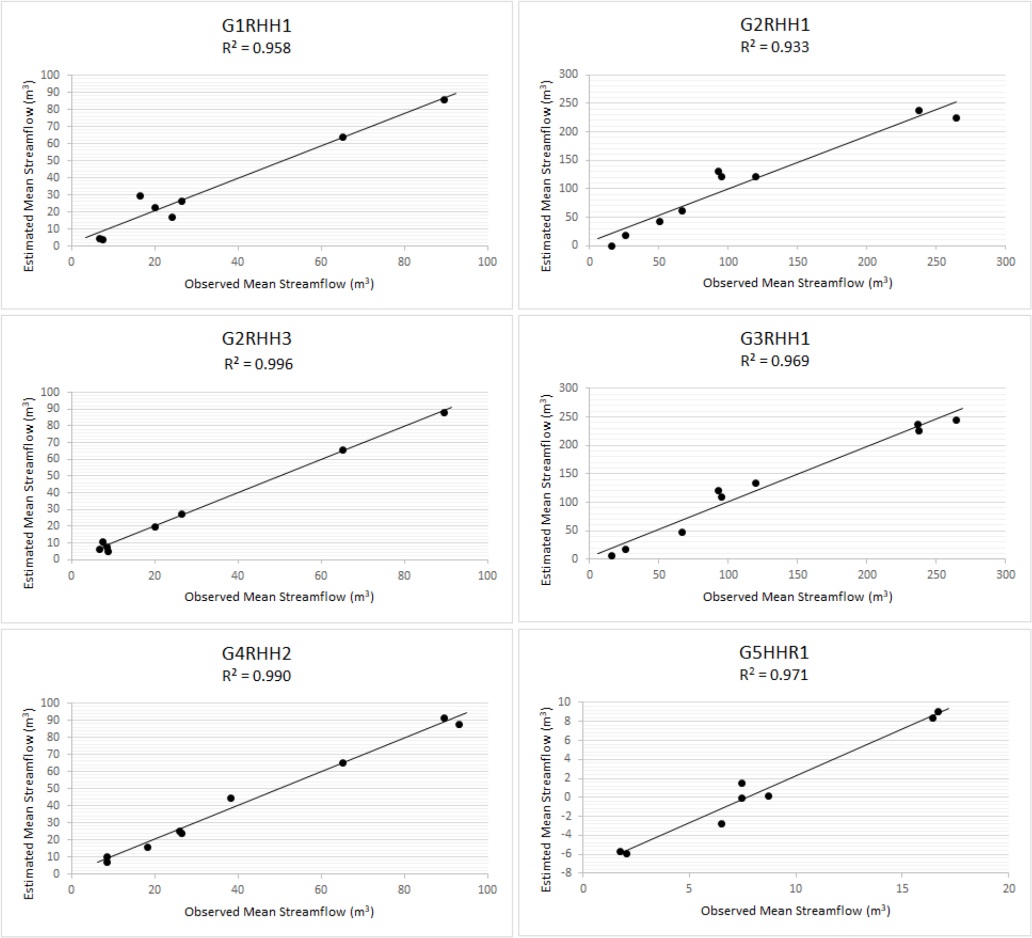


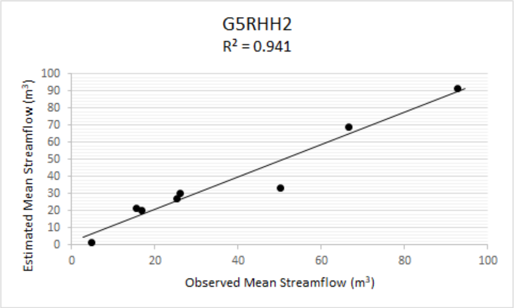
G8HHR1

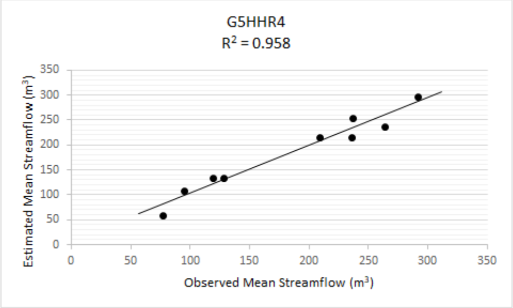
$R^2 = 0.980$

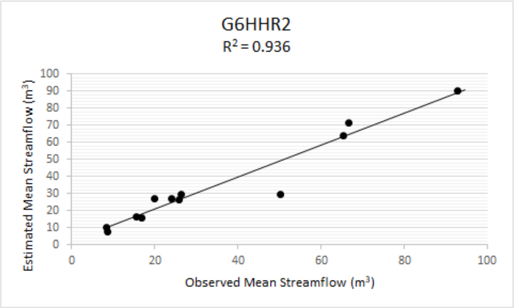
Most Suitable Variables

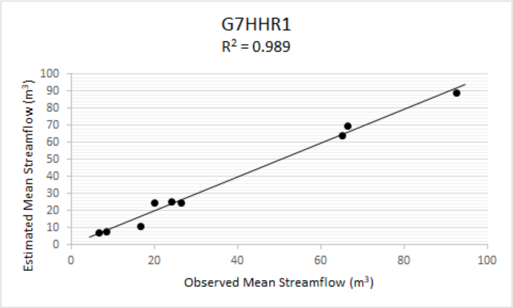
Observed mean annual streamflow compared to estimated mean annual streamflow when models were created using the most suitable predictor variables.

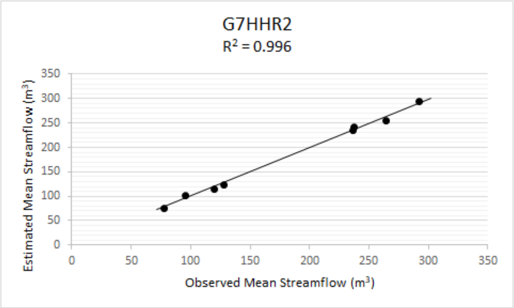


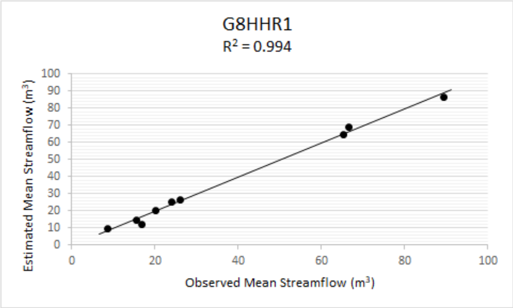

G5RHH2
R² = 0.941
Estimated Mean Streamflow (m³)
0 10 20 30 40 50 60 70 80 90 100
Observed Mean Streamflow (m³)
0 20 40 60 80 100


G5HHR4
R² = 0.958
Estimated Mean Streamflow (m³)
0 50 100 150 200 250 300 350
Observed Mean Streamflow (m³)
0 50 100 150 200 250 300 350


G6HHR2
R² = 0.936
Estimated Mean Streamflow (m³)
0 10 20 30 40 50 60 70 80 90 100
Observed Mean Streamflow (m³)
0 20 40 60 80 100

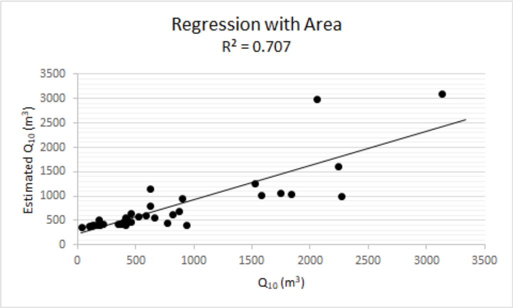

G7HHR1
R² = 0.989
Estimated Mean Streamflow (m³)
0 10 20 30 40 50 60 70 80 90 100
Observed Mean Streamflow (m³)
0 20 40 60 80 100

