

A Statistical Analysis on The Hazard of Drought

The Quantification of Hydrological Drought in
California, United States of America

S. Sanjaya

Master Thesis



A Statistical Analysis on The Hazard of Drought

The Quantification of Hydrological Drought in
California, United States of America

by

S. Sanjaya

in partial fulfillment of the requirement for the degree of
Master of Science
in Civil Engineering

at the Delft University of Technology, Faculty of Civil Engineering and Geosciences
to be defended publicly on Wednesday October 31, 2018.

Student number:	4533313
Project duration:	February 20, 2018 – September 30, 2018
Thesis committee:	Dr. ir. O. Morales-Nápoles, TU Delft, supervisor
	Prof. dr. ir. M. Kok, TU Delft
	Dr. M. Hracowitz, TU Delft
	Dr. Ted Velkamp, VU Amsterdam

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

Cover image: *Lake Oroville, in Oroville, California - Drought in 2014* (Source: <https://www.nbcnews.com/>)

Acknowledgements

It has been a roller-coaster journey of 26 months pursuing my graduate degree. I have to admit this is one of the toughest education levels so far in my life, but it is worth every single tear, energy and thought. Coming to TU Delft was one of the best decisions I have ever made. It enlightens me to another spectrum of perspectives in learning, beyond just studying a knowledge. It has also introduced me to a new horizon of science, that I have never discovered before.

This manuscript is a final chapter of me being a master student. It has taught me way a lot more than I expected: patience, curiosity, delicacy and dedication. I would like to present the best version of it, and at the same time, I am also aware that there is always room for improvement. But most importantly, I would like to express my gratitude to everybody who made it possible through any kind of support. Firstly to all committee members who guided me through the entire process. Dr. Oswaldo Morales-Nápoles, thanks for inspiring me by showing me your enthusiasm as a researcher; it has been a great time working with you for the past year. Thanks to Prof. Kok, Markus and Ted for insightful feedback on my thesis. Special thanks to Anaïs, Marthe and Mike, without your help, my thesis would not been easier.

To my family, your endless and unconditional love mean the world to me. Thanks for always trying to understand my quirkiness, despite our differences. To my other family in UNPAR (Dr. Doddi, Dr. Ida and Finna Fitrianna), their endless care and affection were priceless during my critical time, even before I came here. To my support-system in my master life: Luis, Luisa, Juan and Maria; life in TU Delft would not have been the same without you guys, I just could not imagine it. Thank you for the time we have spent together, and the love you guys shared with this *Sanhecito*. *Muchas gracias, amigos!* To a person, that I could not mention the name, thanks for trying your best to understand this complicated creature. To fellow Indonesian students in Delft, especially Etsa, Bryan, Riris, Helmi, Romi and Fajar, thank you. My thesis life would not be tackier and cheesier without you guys. Also to Tazia Teresa, thanks for the accidental support from the beginning of my master. Thanks to LPDP that supported me financially. And to everyone that I could not mention one by one, thanks for the come-and-go phase in my life.

Lastly, I would like to give a special space to the intangible bond to this country, that I would rather call as my second home. A place that letting me grow in an unique way, and has introduced me to the real adulthood life. To every moment that made me learn to be wiser, to feel the ups-and-downs, to embrace the gloomy and lonely feelings, for that I could not be more grateful. It has unwittingly shaped me to the way I am now. Of course, not every moment would be always '*gezellig*', nevertheless *dank je well Nederland*, you will always be in my heart!

S. Sanjaya
Delft, September 2018

Summary

Drought is a recurrent extreme climate event that strongly affects every spectre of the natural environment and human lives (Madadgar and Moradkhani [42]). Many studies have shown an intensifying trend of drought events throughout the world, notably in drought prone areas (FAO [23], Kundzewicz et al. [38], Wetherald and Manabe [76]), such as Midwestern part of USA. There were numerous drought episodes recorded in Midwestern USA, particularly in California. It has affected many aspects, including water shortages, groundwater overdraft, critically low stream flow (Diffenbaugh et al. [12]) and economical loss. In contrast, California as the third most extensive area in the U.S., is occupied most of the land use for agricultural purposes.

In order to capture the trend of hydrological drought in California, understanding towards the risk of hydrological drought in a comprehensive way should be conducted. In a broader sense, the quantification of hazard of hydrological drought will be a preliminary step mitigating the drought event, as an integral part of risk management. At the same time, understanding drought processes has utmost importance if risk management is to be used for drought combat, mitigation and adaptation tasks (Şen [17]). The urge to explore drought mechanism becomes a critical stage, due to the fact of the complexity of drought and to apply the concept to the model. However, many model were not adequately foreseen or estimate the hazard of drought, precisely hydrological drought. Having said that, this study is carried out to try to close the gap of the knowledge regarding the aforementioned facts in California. This research is aimed (i) to understand the characteristic of hydrological drought using statistical tool; (ii) to build a model to estimate the low stream-flow using the relationship between hydro-meteorological and hydrological drought; and (iii) to quantify the risk ratio of the hydrological drought duration using a statistical method.

In general, the thesis will be divided into two big parts, hydrological drought severity and hydrological drought duration. The former part will be analysed using Bayesian Network to estimate the mean monthly discharge in California. The latter part is evaluated using Generalised Linear Model to study the risk of having the expected number of days without discharge in California under different cases of scenarios.

In the first phase, the BN model were build using the Hydro-climatological datasets in California to estimate the mean monthly stream-flow. The hydro-climatological datasets consist of air temperature, atmosphere water vapor content, evaporation, snowmelt flux, precipitation, snowfall, soil moisture, SOF coefficient, BFI, stream density, and surface runoff coefficient. The model were tested using combination between statistical validation measures (R^2 and NSE) and K-fold. The model results in a very poor estimation of validation ($NSE = 0.165$ and $R^2 = 0.143$). A conclusion can be drawn that the built hydrological drought BN is found to inadequately explain drought, especially to understand thoroughly the hydrological drought in California. Hypotheses, such as, the assumptions of the model and lack of input variables were proposed to be one of many reasons that the model fails to estimate well the mean monthly stream-flow.

Moving towards the second phase, regarding the hydrological drought duration. GLM were tested to fit the time-series dataset in California under several scenarios. The hydrological drought duration which is expressed in expected number of days without discharge in a month, is a GLM function of precipitation, air temperature, atmosphere water vapour content, snowmelt flux, and SPI. In this part, risk ratio was utilised to express the influenced of each explanatory variable towards having days with zero discharge in California. The study, which mainly focus on the precipitation, shows that there is a increasing trend of the risk ratio to the size of the catchment area. Furthermore, in coastal-dominated areas in California, if the mean monthly precipitation is increasing for 1 mm/day, there is a range of 0.88 - 13.63 % reduced risk of having longer hydrological drought duration, where as in river dominated with 0.26 - 1.4 %. On the other hand, model that involves more explanatory variables (model 2) is performed better than model that solely as function of precipitation (model 1). A significant difference between model 1 and model 2 is found using multiple comparison method. Having said that, model 1 tends to underestimate the hydrological drought duration.

Besides gaining information from modelling hydrological drought, this thesis will also try to answer the for-

ulated research questions. The research questions focus upon the statistical perspective towards hydrological drought, by explaining the input data for the model and the advantage(s) and disadvantage(s) of such models.

Contents

Acknowledgements	iii
Summary	vi
List of Figures	ix
List of Tables	xi
Nomenclature	xiii
1 Introduction	1
1.1 Background of the Research.	1
1.2 Objectives and Scope of the Research.	3
1.3 Research Questions and General Approach.	3
1.4 Thesis Outline	4
2 Literature Overview	5
2.1 The Concept of Drought	5
2.1.1 Drought: Definition and Variables of Interest	5
2.1.2 Drought Models: The Concept Beyond.	7
2.1.3 The Concept of Rainfall-Runoff in Hydrology	8
2.1.4 Drought Measurement in The Netherlands	9
2.2 Case Study Area	11
2.2.1 System Description: California in a Nutshell	11
2.2.2 Drought in California: The Current and Historical State	11
2.3 Bayesian Networks	12
2.3.1 Non-Parametric Bayesian Networks	15
2.4 Prediction Power for BNs	16
2.4.1 Method for Result Validation: K-fold Cross-Validation	16
2.4.2 Error Measurements: R^2 and NSE	16
2.5 Generalised Linear Models	17
2.6 ANOVA: Analysis of Variance	18
I Hydrological Drought Severity	21
3 Bayesian Network	23
3.1 Preliminary Stages of Constructing the Bayesian Network	23
3.1.1 Model Assumption for the BN.	23
3.1.2 Input Variables: Extracting Hydrological and Meteorological Variables	24
3.2 Building up the BN for Hydrological Drought	26
3.2.1 Scenario of the Sensitivity Analysis.	28
3.2.2 Statistical Approach	28
3.2.3 Conceptual Approach	31
3.2.4 The Graphical Result of BN for Hydrological Drought	33
3.3 Results and Discussion	35
3.3.1 Inference Result: Comparing Different Data Input	35
3.3.2 Discussion	36
3.4 Conclusion	37

II Hydrological Drought Duration	39
4 Hydrological Drought using Generalised Linear Model	41
4.1 Preliminary Stages on Constructing Linear Model	41
4.1.1 Variable in Datasets: Response and Explanatory Variable	41
4.1.2 Model Structures: Scenarios for the Different Variables of Interest	42
4.1.3 Preliminary Steps: Calculating the Correlation Coefficient of the Datasets	46
4.2 Results and Discussion	46
4.2.1 Fitting Example: Analysing Precipitation to the Response Variable in California	46
4.2.2 Linear Model Equation: Applying GLM in California	48
4.2.3 Hazard Estimation: Covariate Contribution to Drought Hazard Duration in California	51
4.2.4 ANOVA: Comparing the Result of Scenario 3	59
4.3 Conclusion	63
5 Conclusion and Recommendation	65
5.1 Conclusion	65
5.2 Recommendation	66
A Input Variables: Possible Relevant Variables in Drought Mechanism	67
A.1 Spatial Datasets	67
A.2 Hydrological Variables: Drainage Basin Characteristics	68
B Result of Partial Correlation of All Input Variables	69
Bibliography	73

List of Figures

1.1	Map showing the drought hazard in the world. It shows that California is classified as 8th - 10th deciles of drought hazard (source: Dilley et al. [13])	1
1.2	Economic loss due to major droughts worldwide from 1900 to 2016 (in billion U.S. dollars). It shows that drought in the US in 2012 caused a loss of 20 billion USD, as the biggest economic loss since 1900. (source: CRED [9])	2
1.3	A schematic diagram for research methodology	3
1.4	The general structure of the thesis report	4
2.1	Four types of drought categories	5
2.2	Major impact of drought about the different drought categories (source: Van Loon [75], 2015)	6
2.3	Space-time domain of flood and drought in hydrology and meteorology (source: Stahl and Hissdal [67], 2004)	7
2.4	Propagation of drought through the hydrological cycle (source: Stahl [66], 2001)	8
2.5	Descriptive Representation of water fluxes and storages in hydrological processes (source: Savenije [61], 2006)	9
2.6	Accumulated precipitation deficit of 13 stations spread in the Netherlands in 2018 in black line. The x-axis indicates the time of half-year summer (April - September) and the y-axis shows the precipitation deficit in mm. (source: KNMI [36], 2018)	10
2.7	Potential precipitation surplus in half-year summer 2018 in the Netherlands. The negative values shows the shortage of precipitation, whereas the positive values indicates the surplus of precipitation source: KNMI [37], 2018)	10
2.8	Land cover of California (source: USGS, 2016)	12
2.9	Geographical borders of California	13
2.10	Average annual runoff and precipitation for historical droughts in the state of California, data is taken from USGS [71]	14
2.11	Graph of California DSCI for the past two decades. The trend has shown an increasing severity, duration and areal coverage. (source: The National Drought Mitigation Center [69], 2018)	14
2.12	Examples of a simple Bayesian Network Belief with different structuress (source: Couasnon [8])	14
2.13	Example of K-fold cross-validation with $K = 3$	17
3.1	Points of discharge gages in California. There are 810 points along with the drainage area delineated by GAGES II, USGS	24
3.2	Example of time series of variables in station 10258000 (1)	26
3.3	Example of time series of variables in station 10258000 (2)	27
3.4	Example of time series of variables in station 10258000 (3)	27
3.5	Flowchart diagram of sensitivity analysis	28
3.6	The correlation coefficient matrix for variables	29
3.7	Saturated Check on the BN	31
3.8	BN rank correlation matrix	32
3.9	The categories for the input variables	32
3.10	The result of BN graph-1 using UNINET	33
3.11	Box plot of validation measures of BN conceptual model - 1	33
3.12	The result using UNINET	33
3.13	Box plot of validation measures of final conceptual BN	34
3.14	Final result of Bayesian Network for explaining hydrological drought using meteorological and hydrological variables	34
3.15	Inference result for different types of data classification in logarithmic scale. The plot shows between the observed of q_{MM} in y-axis and the estimated value of q_{MM} in x-axis	36

4.1	Example of time series of explanatory variables in station 10258000	43
4.2	Ten hydrological regions in California that delineated based on the main hydrological functions	44
4.3	Station distribution on time period sub-datasets	45
4.4	Result analysis for GLM with precipitation as the explanatory variable	47
4.5	Maps of different risk of hydrological drought duration in the hydrological regions in California using model equation 4.3. The blue hatch indicates the coastal regions, while the green hatch indicates the river regions.	53
4.6	Histogram of risk ratio ($0 < RR < 1$) of days of without Discharge induced by precipitation model 1. The red graph shows the kernel PDF, the greenline indicates the mean value.	55
4.7	Histogram of risk ratio ($0 < RR < 1$) of days of without Discharge induced by precipitation model 2. The red graph shows the kernel PDF, the greenline indicates the mean value.	56
4.8	Mean value of the risk ratio (induced by precipitation) versus time periods of the data for both models.	56
4.9	Spatial distribution map of risk ratio of number of days without discharge induced by precipitation using model 1	58
4.10	Spatial distribution map of risk ratio of number of days without discharge induced by precipitation using model 2	59
4.11	Multicomparison on Model 1 - 1990 and Model 1 - 2005	63
B.1	Partial Correlation of mean monthly dishcarge per catchment area and input variables (1)	70
B.2	Partial Correlation of mean monthly dishcarge per catchment area and input variables (2)	71
B.3	Partial Correlation of mean monthly dishcarge per catchment area and input variables (3)	72

List of Tables

2.1	Summary calculation of the one-way ANOVA	19
3.1	Table of details of the meteorological variables in the model	25
3.2	Table of details of the hydrological variables in the model	25
3.3	Example of technical information and coefficient for Station 10258000	26
3.4	Spearman 's Rank correlation coefficient comparing the result of Discharge and Discharge per Catchment Area	30
3.5	Table of semi-correlations between all input variables and q_{MM}	30
3.6	Validation measures of different types of data	36
4.1	Description of explanatory variables as input for the model	42
4.2	Statistical summary of the response variable in Station Fresno R BL Hidden Dam	42
4.3	Description of hydrological regions in California	43
4.4	Description of class catchment area in California	44
4.5	Table summary of different scenarios	46
4.6	Correlation coefficient of each variables	46
4.7	The GLM regression result retrieved from MATLAB using the whole datasets	48
4.8	Estimated value and standard error (S.E.) of the coefficient for the whole dataset, scenario 1 and scenario 2	50
4.9	Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for coastal-dominated region	51
4.10	Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for river-dominated region	52
4.11	Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for mountainous-dominated region	53
4.12	Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for class catchment area	54
4.13	The statistical summary on the risk ratio result on scenario 3	60
4.14	Table of result for one-way ANOVA model 1 (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the F -statistics)	60
4.15	Table of result for one-way ANOVA model 2 (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the F -statistics)	60
4.16	Table Two-way ANOVA (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the F -statistics)	61
4.17	Table one-way ANOVA of groups of data (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the F -statistics)	61
4.18	Table of mean estimate difference for all combinations of two groups (where Lower C.I. is the lower confidence interval; δ ME is the difference of mean estimate; High C.I. is the high confidence interval)	62

Nomenclature

ANOVA	Analysis of Variance
BN	Bayesian Network
CA	California
CDF	Cumulative Distribution Function
CORDEX	Coordinated Regional Climate Downscaling Experiment
DAG	Directed Acyclic Graph
DSCI	Drought Severity and Coverage Index
FEMA	Federal Emergency Management Agency
FNF	Full Natural Flow
GAGES II	Geospatial Attributes of Gages for Evaluating Streamflow
GLM	Generalised Linear Model
IPCC	Intergovernmental Panel on Climate Change
KNMI	Royal Netherlands Meteorological Institute (<i>Koninklijk Nederlands Meteorologisch Instituut</i>)
NADM	North American Drought Monitor
NPBN	Non-Parametric Bayesian Network
NSE	Nash-Sutcliffe Efficiency
RCA4	Rosby Centre Regional Atmospheric Model
SMHI	Swedish Meteorological and Hydrological Institute
SPI	Standardized Precipitation Index
TWI	Topography Wetness Index
US	United States
USGS	United States Geological Survey
WEF	World Economic Forum
WY	Water Years

Introduction

1.1. Background of the Research

As a natural disaster, drought is a recurrent extreme climate event that strongly affects every spectre of the natural environment and human lives [42]. Drought also affects more people than any other environmental hazard [29]. Among all atmospheric hazards, droughts are possibly the most slowly developing ones, which often have the most prolonged duration, and at the moment one of the least predictability [48]. At present time, droughts duration has happened more frequent with increased intensity on the entire globe [23]. Based on IPCC study, drought during the last millennium has a greater magnitude and longer duration than those observed since the beginning of the 20th century in many countries. Within the same study, it is also projected that by the end of the 21st century, there is a higher risk of having dryer surface [44]. In consequence, a warmer climate, with its increased climate variability, will increase the risk of both floods and droughts [38, 76].

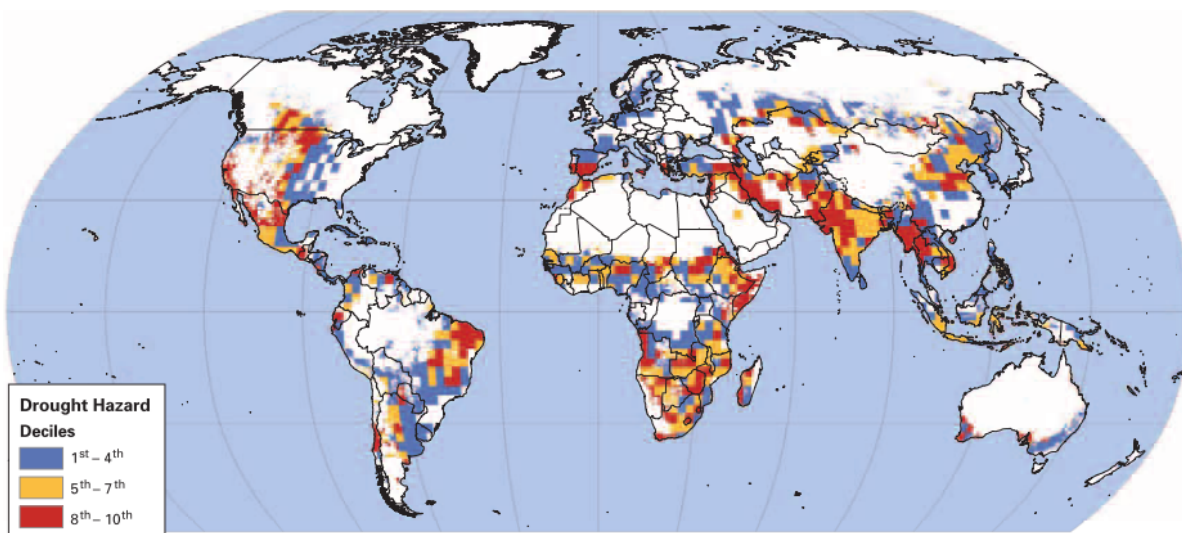


Figure 1.1: Map showing the drought hazard in the world. It shows that California is classified as 8th - 10th deciles of drought hazard (source: Dillely et al. [13])

According to WEF's report in 2018, drought is one of the top five global risk [78]. This trend will take place all around the world and intensify especially in drought-prone areas. Figure 1.1 shows the drought-prone areas in the world classified in three different deciles¹. Regions such as The Midwestern United States, Central

¹Based on Merriam-Webster dictionary, decile is any one of nine numbers that divide a frequency distribution into 10 classes such that each contains the same number of individuals [47]

America, the sub-Saharan belt, southern Spain and Portugal and north-west India are in highest decile[13]. Apart of having higher risk of drought, the impact of drought in human activity is very critical. Drought has an impact on agriculture at various scales of crop production[48], primarily a threat to food and water security [22]. It is also severely aggravated by human activity, such as excessive water use, deforestation etc. The fact that drought has a vast socioeconomic impact. However, there is still a lack of quantification of socio-economic losses [48]. The economic, social and environmental costs and losses associated with drought are also increasing dramatically, even though it is difficult to quantify this trend precisely because of lack of reliable historical estimates of losses [77]. The drought's damages since 1900 that translated into economic loss are shown in Figure 1.2. The statistic shows that drought event in the US in 2012 has caused the biggest economic loss. Additionally, there were three drought events in the US in the past two decades that were recorded as the top ten economic loss. It mainly occurred in the Midwest part of the US, precisely in the state of California.

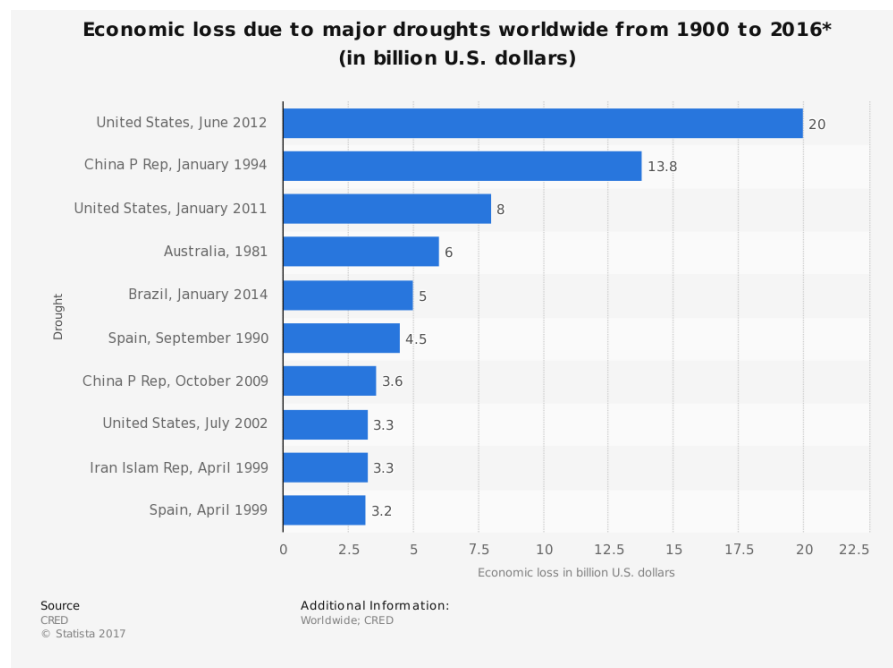


Figure 1.2: Economic loss due to major droughts worldwide from 1900 to 2016 (in billion U.S. dollars). It shows that drought in the US in 2012 caused a loss of 20 billion USD, as the biggest economic loss since 1900. (source: CRED [9])

California, as one of the biggest drought-prone areas, has had several episodes of drought period. Since 1895, there have been six prolonged dry periods lasting for two or more years (USGS, 2017). Those were water years ² (WY) 1928–1934, WY 1976–1977, WY 1987–1992, WY 2001–2002, WY 2007–2009, and WY 2012–2016. Study in 2014 has shown that the latest 3-year episode of drought, from 2012 until 2014, has surpassed the historical drought in 1976–1977 and in late 1980s (see Griffin and Anchukaitis [28]). Consequently, the state is currently experiencing a record-setting drought, which has led to acute water shortages, groundwater overdraft, critically low stream flow, and enhanced wildfire risk [12].

This condition caused economic losses, especially in the agricultural sector. FEMA has estimated that the annual drought losses in the United States are at \$ 6-8 billion[24], whereas the reported costs for the recent drought in the western United States exceed \$10 billion merely in 2002 [41]. In 2014, California's drought caused approximately \$ 2.2 billion in total economic losses, with \$ 810 million estimated of crop revenue loss [5]. In the following year, according to a UC Davis study, drought caused \$ 2.7 billion of economic loss which is equivalent to about five percent of annual agricultural production in the US [35].

Obviously, the loss can be reduced or even prevented once one understands the process of drought. Un-

²Water year is a term to differ between dry and wet periods. This period starts on 1st October of the preceding year and ends on 30th September of the water year. For instance, Water Year 2016 starts on 1st October 2015 and ends on 30th September 2016.

derstanding drought processes has utmost importance if risk management is to be used for drought combat, mitigation, and adaptation tasks [18]. Hence, in a preliminary stage of risk management, the risk quantification of drought is an important action to cope with the current problem. Nevertheless, drought is a complex natural hazard, since there is no clear pattern in explaining the mechanism of drought. It becomes another challenge in quantifying the risk of drought since the process of drought has not been clearly defined.

1.2. Objectives and Scope of the Research

The mechanism of drought is hard to explain due to its complexity. It depends not only on the atmosphere but also on the hydrological processes which supply moisture to the atmosphere [48]. It emphasises the interrelationship between numerous variables, which brings the urgency to have a better understanding of drought. Hence, the objective of this study is (i) to understand the characteristics of hydrological drought using statistical tools. The term of hydrological drought will focus on the low stream-flow of the river. It is expected that after this study, one could understand better the relation between drought towards hydrological and meteorological variables in nature, as well as be able (ii) to build a model to estimate the low stream-flow using this relationship. Besides the severity of drought, drought duration also plays an important role in characterising drought. Drought duration has shown to affect its severity too (see study of Rajsekhar et al. [55]). This study will (iii) quantify the risk ratio³ (see section 4.2.1 for explanation) of the hydrological drought duration, also using a statistical method. Risk ratio will help to explain the sensitivity of relevant variables to the hydrological drought duration. As explained in section 1.1, California has several episodes of drought with a severe loss in the past decade. Thus, California is selected as the focal point of the study.

1.3. Research Questions and General Approach

As previously mentioned, there is no clear pattern on how hydro-climatological variables interact in nature, precisely during and prior to the drought. This makes modelling drought becomes more challenging. At the same time, this gives an opportunity to explore drought modelling using a different approaches, such as statistical models. In this study, the statistical tools that will be used are Bayesian Networks (BN) and Generalised Linear Model (GLM). Figure 1.3 depicts the general structure on how to obtain the outcome this thesis. The output can be derived using the mentioned models with hydrological and meteorological variables as the input. The detailed methodology of BN and GLM will be further elaborated in Chapter 2.

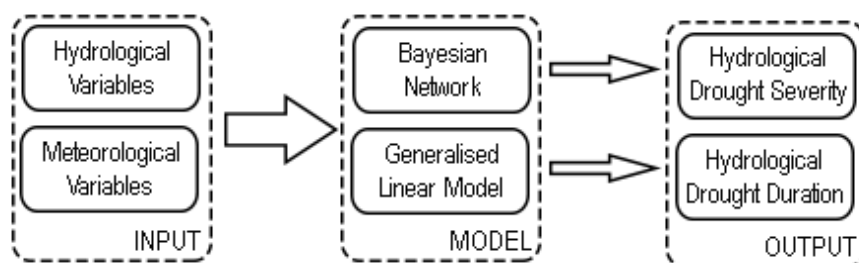


Figure 1.3: A schematic diagram for research methodology

Based on the objectives, the following research questions have been developed for this study:

What are the hydro-meteorological variables that have a strong influence in characterising hydrological drought parameters?

This question is aimed to answer the gap upon the understanding of hydrological drought with respect to hydro-meteorological variables. The recent development (that will be explained in section 2.1) shows that many droughts modelling rely solely on precipitation or evaporation, or combination of both. This is not entirely correct, as another variables might also control the drought characteristic. In this research, the main focus of the hydrological drought parameters are its severity and duration. These parameters will be quantified using different statistical tools. It will be started with collecting the possible hydro-meteorological data, sensitivity analysis to find out the relevance of the variables, and validation to measure the reliability of the

³In a rough definition, the risk ratio in this research is the ratio of expected number of days without discharge (or with zero discharge) to the expected number of days with discharge

results. Throughout these processes, one could derived which hydro-meteorological variables explain better the hydrological drought parameters.

What are the advantages (and disadvantages) of hydrological drought models, particularly using statistical models?

The application of statistical methodology for hydrological drought models will answer this question, in a preliminary manner. Throughout the processes, it will disclose the drawbacks and challenges when it comes to building up the model. On the other hand, it would give another fresh perspective towards the development of drought model. Of course, we will be able to answer this question, once the outcome of the model has been interpreted.