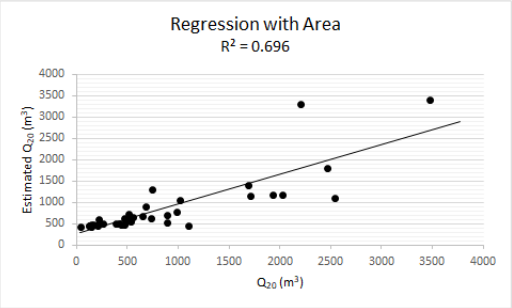

G7HHR2
R² = 0.996
Estimated Mean Streamflow (m³)
0 50 100 150 200 250 300 350
Observed Mean Streamflow (m³)
0 50 100 150 200 250 300 350

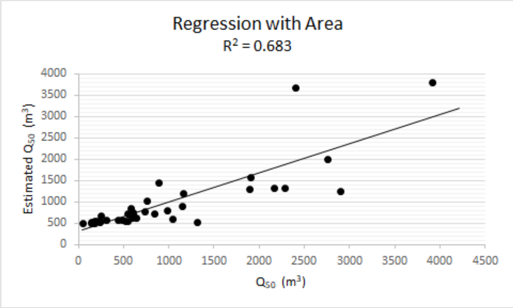

G8HHR1
R² = 0.994
Estimated Mean Streamflow (m³)
0 10 20 30 40 50 60 70 80 90 100
Observed Mean Streamflow (m³)
0 20 40 60 80 100

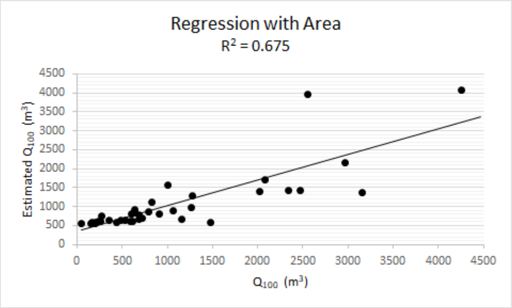
Appendix IV - Comparison of Calculated and Estimated Design Flow

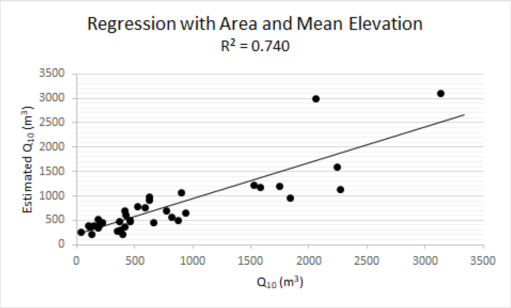
This appendix presents graphs comparing calculated design flow to estimated design flow for each of the four return periods analyzed and with the different combinations of predictor variables.