1.4. Thesis Outline

The main report consists of five chapters. It begins with this Introduction, followed by Chapter 2 which presents the state-of-the-art of drought modelling, a short description of the drought in California, as well as a brief explanation of Bayesian Networks and Generalised Linear Models. There will be two main parts, the hydrological drought severity (Chapter 3) and hydrological drought duration (Chapter 4). Chapter 3 presents the model of hydrological drought severity in California using Bayesian Network with hydrological and meteorological variables as input. Chapter 4 presents the risk ratio quantification of hydrological drought duration using GLM. Lastly, Chapter 5 presents the main findings of the thesis by answering the formulated research questions and recommendations for future research. Figure 1.4 depicts the general structure of the thesis report. Later, the appendix also present the explanation of input variables in the model (appendix A) and the

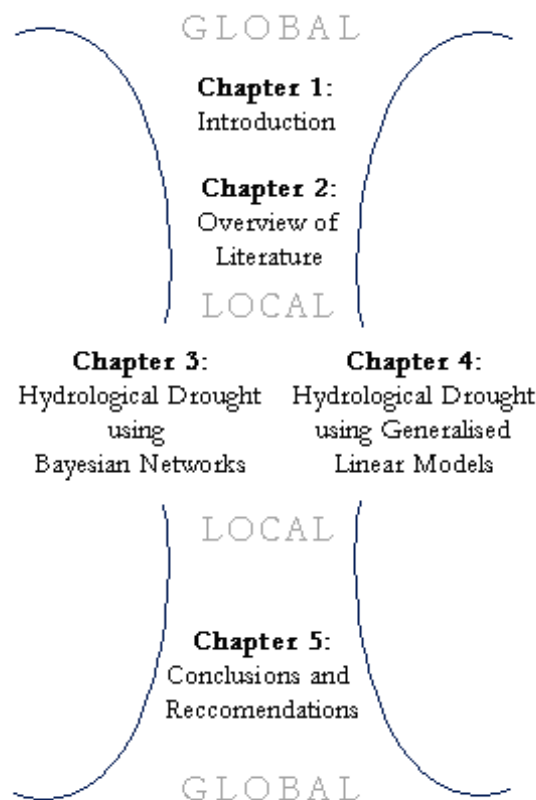


Figure 1.4: The general structure of the thesis report

figures for partial correlation of input variables in the model (appendix B).

2

Literature Overview

This chapter presents the general idea of theories and concepts that will be used in the thesis. Starting with the general definition of drought and its modelling components; continues with a brief explanation of California as the case study and ends with an introduction to the statistical model.

2.1. The Concept of Drought

2.1.1. Drought: Definition and Variables of Interest

A definition of drought proposed by Tallaksen and Van Lanen [68], 2004 is:

Drought is a sustained period of below-normal water availability. It is a recurring and worldwide phenomenon, with spatial and temporal characteristics that vary significantly from one region to another.

Drought cannot be defined in a universal context. It is easier to observe drought based on a particular characteristic. Therefore, droughts are commonly classified into four big categories, each containing different measure variables to quantify drought, as depicted in Figure 2.1

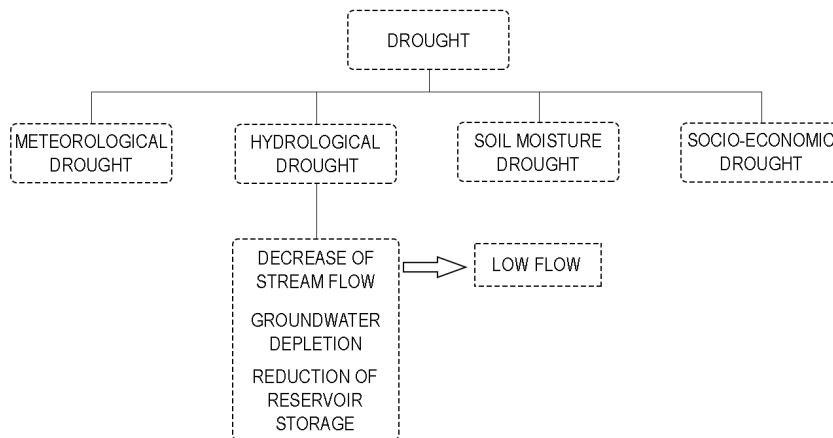


Figure 2.1: Four types of drought categories

In the following section, each type of drought will be shortly explained:

1. Meteorological drought

This type of drought is the most common perception of drought. Usually, it refers to the day without precipitation [48]. Mishra and Singh [48] define meteorological drought as the lack of precipitation over

a specified region for a period of time. In other words, meteorological drought is measured by the level of deficiency of precipitation in a particular region in an extent of time. Having said that, the level of threshold of meteorological drought varies depending on the characteristic of the region. Meteorological drought takes place due to many climatic factors. However, it is considered as the initiator of other types of droughts [17]. Precipitation has been commonly used in many drought analyses, e.g. Chang [3], Eltahir [16], Pinkayan [54], Santos [60]. These droughts analyses dealt with extreme value theory to express the deficit of precipitation in a certain return period.

2. Hydrological drought

Hydrological drought refers to the shortfall of water availability in surface and subsurface with respect to a certain level of threshold. The characterisation of the threshold depends on either water demand, engineering design discharge, or statistical exceedance levels [17]. Hydrological drought usually indicated by the reduction in lake storage, lowering of groundwater levels and a decrease of streamflow discharge that happened in one period of time and affect large areas [64]. Streamflow time series has been applied in analysing drought, likewise precipitation. For instance Chang and Stenson [4], Clausen and Pearson [7], Dracup et al. [15], Frick et al. [25], Mohan and Rangacharya [49], Zelenhasić and Salvai [79]. Based on Figure 2.2, it is clearly seen that compared to other types of drought, hydrological drought has affected many level of the categories.

3. Soil moisture drought

The soil moisture drought is a deficit of soil moisture content in the unsaturated zone⁴. As soil moisture is crucial to supply water to vegetation, this type of drought is also recognised as agricultural drought. Many studies have been conducted to understand agricultural droughts, by combining precipitation, temperature and soil moistures [48], such as PDSI (Palmer Drought Severity Indices) which are proper representation in explaining the impact of agricultural drought.

4. Socio-economic drought

Unlike another type of drought, socioeconomic drought is related to the impact of having drought disaster, as provided in Figure 2.2. Some also refer socioeconomic drought as the failure of water supply to meet the demand [75].

Impact Category		Drought Category		
		Meteorological Drought	Soil Moisture Drought	Hydrological Drought
Agriculture	Rainfed	x	x	
	Irrigated		x	x
Ecosystems	Terrestrial	x	x	
	Aquatic			x
Energy and industry	Hydropower			x
	Cooling water			x
Navigation				x
Drinking water				x
Recreation				x

Figure 2.2: Major impact of drought about the different drought categories (source: Van Loon [75], 2015)

To characterize the term of drought itself, a certain indicator need to be set, which is the variable of interest. In hydrological drought, the decrease of the streamflow could be express in low flow. Thus, the quantification of low flow would be the variable of interest in expressing the term drought.

LOW FLOW - The low flow regime of a river or time series of low flow values resembles the characteristic of the drought behaviour of a river [32]. However, the terminology of low flow has many interpretations, depending on the situations. Therefore, it is important to describe the concept of low flow clearly. Low flow in this research is not limited to a prevailing discharge during a dry season, nevertheless also as a reduction in various aspects of the overall flow regime [64]. It is evident that there is no separation between dry and wet period. Nevertheless, the issue will be treated by including the seasonality factor in the model, as will be explained in section 2.1.2

⁴Unsaturated zone is the zone between the land surface and the ground water table [51]. Unsaturated zone is also known as root zone or vadose zone. The water content in this zone is held by capillary and adsorptive forces [33]. It has significant role for the life of vegetation being the place for water uptake by plant roots.

The next stage would be the main driving force of drought and low flow. Since drought will be expressed in low flow, thus, any driving force or mechanism of low flow will be taken into account in defining drought. The ground basis of this study will be a combination concept between low flow and drought propagation (that will be discussed in the following subsection). Study by Smakhtin [64] shows that the natural factors that influence the low flow in the river include the distribution and infiltration characteristics of soils, the hydraulic characteristics and extent of the aquifers, the rate, frequency and amount of recharge, the evaporation rates from the basin, distribution of vegetation types, topography and climate. All the mentioned factors could be classified in gains and losses. In addition to that, any anthropogenic influences should be treated separately.

2.1.2. Drought Models: The Concept Beyond

SPATIAL AND TEMPORAL VARIABILITY - The development of drought takes place slowly and usually over large areas that prolong the period of below-normal precipitation [73]. A certain drought event has different temporal and spatial scale. Therefore, to build a drought model, it is important to be aware of the variability. The following Figure 2.3 shows the time and space domains of flood and drought in hydrology and meteorology. Compared to flood, drought has larger affected area and longer time scale. A drought event develops as a result of an accumulated water deficit [67]. In contrary, a flood occurs due to distinctive events, such as rain, storms, or snow melt. Therefore, having water depletion in a particular time extent over a certain region might lead to a drought event.

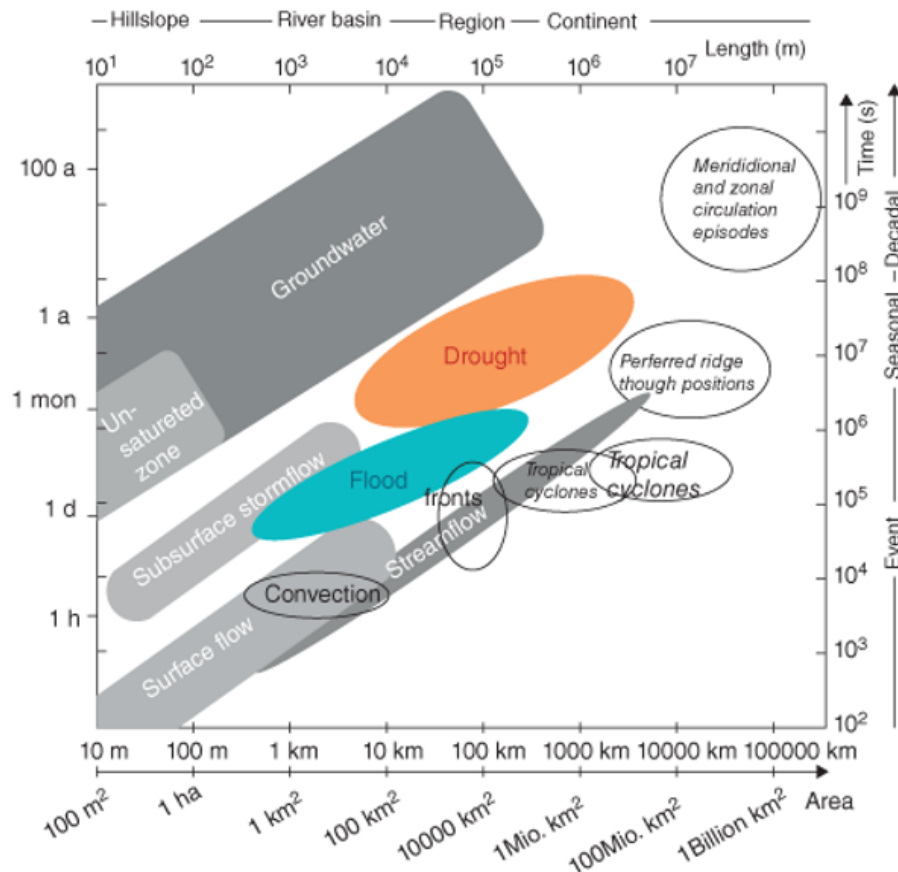


Figure 2.3: Space-time domain of flood and drought in hydrology and meteorology (source: Stahl and Hisdal [67], 2004)

Even though drought usually happens on a regional scale, there is still a heterogeneity in the drought-affected areas. Stahl and Hisdal have studied that when considering spatial variations in their drought model, according to them, there are four important characteristics to be included [67]. The characteristics of spatial variability in drought event are, as follows:

1. The spatial extent of the individual event
2. The variability within the affected area

3. The dynamics of the spatial extent

4. A recurrent pattern in space

Temporal variability exists due to different the time scale variation of hydrological behaviour. To analyse the temporal variation in drought events for example, one should have sufficient availability of the data. By having so, it is possible to assess the trend of a drought event. Commonly, it is suggested to have 30 years or more time series to define the particular characteristic of the hydrological regime [67]. To obtain a consistent result, it is important to be aware of the seasonality under the study region, when calculating hydrological drought characteristic [32]

DROUGHT PROPAGATION - The starting point of drought development is natural climate variability as a part of atmospheric processes [62, 75]. A lack of precipitation over an extensive area and period of time would lead to meteorological drought. This lack of water input will affect the hydrological cycle. If this condition continues, along with the combination of having increased evaporation and transpiration, it will generate the soil moisture deficiency. In this stage, the soil moisture is not able to support crop (or the so-called agricultural drought). Subsequently, the streamflow and groundwater depletion will occur in different timescales. The lag time from having meteorological drought to a groundwater drought may take months or even years, whilst streamflow drought occurs faster, in daily or monthly lags. The explained mechanism is illustrated in Figure 2.4. This process is the so-called drought propagation or defined as the movement process from meteorological to hydrological drought. Furthermore, the benefit of drought propagation is able to explain the dependency of different type of droughts in the hydrological cycle [75].

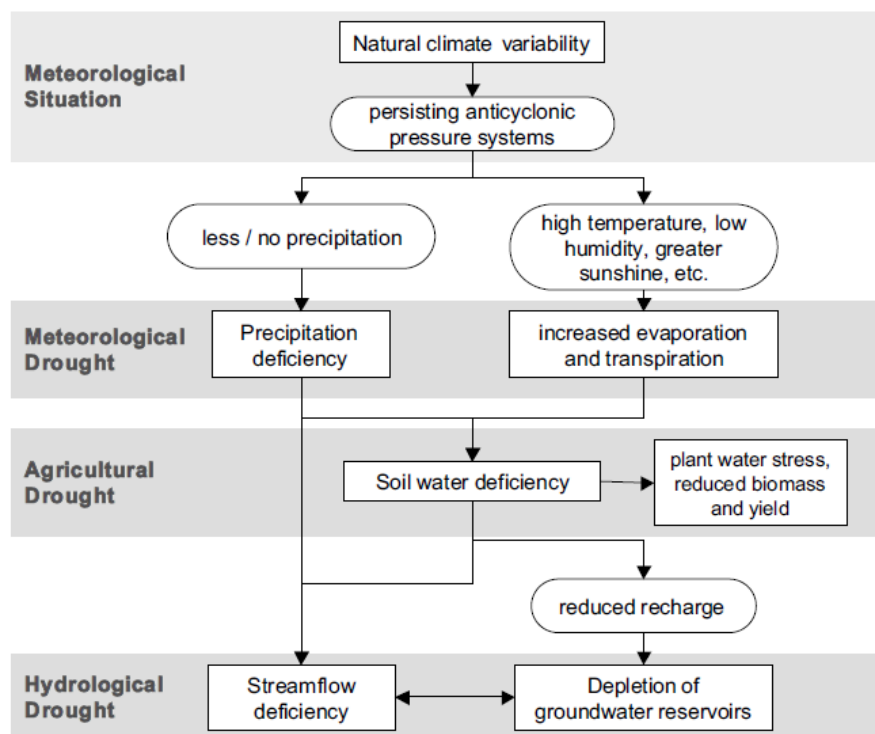


Figure 2.4: Propagation of drought through the hydrological cycle (source: Stahl [66], 2001)

2.1.3. The Concept of Rainfall-Runoff in Hydrology

This section becomes one of the important parts as the statistical model will try to imitate the run-off mechanism to estimate the low-flow. Generally, the proportion of the water volume of such mechanism is computed using water balance concept. A water balance is a general control water volume equation using the basis of continuity, momentum and energy equations for various hydrology processes [61]. In a catchment scale, the

continuity equation formulates the water balance as follows:

$$\frac{dS}{dt} = (P - E) \cdot A - Q \quad (2.1)$$

where P is the precipitation flux, E is the evaporation flux, A resembles the size of focused catchment area, and Q is the river discharge within the catchment. This equation indicates the volume changes in water storage over time by considering water fluxes inside the catchment. This assumption, however, fails to accommodate the anthropogenic variables, as this term might affect the storage capacity significantly. In a catchment scale, the water fluxes and storages in hydrological processes are illustrated in figure 2.5. These components in the hydrological cycle help to initiate the potential hydro-climatological variables to be feed to the statistical model.

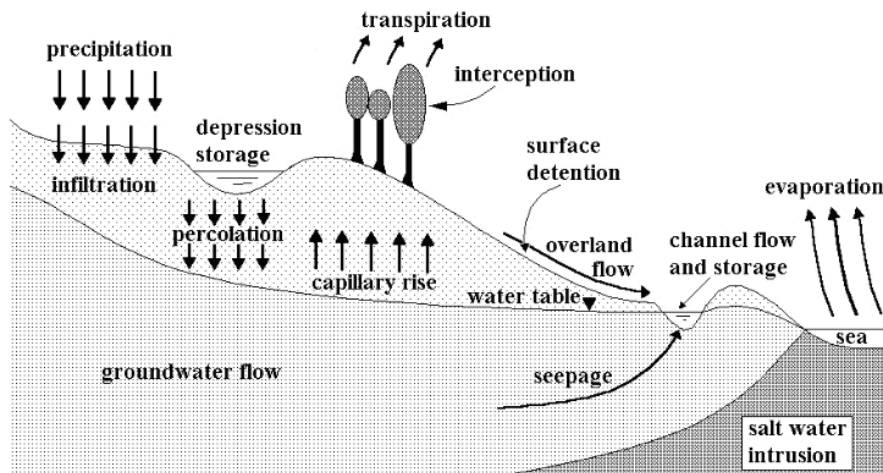


Figure 2.5: Descriptive Representation of water fluxes and storages in hydrological processes (source: Savenije [61], 2006)

As the focus of the study is to study the low flow in the river catchment, it is also critical to consider the adaptation behaviour of the run off due to its topography, soil characteristic and land cover. The aforementioned categories, in a catchment scale, allow us to understand more the alteration of the storage volume. Nonetheless, the parametrisation of these features is somewhat difficult to be obtained due to its variability. There is no measurement on the temporal development of these value, as this might changes also over time. The spatial variability could be answered by utilising the spatial datasets in a high resolution. However this problem is that a high-resolution dataset is costly and highly computational.

2.1.4. Drought Measurement in The Netherlands

In this subsection, an overview on how the drought was measured in the Netherlands will be explained. In general, the Netherlands characterises drought using two variables: (i) the mean precipitation deficit within the country and (ii) the discharge deficit of the Rhine River.

The precipitation deficit is defined as the cumulative difference between precipitation and grass reference evaporation, from 1 April onwards [2]. Any negative precipitation deficit during measured period will be set to zero. The annual maximum precipitation deficit happens during April until September, which also considers as the largest precipitation deficit. For example, figure 2.6 shows the graph of precipitation deficit in the Netherlands from April 2018 until September 2018 (half-summer year) (*neerslagtekort in Nederland in 2018*), which is indicated in the black line. The red line shows the extreme driest year recorded in the Netherlands in 1976; The green line indicates 5% exceedance of the years of precipitation deficits; The blue indicates the average recorded precipitation deficits in the Netherlands. 297 mm stated in the graph indicates the current updated actual measurement of precipitation deficit. The positive slope of the graphs exhibits the increasing trend of the drought. On contrary, the negative slope shows that the measure precipitation is greater than evaporation. Average precipitation were obtained by averaging the accumulated measure precipitation from 13 stations throughout the country. Whereas the grass reference evaporation were extracted using empirical formula (Makkink formula) that involves temperature and solar radiation. For further explanation of Makkink

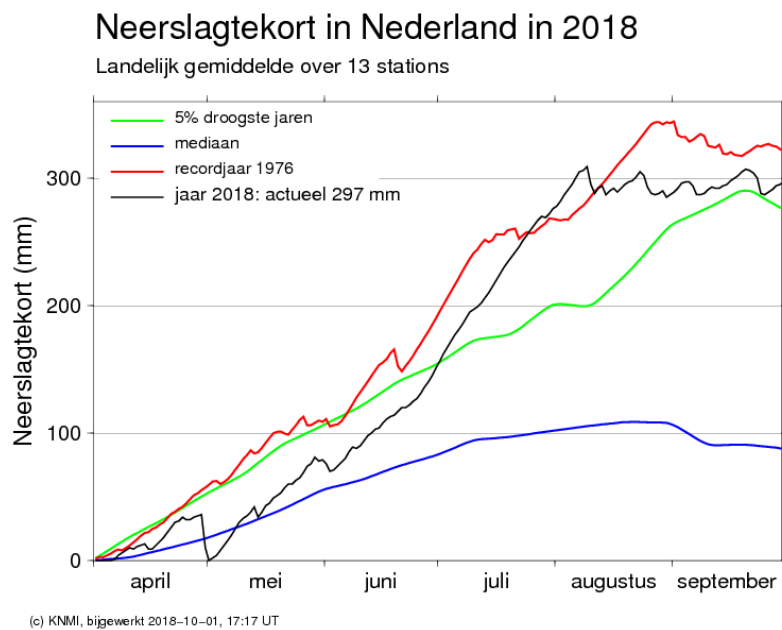


Figure 2.6: Accumulated precipitation deficit of 13 stations spread in the Netherlands in 2018 in black line. The x-axis indicates the time of half-year summer (April - September) and the y-axis shows the precipitation deficit in mm. (source: KNMI [36], 2018)

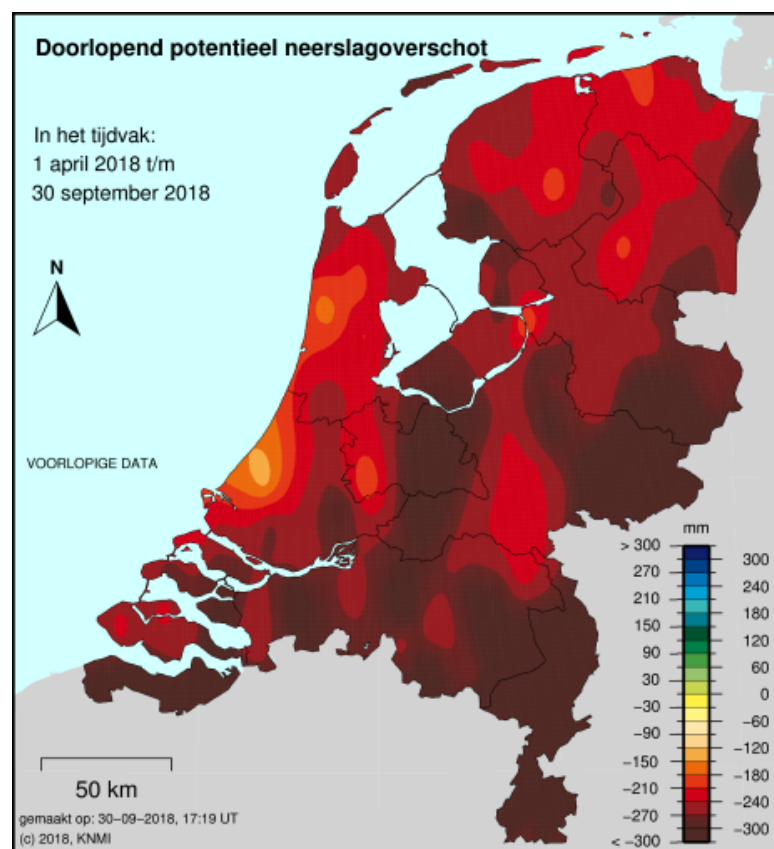


Figure 2.7: Potential precipitation surplus in half-year summer 2018 in the Netherlands. The negative values shows the shortage of precipitation, whereas the positive values indicates the surplus of precipitation source: KNMI [37], 2018)

formula, reader may refer to De bruin and Stricker [11].

As a general overview of how the drought spatially distributed in the Netherlands in 2018, KNMI also produces

a map of potential precipitation surplus (*doorlopend potentieel neerslagoverschot*). This map is depicted in figure 2.7. The legend indicates the precipitation surplus which was derived using also the difference between average precipitation and the loss of moisture expresses in grass reference evaporation. The positive value shows that there is immense of average precipitation compared to the evaporation, whilst the negative values indicates the shortage of average precipitation. From figure 2.7, it can be shown that east part of the Netherlands experience a higher shortage than the rest of the country. A long period of heatwave also hit the Netherlands in 2018 that makes the country had precipitation deficit almost throughout the half-year summer.

An additional variable to characterise drought in the Netherlands is the discharge deficit in the Rhine River. The measurement of this parameter takes place in the German-Netherlands border (Station Lobith). The discharge deficits were derived if only in the period of the prevailing measured discharge is below approximately 20% half-year summer of average discharge. This is one of examples of indicating hydrological drought using threshold method.

2.2. Case Study Area

As mentioned in the previous chapter, the focus point of the study is the State of California, US. In this thesis, a statistical model will be applied in California due to several reasons. California has suffered numerous episodes of drought events. Furthermore, California has a long historical time series of prospective variables, which are also openly accessible making the area appropriate for investigating new statistical techniques. These following sections will explain the system description of the State of California and a brief introduction to the drought in California.

2.2.1. System Description: California in a Nutshell

The state of California lies between 32° 32' to 42° north-latitude and 114° 8' to 124° 26' west-longitude [50]. Located on the Midwest of the US, California is bordered to the north by the state of Oregon, to the east by the state of Nevada and Arizona, by Mexico's state to the south, and by the Pacific Ocean to the west. The state boundaries of California is illustrated in Figure 2.9. The area of California is 160,000 square miles or equal to 414,000 square kilometres, which made California as the third most extensive by area in the US [50]. Based on Köpen-climate classification, California has majority hot/warm to summer Mediterranean climate. California's climate is varied and tends towards the extreme. In general, there are two major seasons, dry summer with low humidity and cool evenings and a mild-rainy winter. However in the high mountain areas, four seasons prevail, and the snow lasts from November to April [20]. The diversity of the landscape made California is the only state in the US with an extensive sea-coasts, high mountains and deserts [20]. Based on USGS land cover institution, the State of California is predominantly covered with shrub with around 40 percent of the state. Followed by evergreen forest with approximately 20 percent, it mostly covers the northern part of California. The area in the south west of the state is occupied by a medium to the high intensity of urbanised area. The most prominent area for agriculture, where utilised almost 10 percent of the area, is located in the middle part of California, or the so-called Central Valley. Figure 2.8 describes California land cover provided by USGS.

Although agriculture is accounted less than ten percent of the state's income, California provides more than half of the country's vegetables and fruits [19]. Nevertheless, the primary issue in California is the availability of the water resource, especially in the southern part. The annual water consumption of two-thirds of the southern part of the state is about 75%, whereas the supply is merely accounted for 30% [20]. Due to lack of water resources, more than 1,000 dams and reservoirs have been built in California to fulfil the water demand for agriculture and industry.

2.2.2. Drought in California: The Current and Historical State

USGS classifies drought in California in four ways, as discussed in section 2.1.1, and hydrological drought is defined as a period of one, or more water years, in which the below-normal availability of surface water and groundwater [72]. California has suffered from periodic droughts, which have had an impact on agricultural production which leads to the reduction of acreage and undergone commercial and residential development [19]. Figure 2.10 shows the historic California droughts since 1895. There were six drought episodes with two or more years event, except drought in 1976–1977. Albeit drought in 1976–1977 WY, which occurred in less than 2 years, was the most severe drought; the water supply was partially mitigated by surface and ground-

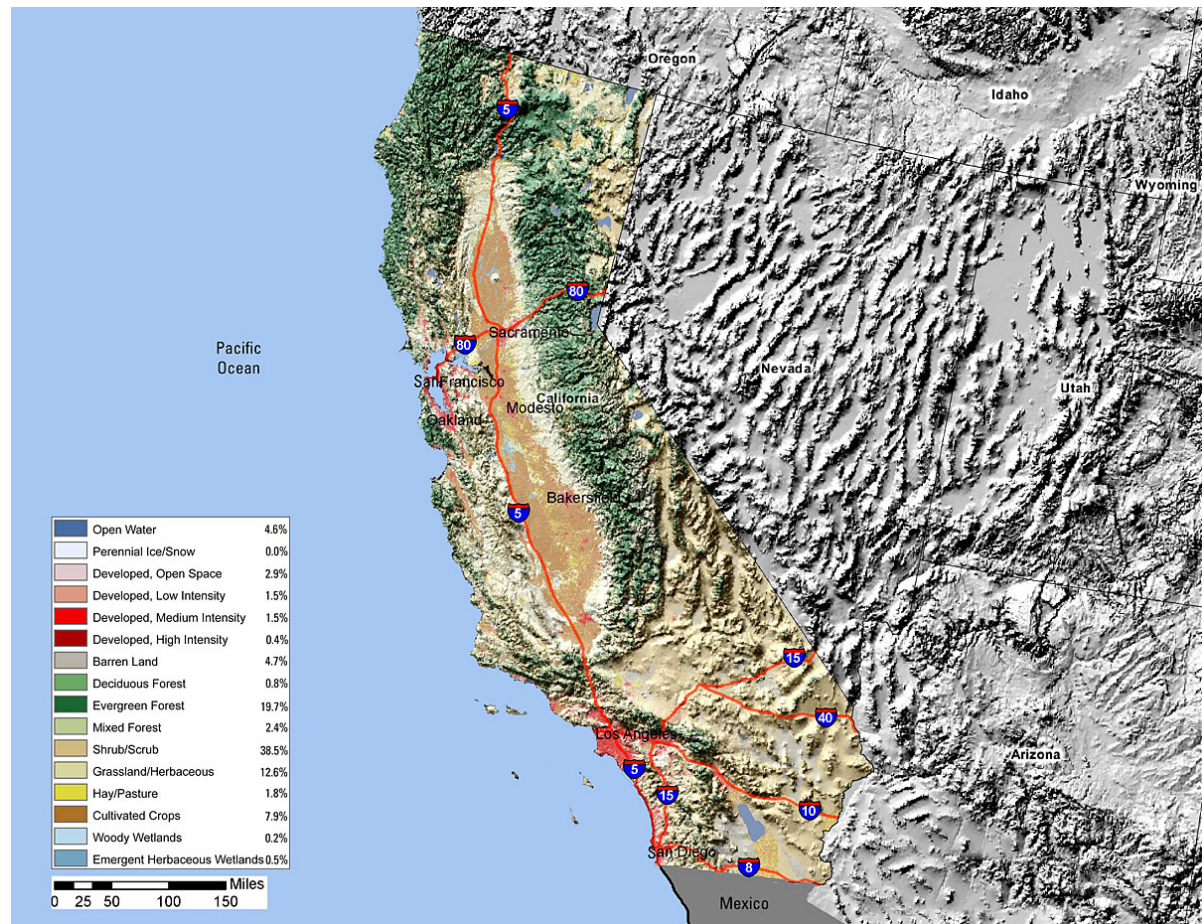


Figure 2.8: Land cover of California (source: USGS, 2016)

water storage [72]. From the figure, it could be concluded that the longer the duration of drought, the impact is more aggravated. For instance, within the same precipitation order of magnitude between drought of WY 1987–1992 and WY 2001–2002, the former drought event resulted in less run-off.