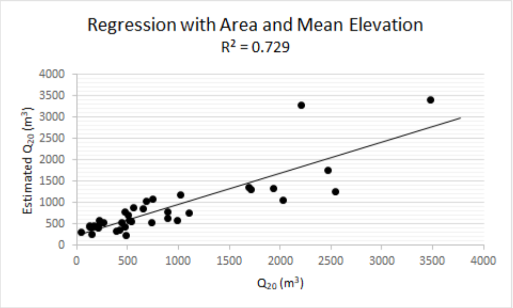


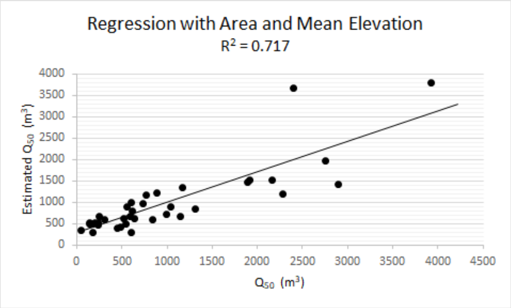


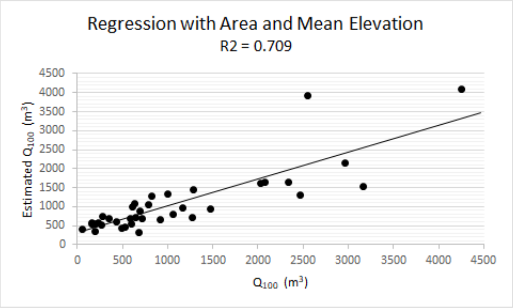


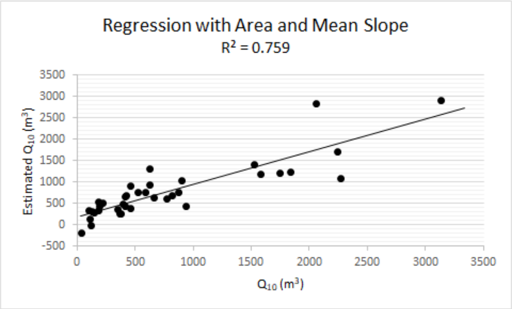


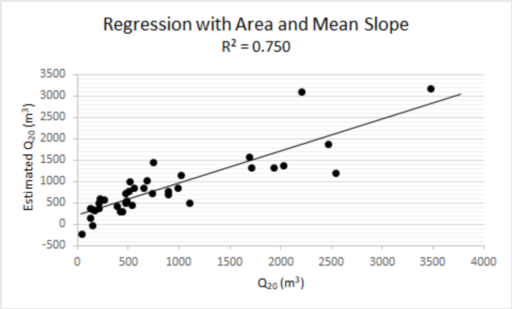


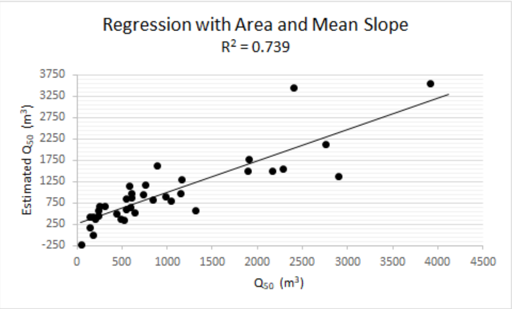


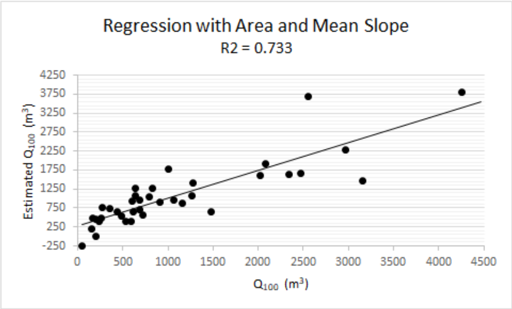





Regression with Area and Mean Slope
R^2 = 0.759
Estimated Q_{10} (m^3)
3500
3000
2500
2000
1500
1000
500
0
-500
0
500
1000
1500
2000
2500
3000
3500
Q_{10} (m^3)


Regression with Area and Mean Slope
R^2 = 0.750
Estimated Q_{20} (m^3)
3500
3000
2500
2000
1500
1000
500
0
-500
0
500
1000
1500
2000
2500
3000
3500
4000
Q_{20} (m^3)


Regression with Area and Mean Slope
R^2 = 0.739
Estimated Q_{50} (m^3)
3750
3250
2750
2250
1750
1250
750
250
-250
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Q_{50} (m^3)


Regression with Area and Mean Slope
R2 = 0.733
Estimated Q_{100} (m^3)
4250
3750
3250
2750
2250
1750
1250
750
250
-250
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Q_{100} (m^3)