Runoff is one of a good parameter to measure the severity of drought impact since it is the final feedback from the hydrological cycle. The total runoff is accumulated from the preceding processes and conditions, e.g. the amount, location and type of precipitation, the amount of base flow, the prevailing soil moisture, the reservoir storage, the type of landscape and plant cover, and many more. Therefore, runoff indirectly resembles the impact of snowfall, rainfall and reservoirs. The 2012–2016 drought was reported as one the extreme proportions, with a high record of temperatures and low levels of snowpack and precipitation [52]. This event appears to be most severe drought event in the past two decades. It is depicted in Figure 2.11 that the drought severity and coverage index⁵ (DSCI) between 2012–2016 reached more than 375 and was the highest amongst the past two drought events. In January 2014, the state government of California declared a drought state of emergency. Nevertheless, in April 2017, the state government terminated the drought emergency, and at the same time, emerged the water conservation to maintain the water availability.

2.3. Bayesian Networks

Bayesian Networks (BN) are a class of probabilistic graphical models. They are represented by a Directed Acyclic Graph (DAG) that consists of nodes and arcs. The nodes represent random variables that can take on two or more possible values, whilst the arcs represent the existence of direct casual influences between the linked variables (nodes). In general, constructing a BN could be done by a simple procedure. That is (i) assigning a vertex to each variable in the domain; and (ii) drawing arrows towards each variable from a set

⁵DSCI or Drought Severity and Coverage Index is the weighted sum of drought severity category by United States Drought Monitor. (retrieved from: <http://droughtmonitor.unl.edu/AboutUSDM/DroughtClassification.aspx>)

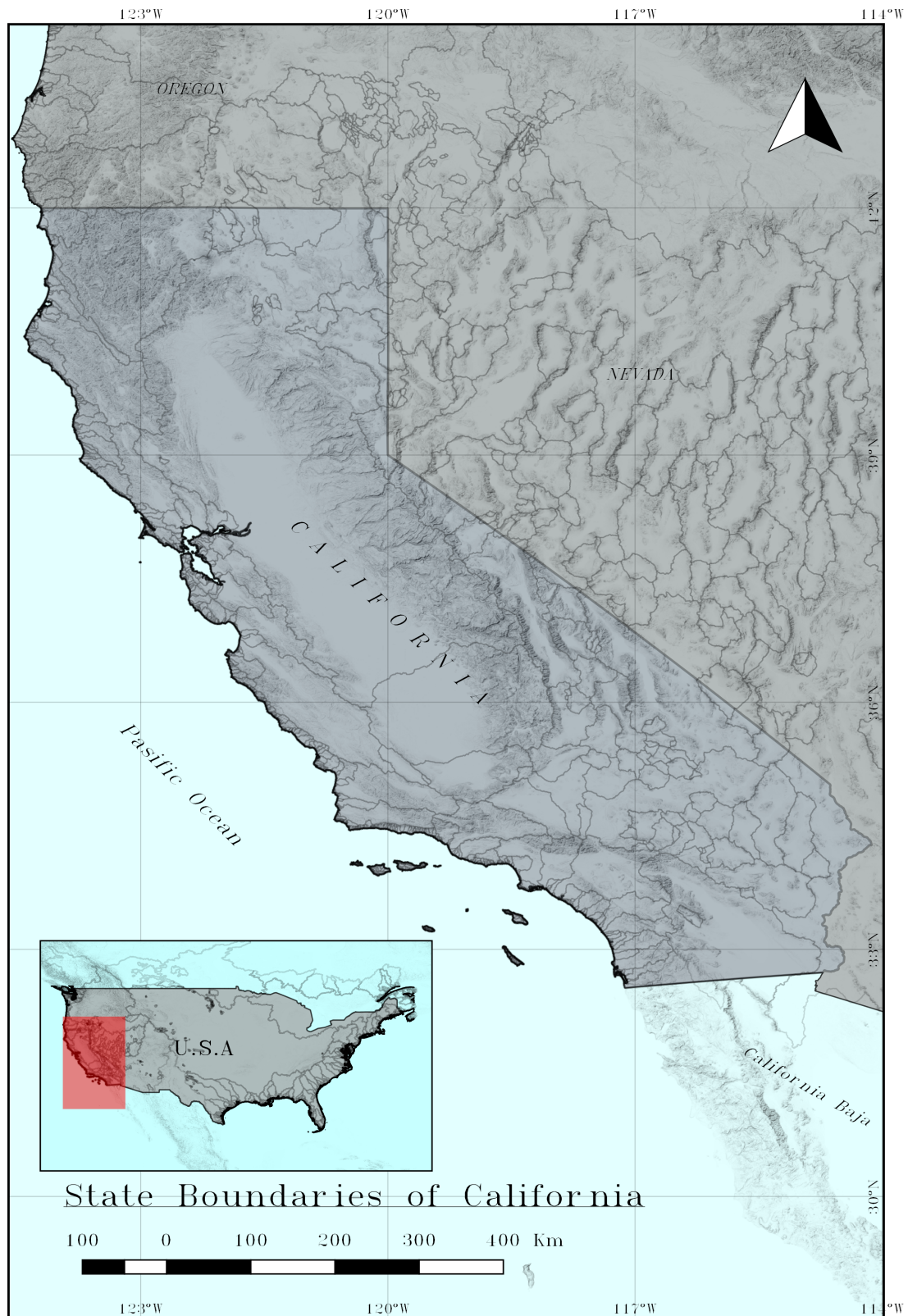


Figure 2.9: Geographical borders of California

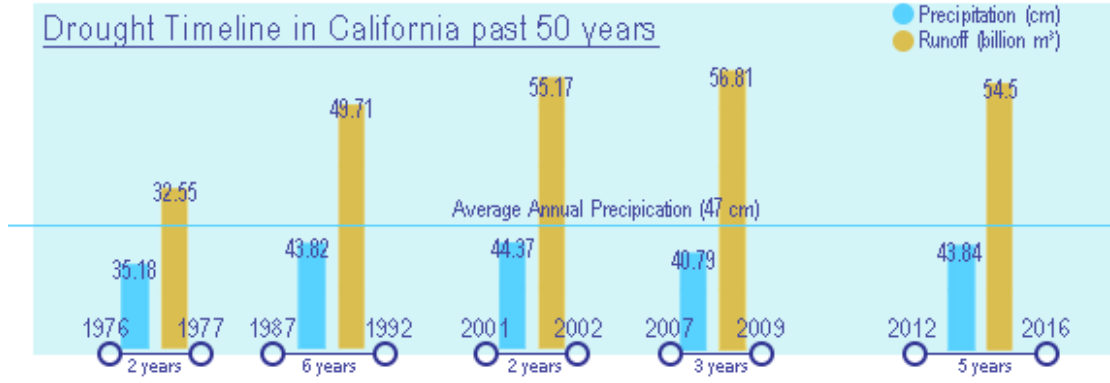


Figure 2.10: Average annual runoff and precipitation for historical droughts in the state of California, data is taken from USGS [71]

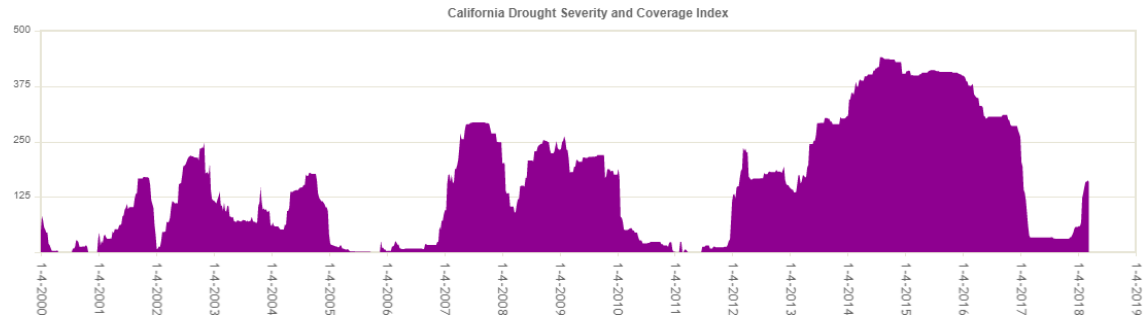


Figure 2.11: Graph of California DSCI for the past two decades. The trend has shown an increasing severity, duration and areal coverage. (source: The National Drought Mitigation Center [69], 2018)

of other variables to represent the causal relationships. The significances of the influence are shown by the quantification of conditional probability functions. The joint distribution over random variables $X_1, X_2, \dots, X_n$ is stated in the following equation:

$$f_{X_1, \dots, X_n}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \prod_{i=2}^n f_{X_i|Pa(X_i)}(x_i|X_{Pa(X_i)} = x) \quad (2.2)$$

where $f_{X_1, \dots, X_n}$ denotes the joint probability function of the n variables, f_{X_i} is their marginal probability function, $f_{X_i|j}$ denotes the conditional probability function, $X_{Pa(X_i)} = x$ is short hand notation for $X_{Pa_1(X_i)} = x_{Pa_1(X_i)}, \dots, X_{Pa_m(X_i)} = x_{Pa_m(X_i)}$ and $Pa(X_i)$ is a set of m parents of node X_i .

As a brief explanation of DAG, figure 2.12 demonstrates several simple BN structures (left: converging connection; middle: diverging connection; right: serial connection). In general, these graphs present the acyclic relation, which imply the joint probability function between variables. In the hierarchical perspective, the immediate predecessors are the *parents* of the immediate successor nodes, whereas the successor nodes are the *children*. For example in figure 2.12, the middle graph shows node F as the child of parent variables, node D and E. It shows a conditional dependence between D and E, given F is known. Another interesting

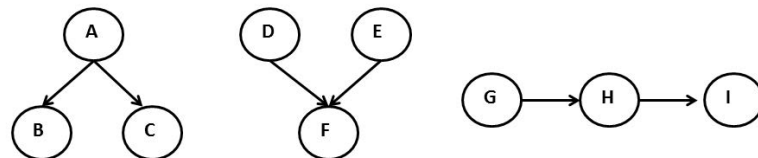


Figure 2.12: Examples of a simple Bayesian Network Belief with different structures (source: Couasnon [8])

information that could be inherited is from the right graph in figure 2.12, is that having information about G

does not change the information about I, once H is known. In other word, G is conditionally independent of I.

Later in the application, these type of graphs will be constructed in more robust networks. The advantages of network representation is that it allows us to express directly the fundamental qualitative of "direct dependency". The network then displays a consistent set of additional direct and indirect dependencies and preserves it as a stable part of the model, independent of the numerical estimates. This explicit dependence is appropriate for the complexity of drought studies. Bayesian Networks have been applied in many projects in different fields [30], e.g. risk analysis, reliability analysis, prediction model, etc. The use of BN in this study is inspired by the application to quantify the maximum riverine discharge in Europe, done by Paprotny and Morales in 2015. BN has shown a compelling model in statistical analysis, in combination with the conceptual studies, to understand the risk.

2.3.1. Non-Parametric Bayesian Networks

As the concept of BN has been introduced in the previous subsection, Non-Parametric Bayesian Networks (NPBN) are part of continuous BNs. In NPBN, there is no assumption to be made about the marginal distribution of the variables (nodes), rather extracted from the data or elicited from the experts [8]. The arcs are expressed in the (conditional) rank correlations between the variables that is established by bivariate copulas.

In a smaller scale, it is common to use bivariate analysis to test the dependency between two random variables. Generally, this analysis is easier to perform if two variables follow the same distribution, which is not always the case. This problem could be solved using Copula, that avoid this restriction. The copula approaches to dependence modelling is rooted in a representation theorem due to Sklar in 1959 [27]. Copula normalized the variables into a unit square distribution, or alternatively saying that copula transformed the variables into uniform square unit [0,1]. Under bivariate analysis of continuous random variables, the cumulative distribution function (CDF) of two variables is expressed as a copula in the following equation

$$H(x, y) = C\{F(x), G(y)\}, \quad x, y \in \mathbb{R} \quad (2.3)$$

where $H(x, y)$ is the bivariate CDF; $F(x)$ and $G(y)$ are the marginal distribution of two random variables. When Standard Normal Copula is not able to represent the data distribution, there are also another different type of copulas under Archimedean Family, such as Frank Copula, Clayton Copula, Gumbel Copula, etc. However to date, the most renowned copula is Gaussian or Standard Normal Copula, due to its rapid calculation in large multivariate model. The Bivariate Gaussian Copula is presented in the following equation:

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) \quad (2.4)$$

where Φ is the standard normal distribution, Φ^{-1} is its inverse and Φ_ρ is the bivariate Gaussian cumulative distribution with (conditional) product moment correlation ρ of u and v are the uniform variates in the interval [0,1].

Moving forward to the correlation between the copula, this part will the translate the mathematical computation of (conditional) rank correlation. The arc j between variable X_i and its n parent $Pa_1(X_i), \dots, Pa_n(X_i)$, $Pa_j(X_j) \rightarrow X_i$ is parametrized as follows:

$$\begin{cases} r(X_i, Pa_j(X_i)) & j = 1 \\ r(X_i, Pa_j(X_i) | Pa_1(X_i), \dots, Pa_{j-1}(X_i)) & j = 2, \dots, n \end{cases} \quad (2.5)$$

where j is the index in the non-unique sampling order. The empirical normal rank correlation approaches the empirical rank correlation if the normal copula assumption holds and given a large enough number of observations [8]:

$$r(X_i, X_j) = \rho(F_{X_i}(X_i), F_{X_j}(X_j)) \quad (2.6)$$

r is the normal rank correlation, and ρ is the product moment correlation. By definition, product moment correlation between two random variables is the ratio between covariates of two random variables and the product of the standard deviation of two random variables. Therefore, the product moment correlation between X_i and X_j is formulated as follows

$$\rho_{X_i, X_j} = \frac{cov(X_i, X_j)}{\sigma_{X_i} \cdot \sigma_{X_j}} \quad (2.7)$$

where cov is the covariance between X_i and X_j , σ is the standard deviation. Under bivariate normal copula assumption, a simple characteristic relation between the Spearman's rank correlation r and the Pearson's product moment correlation coefficient ρ exist:

$$r = \frac{6}{\pi} \cdot \arcsin\left(\frac{\rho}{2}\right) \quad (2.8)$$

The benefit of using joint normal distribution is that the conditional and partial correlation are equal [30]. Therefore the correlation matrix could be computed from the partial correlation. Partial correlation defines the correlation between two variables given certain circumstances to be hold, as this is relevant to the conditional rank correlation. Partial correlation is formulated in the following equation:

$$\rho_{12;3\dots n} = \frac{\rho_{12;3\dots n-1} - \rho_{1n;3\dots n-1} \cdot \rho_{2n;3\dots n-1}}{\sqrt{1 - \rho_{1n;3\dots n-1}^2} \cdot \sqrt{1 - \rho_{2n;3\dots n-1}^2}} \quad (2.9)$$

Inference in NPBNS

Performing inference in NPBNS could be handled by sampling it. In the case that other types of copulas than the Gaussian Copula are used, performing inference becomes more complicated. This condition may lead to a big drawback when it comes to a real-time decision making. However, this problem could be solved if one use the normal copula. Using normal copula allows the user to operate the joint normal distribution with the same rank correlation structure as the original one [30]. Therefore, the properties of the joint normal distribution could be used to transform the conditional rank correlation into the conditional product moment correlation. As mentioned earlier in section 2.3.1, the conditional and partial correlation are equal. Furthermore, one could sample from these normal joint distribution given the correlation matrix, then transform it back to the marginal distribution. By performing this method, a sample from the joint distribution of the original variables and the dependence structure realized by the normal copula become attainable.

2.4. Prediction Power for BNs

In this section, the technique to perform validation will be explained as an important part of identifying whether the model is a fair representation of the data. The validation method will be used in combination with the error measurement to quantify the acceptable error in the model. Validation also helps to justify the acceptable error in the model, in which the result still considers in a good performance. This validation method and measure will be used only to test Bayesian Networks in this research.

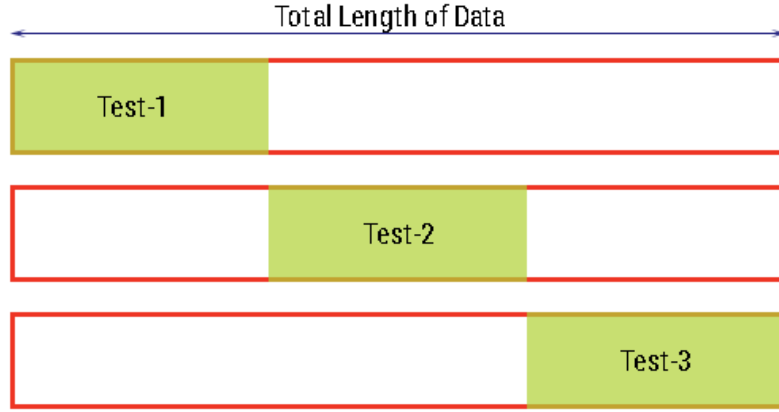
2.4.1. Method for Result Validation: K-fold Cross-Validation

There are many method to validate the model performance. In this research, K-fold cross-validation is utilized as it is the most common sub-sampling techniques[57]. The basic concept of K-fold cross-validation is that all data will be divided into K equal parts, as one part will be used as the test data, whereas the remaining will be treated as the training data set. This process will be repeated until every equal parts are being tested or K times. For instance, as illustrated in figure 2.13, K value is equal to 3, meaning the data will be divided into three parts. In other word, there are three different scenarios to be validated in the model. The first scenario defines the test-1 as the testing data set (green colour) and the rest as the training data set (white colour). All scenarios will tested and deliver different error measurement.

Furthermore, many researches have been conducted in order to determine the value of K for which K-fold cross-validation fits the best. In contrary, many also found satisfactory results of K-fold using arbitrary number, depending on the setting of the model [57]. The benefit of using K-fold cross-validation is that using larger the K value will reduce the bias of true error rate estimator, while the variance of true error rate estimator will be large [57].

2.4.2. Error Measurements: R^2 and NSE

As explained in section 2.4.1, the validation of the BN model will be quantified by estimating the true error value, as it is also aimed to set a threshold for indicating the realization of the model. There are many performance measures to validate the result of the model, however, in this research, two methods will be used, namely coefficient of determination (R^2) and Nash-Sutcliffe Efficiency (NSE). The coefficient of determination represents the linear relation between the estimated and the observed data. It describes the proportion of the variance measured data explained by the model [10]. R^2 has 0 to 1 value range, with higher value indicates

Figure 2.13: Example of K-fold cross-validation with $K = 3$

the less error variance. Nash-Sutcliffe Efficiency defines the relative values of the residual variance compared to the measured data variance [10]. NSE ranges between $-\infty$ to 1.0, a value between 0 and 1.0 represents an acceptable performance, as 1.0 is the optimal value, meanwhile less than 0 signifies a poor performance [10]. The both error measurements are formulated in the following equations:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Q}_i - \bar{Q})^2}{\sum_{i=1}^n (Q_i - \bar{Q})^2} \quad (2.10a)$$

$$NSE = 1 - \frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^n (Q_i - \bar{Q})^2} \quad (2.10b)$$

where $\hat{Q}_i$ is the estimated value of the discharge from the model, Q_i is the measured value of the discharge, and $\bar{Q}$ is the mean value of the observed value.

2.5. Generalised Linear Models

Generalised linear model (GLM) ⁶ is a flexible extension model of classical linear regression model. The classical linear regression model could be expressed mathematically in a general form of

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon; \quad Y_i \sim N(\mu_i, \sigma^2) \quad (2.11)$$

where random variables Y_i are independent and mostly in continuous form. However, GLM enables responses variables to have distribution other than normal and not necessarily have linear relation to explanatory variables. This relationship is expressed via link function. Principally, a GLM has three main components:

1. Variates that denote as $Y_1, \dots, Y_N$ are assumed to have same distribution from the exponential family in the canonical ⁷ form.
2. A set of parameters $\boldsymbol{\beta}$ and explanatory variables

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & & x_{Np} \end{bmatrix}$$

⁶In general linear model, some terminologies should be introduced. Assumed that data is represented by a data matrix, a two dimensional array in which the rows are indexed by experimental or survey units. These units are the physical items on which observations are made [45]. The column in data matrix are so called the variates, or the variable of interest. Variates are associated as responses or dependent variates, that might be influenced by the explanatory variables or covariates. The form of covariates is flexible, either qualitative or quantitative.

⁷Canonical indicates a particular choice from a number of possible convention, which allows a mathematical object has a unique standardization [59].

3. A monotone link function g such that

$$g(\mu_i) = \mathbf{x}_i^T \beta + \epsilon$$

where

$$\mu_i = E(Y_i)$$

There are two important characteristics of GLM. Firstly, it assumes that the observations are independent of each other or at least uncorrelated. This assumption restricts data that exhibit autocorrelations of time series and spatial processes. Secondly, regarding error structure, it is assumed that there is only a single error term in the model. As a result, this model excludes models with more than one explicit error term. Another advantage using GLM is the reduction of scaling problem. Since normality and constancy of variance are no longer required, it allows us to only focus on the estimation of the additivity of effects. In this thesis, the GLM will be focused on logarithmic link function, which will transform the general equation (2.11) into:

$$E(Y_i) = \mu_i = \exp^{x_i^T \beta}; \quad Y_i \sim \text{Poisson}(\mu_i) \quad (2.12)$$

where the link function is

$$\ln(\mu_i) = x_i^T \beta \quad (2.13)$$

GLM with logarithmic link function follows a Poisson Distribution. The Poisson Distribution is often used to model count data [14]. Usually in order to model using GLM, an adequate length of time series are needed to yield an accurate fit and result, both for the response and the explanatory variables.

2.6. ANOVA: Analysis of Variance

This section will be divided into three parts to explain the types of ANOVA that will be used later in explaining the result of GLM. The underlying concept of ANOVA is analysing the variability in the data to see how much it can be attributed to difference in true mean value and how much is due to variability in the individual populations [56]. It looks at each of the distributions and compares the between-group differences with within-group differences [56]. This can be done by computing two type of sum of squares from the data: sum of squares within groups and sum of squares between groups (that will be demonstrated later). Hence, ANOVA is a tool to analyse the means of several groups by identifying sources of variability in the data. ANOVA also enables to evaluate more than two independent groups that have different means on a particular measure and eliminates the error rates resulting from multiple t-tests⁸ (or testing of hypothesis for two samples). In this study, ANOVA will test the formulated hypothesis using p-value⁹.

One-way ANOVA

A one-way ANOVA is a statistical test to compare the variance of the mean in different groups by considering one independent factor. A one-way ANOVA holds a null hypothesis that there is no difference between means of different groups. The null hypothesis of one-way ANOVA is stated mathematically as follows:

$$\begin{aligned} \text{Null Hypothesis}(H_0) & : \mu_1 = \mu_2 = \dots = \mu_k \\ \text{Alternative Hypothesis}(H_a) & : \text{not all the } \mu'_i \text{'s are equal} \end{aligned}$$

where $\mu_1, \dots, \mu_k$ is the mean of k normal populations with unknown but equal variance σ^2 .

The procedure of one-way analysis for $k \leq 2$ populations

The following steps are taken from Ramachandran and Tsokos [56], for completely randomized design ANOVA.

1. Assumed that Y_{ij} , $i = 1, \dots, k$, and $j = 1, \dots, n_i$ be the data of completely randomised sample with k is the number of randomised samples, whereas n_i is the size of the randomised sample- i . That becomes, $y_{11}, y_{12}, \dots, y_{1n_1}$ for population 1, and $y_{k1}, y_{k2}, \dots, y_{kn_k}$ for population k . Therefore, compute:

$$T_i = \sum_{j=1}^{n_i} (y_{ij}); \quad T = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij}); \quad \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij})^2$$

⁸For further explanation of t-test, reader may refer to Ramachandran and Tsokos [56] on Chapter 7, section 7.5 regarding testing of hypothesis for two samples.

⁹By convention, if the p-value is less than 0.05 ($p < 0.05$), we can conclude that the null hypothesis (H_0) can be rejected. Alternatively saying that when p-value is below 5%, the result is statistically significant, or there is evidence(s) to suggest that the null hypothesis is false.

where T_i is the sum of all the observations in the i th sample. Let also CM be the correction factor for the corrected means, and is given by:

$$CM = \frac{(\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij})^2}{N} = \frac{T^2}{N}, \quad \text{where } N = \sum_{i=1}^k (n_i),$$

and we could also stated the mean of the observations in the i th sample, as: $\bar{T}_i = \frac{T_i}{n_i}$, combining the previous definition, the total sum squares is

$$Total\ SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij})^2 - CM$$

2. Compute the sum of squares between samples (SST) (treatments),

$$SST = \sum_{i=1}^k \frac{T_i^2}{n_i} - CM$$

hence, the sum of squares within sample (SSE) becomes

$$SSE = Total\ SS - SST$$

3. Having the sum of squares between and within the sample, we are able to calculate the mean of both sum of squares. That is:

$$\begin{array}{ll} \text{Mean square between groups} & MST = \frac{SST}{k-1} \\ \text{Mean square within groups} & MSE = \frac{SSE}{n-k} \end{array}$$

4. Then the test statistic is

$$F = \frac{MST}{MSE}$$

Normally using F-stats we could check whether the value falls in the rejection regions. However, in this research we are going to transform the F-stats to p-value, and use p-value to test the null hypothesis. The computation of p-value is done using statistical software (MATLAB). Thenceforth, the p-value will be retrieved from MATLAB result

Based all the aforementioned procedures, the calculation of one-way ANOVA can be displayed in table 2.1. This table will be also later used in presenting the result of ANOVA in section 4.2.4.

Table 2.1: Summary calculation of the one-way ANOVA

Source of Variation	Degree of Freedom	Sum of Squares	Mean Squares	F-Statistics
Treatments	$k - 1$	$SST = \sum_{i=1}^k \frac{T_i^2}{n_i} - CM$	$MST = \frac{SST}{k-1}$	$\frac{MST}{MSE}$
Error	$n - k$	$SSE = Total\ SS - SST$	$MSE = \frac{SSE}{n-k}$	
Total	$n - 1$	$Total\ SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij})^2 - CM$		

Two-way ANOVA

Similar to one-way ANOVA, two-way ANOVA enables to measure response at combinations of levels of two or more factors considered simultaneously and to give knowledge about the likely interactions among different factors. Having said that, the concept of the calculation is similar to one-way ANOVA, and expanded to have more source of variations in the calculation. Two-way ANOVA also checks the means equality for considered factors. Therefore, the null hypothesis for each variation is that there is no difference between means in the group. It also adds the null hypothesis about interactions of the factors, that formulated as follows:

$$\begin{array}{ll} \text{Null Hypothesis}(H_0) & : \text{there is no interaction among different factors} \\ \text{Alternative Hypothesis}(H_a) & : \text{there is interaction among different factors} \end{array}$$

Multiple Comparisons

In one-way ANOVA, if there is no equality of means among the groups (*null hypothesis is rejected*), there is no further information regarding which of the given group of means differs significantly from another give group of means. This missing information could be answered using multiple comparisons. Multiple comparisons is a method to compare each pairwise combinations.

In general, the null hypothesis for multiple comparisons is formulated as follows:

$$\begin{aligned} \text{Null Hypothesis}(H_0) &: \mu_i = \mu_j \\ \text{Alternative Hypothesis}(H_a) &: \mu_i \neq \mu_j \end{aligned}$$

for all $i \neq j$. The concept of testing pairwise combination in Tukey's Method may sounds similar to the testing of hypothesis for two samples. However, the difference that the estimated p-value for Tukey's Method is corrected for the comparison of multiple groups, since performing multiple times of pairwise combinations from populations using t-test might cause a bias to reject the H_0 . Tukey's method helps to avoid the incorrect finding of a significant difference as the number of comparisons increases. Since the number of sample of each groups are different, therefore *Tukey-Kramer Method*¹⁰ is to be used in this analysis.

This part is taken from Ramachandran and Tsokos [56]:

Theorem of Tukey's Method: Let $\bar{T}_i, i = 1, 2, \dots, k$ be the k sample means in a completely randomized design. Let $\mu_i, i = 1, 2, \dots, k$ be the true means and let $n_i = n$ be the common sample size. Then the probability that all $\binom{k}{2}$ differences $\mu_i - \mu_j$ will simultaneously satisfy the inequalities

$$(\bar{T}_i - \bar{T}_j) - q_{\alpha, k, v} \frac{s}{\sqrt{n}} \leq \mu_i - \mu_j \leq (\bar{T}_i - \bar{T}_j) + q_{\alpha, k, v} \frac{s}{\sqrt{n}} \quad (2.14)$$

is $(1 - \alpha)$, where $q_{\alpha, k, v}$ is the upper α critical value of the Studentized range distribution. If, for a given i and j , zero is not contained in the preceding inequality, $H_0 : \mu_i = \mu_j$ can be rejected in favor of $H_a : \mu_i \neq \mu_j$, at the significance level of α .

¹⁰It is an extension method of original *Tukey Method* for unequal sample data

I

Hydrological Drought Severity

Hydrological Drought using Bayesian Network

This chapter describes the result of hydrological drought modelling using Bayesian Network, which implies the analysis of hydrological drought severity. The motivation of using such tool is that BN has been applied in many different fields, and is an ample model to understand risk (see section 2.3). BN has also been applied to understand the characteristic of drought. For example, study in 2016, by Shin et al. [63] using BN, tried to forecast the meteorological drought in which limited to the use of drought indices. Thus, the innovation of this research is to include other types of parameters (meteorological and hydrological variables) to study the characteristic of hydrological drought (see section 1.2). The output of the analysis is the BN graph to estimate mean monthly discharge. This chapter will be dissected into four parts. It will begin with preliminary stages of constructing the BN. This section explains on which data to be used in the model to estimate the low flow, as well as defining the main assumptions beyond the model systems. The following section - building up the BN for hydrological drought - presents the sequence of steps on how the BN is built. The third section which is result and discussion, will elaborate the small scenarios to study the model behaviour of the BN final graph. Later, the discussion part will explain several hypotheses of the drawbacks of the model and several recommendations to improve the model performance. The last section - conclusion - delivers the main findings of building up the final BN hydrological droughts.

3.1. Preliminary Stages of Constructing the Bayesian Network

This section delivers the description of the model system for the BN and the stage of preparing the datasets of the input variables for the model.

3.1.1. Model Assumption for the BN

This subsection will explain the assumption to be made beyond the hydrological drought BN and the system description of the input variables in California. In general, the hydrological drought BN makes use of the combination between the concept of rainfall-runoff and drought propagation. On one hand, the concept rainfall-runoff is exhibited by the input and the output of the model. On the other hand, drought propagation plays a role in expressing the structure of the BN, that will be shown later. The outcome of the model is the mean monthly discharge, in which represent the low flow expressing the hydrological drought severity. Additionally, the scope of the input variables will be limited to the meteorological and hydrological variables. Therefore, any variables that are considered or classified as anthropogenic variable will not be included. Also, any depletion of fresh water in the river due to water quality problem is not treated in this research. As indicated in section 2.1.2, to establish the model, one should consider the spatial constraint of the catchment. One of the projects that provides the information of watersheds, is called Geospatial Attributes of Gages for Evaluating Streamflow II (GAGES II). There are 810 point of gages along with the drainage basins in California, as depicted in figure 3.1. This product of drainage basin makes a distinction between catchment with and without human influence. There are 158 stations and watersheds in California that are considered reference gages. Reference gages are the catchments that are classified as full natural flow (FNF), it has no canal, flumes, or man-made channels built within the catchments.[21]

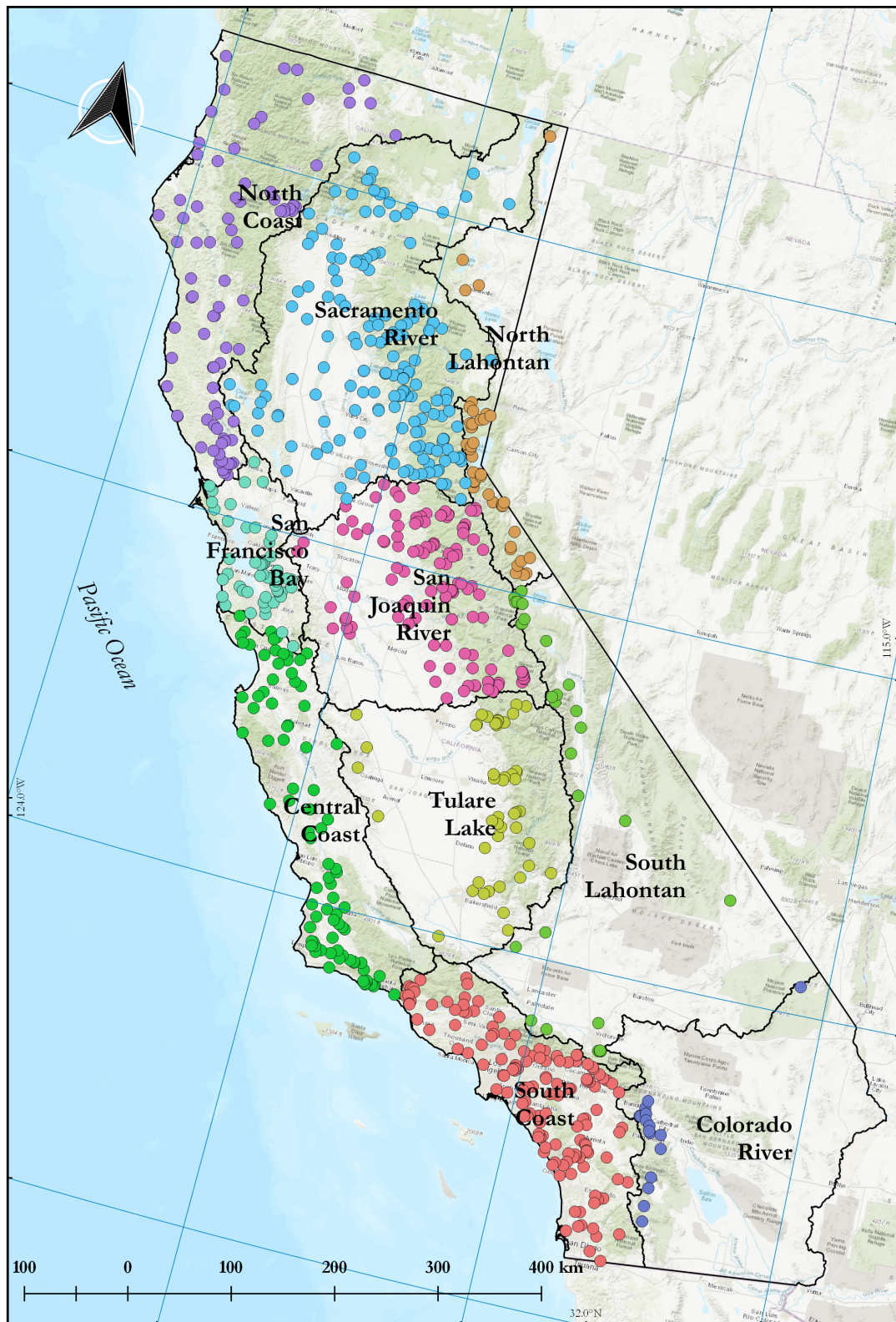


Figure 3.1: Points of discharge gages in California. There are 810 points along with the drainage area delineated by GAGES II, USGS

3.1.2. Input Variables: Extracting Hydrological and Meteorological Variables

This part presents the details of the input variables that are possibly relevant in estimating the variable of interest. As mentioned in section 1.3, the input variables will be classified into two classes, meteorological

and hydrological variables. For further explanation on the concept or definition of the variables, the reader may refer to appendix A. An example of one station will be also presented to give the reader idea about the data to be used in the model.

Meteorological Variables

There are many meteorological variables in nature that are involved in the hydrological cycle. In general, meteorological variables in the hydrological cycle are formulated in the water balance equation, which implies the gain and loss of water fluxes in a catchment scale. This study makes use of spatial database from CORDEX. Table 3.1 summarized the variables that are possibly relevant for explaining drought¹¹.

Table 3.1: Table of details of the meteorological variables in the model

Data Type	Raw Unit	Nodes	Data Source	Year	Resolution
Mean Monthly Precipitation Flux	kg/m^2s	PR_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Snowfall Flux	kg/m^2s	PRSN_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Evaporation	kg/m^2s	EV_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Air Temperature	K	TAS_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Surface Temperature	K	TS_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Vapour Content	kg/m^2	PRW_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Snow Melt Flux	kg/m^2s	SNM_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
SPI	-	SPI	NADM	1950-2017	N/A

Hydrological Variables

This category of variables is describing the catchment characteristic such as geometrical components or the catchment control. These variables are an important part to characterise the feedback of the drought mechanism. Catchment control could be derived from the hydrological variables mostly related to the catchment characteristic. In this research, all the data were taken from GAGES II database. Table 3.2 summarized the hydrological parameters that used in the model

Table 3.2: Table of details of the hydrological variables in the model

Data Type	Raw Unit	Nodes	Data Source	Year	Resolution
Mean Runoff Coefficient	-	MRRO_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Mean Monthly Soil Moisture	kg/m^2	MRSO_MEAN	CORDEX NAM-44	1951-2005	26.5 arcminutes
Catchment Area	m^2	Area	GAGESII	-	N/A
Compactness	-	Compact	GAGESII	-	N/A
Stream Density	-	Density	GAGESII	-	N/A
Base Flow Index	-	BFI	GAGESII	-	N/A
Saturation Overland Flow	-	SOF	GAGESII	-	N/A
Infiltration Overland Flow	-	HOF	GAGESII	-	N/A
Topographic Wetness Index	-	TWI	GAGESII	-	N/A
Subsurface Contact Time Index	-	SCTI	GAGESII	-	N/A

Example of the Input Data

In order to give an idea of the type of data to be used in the model, an example of the dataset for Station Tahquitz C NR Palm California is presented. Table 3.3 reveals some technical information regarding the station and the several basin characteristics parameters. Located in Riverside county in the southeast California, station 10258000 is considered as the reference gages with full natural flow within the catchment. Moreover, figure 3.2, 3.3 and 3.4 illustrate the time series variation for Station 10258000. There are 14 time series of different variables.

¹¹ 26.5 arcminutes equals to approximately 49 km

Table 3.3: Example of technical information and coefficient for Station 10258000

Station	10258000
Latitude	33.805
Longitude	116.558
County	Riverside
State	California
Class	Reference
Area	43.691
Compactness	2.146
Stream Density	0.871
BFI	49.638
SOF Coefficient	0
HOF Coefficient	0
Topography Wetness Index	8.239
Subsurface Contact Time Index	2.070

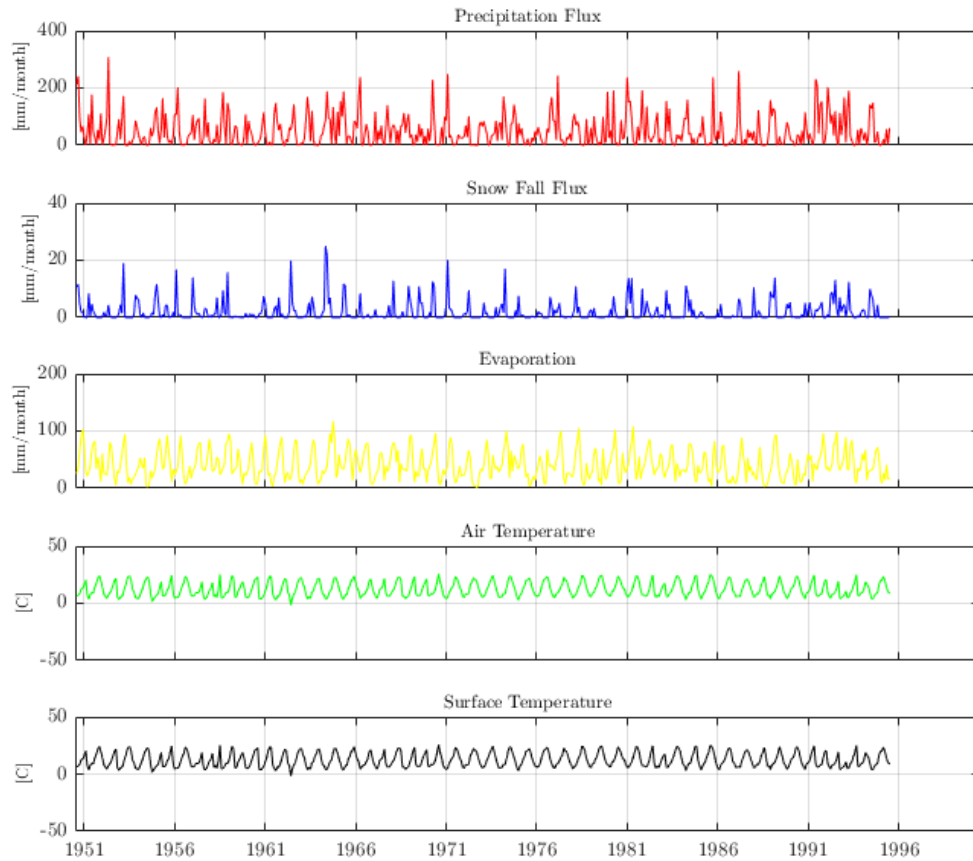


Figure 3.2: Example of time series of variables in station 10258000 (1)

3.2. Building up the BN for Hydrological Drought

In this section, the process in obtaining the final BN for hydrological drought is explained. In general, the final hydrological drought BN is obtained using following procedures: scenario sensitivity analysis, statistical investigation and conceptual approach. The final BN graph shows the most relevant variables in explaining the hydrological drought, and the final outcome of this analysis is try to estimate the low flow.

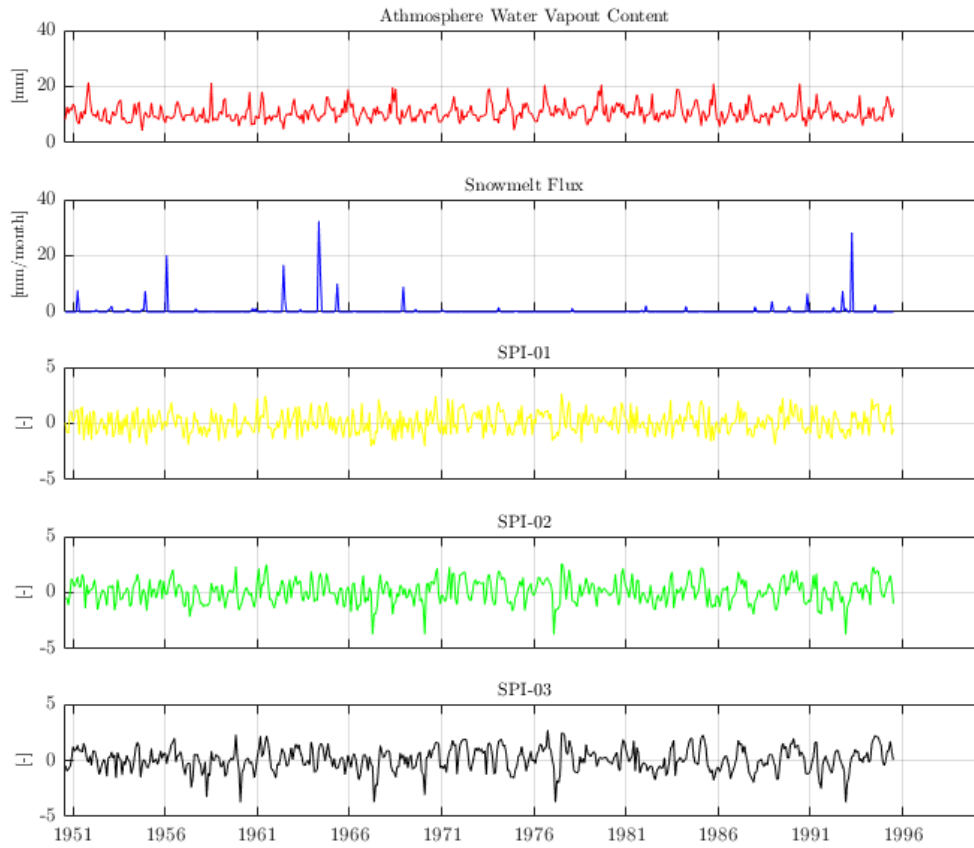


Figure 3.3: Example of time series of variables in station 10258000 (2)

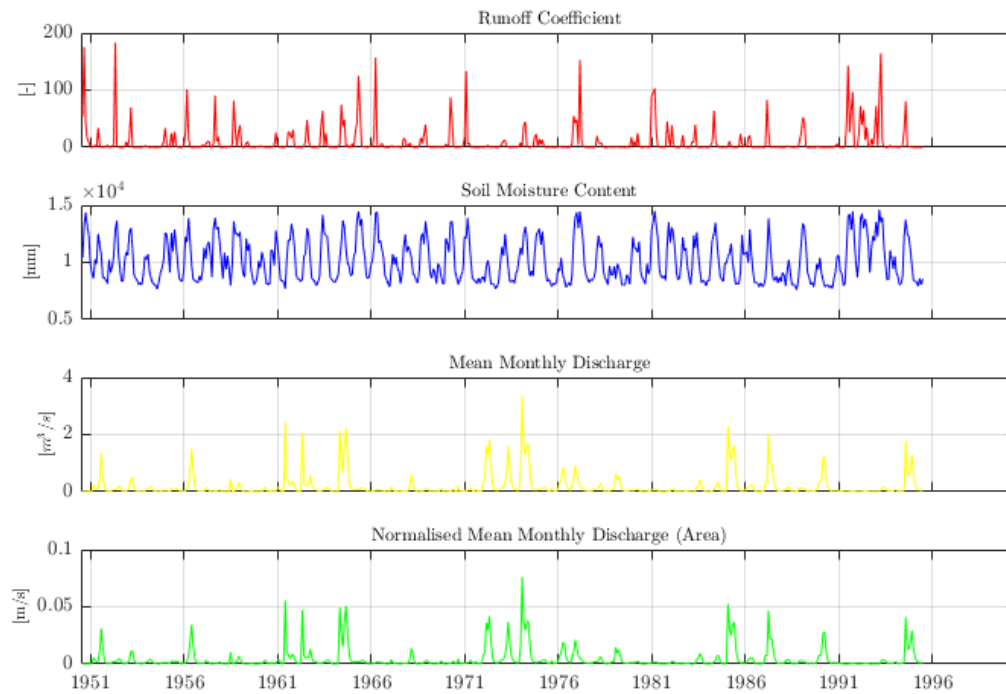


Figure 3.4: Example of time series of variables in station 10258000 (3)

3.2.1. Scenario of the Sensitivity Analysis

The scenario is made to provide the framework of the stepwise analysis in gaining the most relevant variables for best describing drought mechanism. In general, the sensitivity analysis is classified into two big stages, the statistical stage and conceptual validation. The iterative process on selecting the variables is performed by eliminating the least correlation coefficient in the BN, in combination with the study of drought processes. Thence, it allows the model to behave as identical as possible to the drought mechanism. To have a better picture of the structure of sensitivity analysis, figure 3.5 presents the flow diagram.

Sensitivity Analysis: Selecting the Relevant Variables

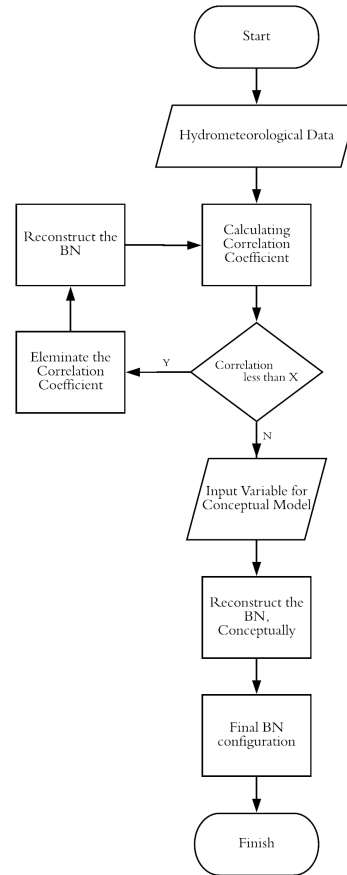


Figure 3.5: Flowchart diagram of sensitivity analysis

3.2.2. Statistical Approach

In order to investigate on which input variables are related to generation of low stream-flow, the first step to start with, is to compute the correlation coefficient matrix. This would be a very handy tool to see the statistical correlation between two variables. The matrix of correlation coefficient is depicted in figure 3.6. The legend of the matrix indicates from -1 (blue) to 1 (yellow). The attribute of the value shows the relationship between variables. Positive value means that both variables grow in the same direction, while negative value means that one variable grows in different direction towards the other variable. Figure 3.6 shows the correlation coefficient for classes such as meteorological variables, index, hydrological variables and variable of interest. For variable of interest, it consists of the mean monthly discharge and normalised mean monthly discharge by area (this variable will be explained in the following part).

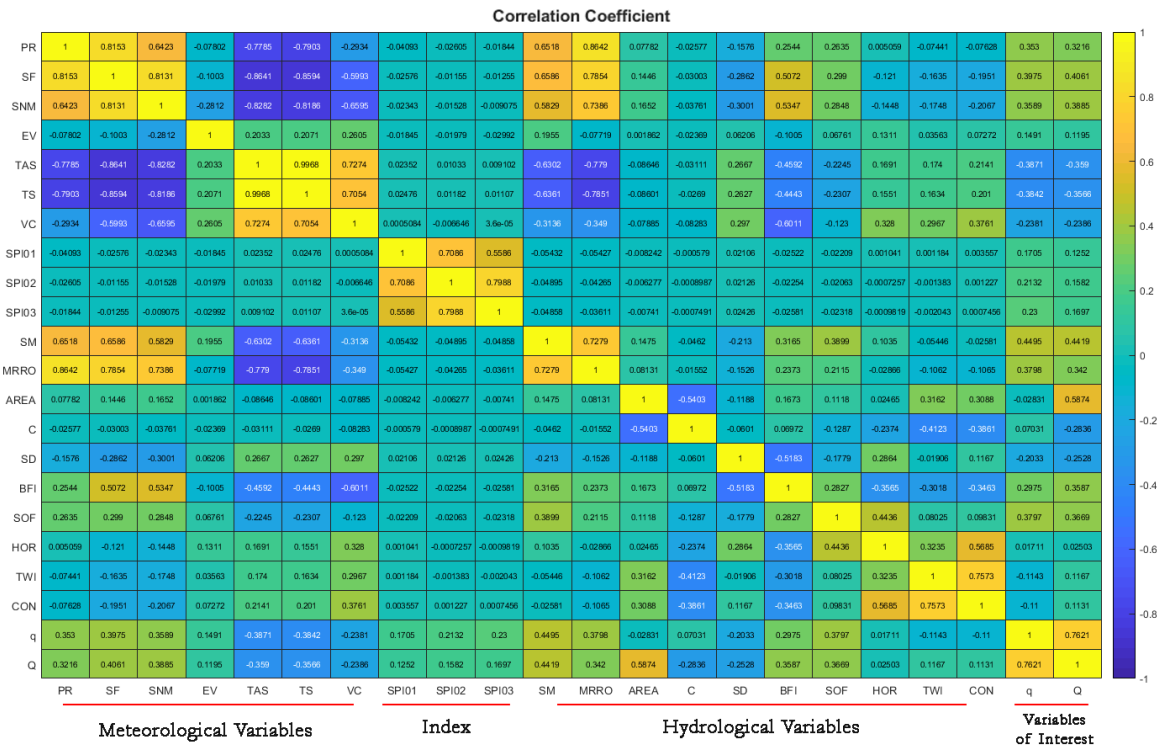


Figure 3.6: The correlation coefficient matrix for variables

Scaling Factor for the Variable of Interest

As has been stated in section 1.2, that hydrological drought in this context will be focussed in analysing the low flow. Therefore, the variable of interest will be expressed in mean monthly discharge (Q_{MM}). The time series of mean monthly discharge was extracted from USGS, based on the 810 points of drainage basin shows in figure 3.1. In this subsection, the variable of interest will be normalised into discharge per catchment area. Normalising the mean monthly discharge by dividing it with its catchment area will help to disregard the bias of the size of the catchment in the estimation of the low-flow. In addition to that, scaling relationships between turnover or discharge and water system size may help to reveal and understand the general patterns and processes in global and regional hydrological systems [26, 31, 58]. The normalised mean monthly discharge denotes as mean monthly discharge per catchment area (q_{MM}) or specific discharge. Using discharge per catchment area has shown a moderate increased in terms of correlation coefficient between the input variables and variable of interest. The data in table 3.4 were extracted from figure 3.6. From table 3.4, it is shown that meteorological class is correlated better with this stream flow per catchment area, whereas the hydrological class has a better relationship with original discharge. Starting from this point, one is able to reduce variables that has correlation coefficient less than 0.1. Those variables are Area, Compactness, and Hortonian Overland Flow. This made the number of input variables decreased to 17 variables from 20 initial input variables.

Semi-correlations for variables

Moving forward to the next statistical check, the remaining variables will be tested using semi-correlation. The semi-correlations are the Pearson's product moment correlation coefficients computed in the upper and lower quadrants of the normal transforms of the original variables [53]. Using this result of analysis, one could check the tail dependence for the considered bivariate analysis. Furthermore, the assessment could also imply whether the Gaussian Copula (see subsection 2.3.1) used in BN is able to represent the bivariate distribution. The tail dependence could be evaluated by comparing between the overall product moment correlation and the semi-correlations. If at least one of the absolute semi-correlations is larger the overall correlation, one could indicate that there is tail dependence in a certain quadrant. Having said that, Gaussian Copula might fail to represent the relationship between two random variables. A further information of the analysis is elaborated in appendix B

Table 3.4: Spearman's Rank correlation coefficient comparing the result of Discharge and Discharge per Catchment Area

Input Variable	Q_{MM}	q_{MM}
Meteorological Variables:		
Precipitation Flux	0.3216	0.3530
Snowfall Flux	0.4061	0.3975
Snow Melt Flux	0.3885	0.3589
Evaporation	0.1195	0.1491
Air Temperature	-0.3590	-0.3871
Surface Temperature	-0.3566	-0.3842
Atmosphere Water Vapour Content	-0.2386	-0.2381
Index:		
SPI-01	0.1252	0.1705
SPI-02	0.1582	0.2132
SPI-03	0.1697	0.2300
Hydrological Variables:		
Runoff Coefficient	0.3420	0.3798
Soil Moisture Content	0.4419	0.4495
Area	0.5874	-0.0283
Compactness	-0.2836	0.0703
Stream Density	-0.2528	-0.2033
Base Flow Index	0.3587	0.2975
Saturated Overland Flow	0.3669	0.3797
Hortonian Overland Flow	0.0250	0.0171
Topography Wetness Index	0.1167	-0.1143
Subsurface Contact Time Index	0.1131	-0.1100

In the pre-analysis of the variable of interest, several variables were discarded due its low correlation. It resulted in 17 variables to be tested using semi-correlations. Table 3.5 shows the semi-correlations of 17 variables with q_{MM} that were inherited from figure B.1, B.2, and B.3. It is evident that some variables, show

Table 3.5: Table of semi-correlations between all input variables and q_{MM}

Input Variables	ρ	ρ_{NW}	ρ_{SW}	ρ_{NE}	ρ_{SE}
Precipitation Flux	0.3530	0.1371	0.1217	0.0666	0.1513
Snowfall Flux	0.3975	0.1325	0.1494	-0.0458	0.1753
Evaporation	0.1491	0.0668	0.1037	0.0086	0.0883
Air Temperature	-0.3871	-0.1983	-0.1502	-0.2225	0.1144
Surface Temperature	-0.3842	-0.2123	-0.1444	-0.2249	0.1284
Vapour Content	-0.2381	-0.0830	-0.0971	-0.1195	0.0780
Snow Melt Flux	0.3589	0.1590	-0.0016	0.0098	-0.0046
Runoff Coefficient	0.3798	0.1451	-0.0700	0.1125	0.0678
Soil Moisture Content	0.4495	0.1182	0.1768	0.2117	0.1052
Stream Density	-0.2033	-0.0550	-0.1193	-0.0907	-0.0035
Base Flow Index	0.2975	0.0405	0.1225	-0.1319	0.0105
SOF	0.3797	0.1071	-0.0634	0.0883	0.1019
Topography Wetness Index	-0.1143	-0.0039	-0.0792	-0.1186	0.0644
Subsurface Contact Time Index	-0.1100	0.0050	-0.1264	-0.0904	0.0846
SPI - 01	0.1705	0.0266	-0.0145	0.1985	0.0472
SPI - 02	0.2132	0.0255	0.0077	0.2101	0.0836
SPI - 03	0.2300	0.0441	0.0238	0.1929	0.0926

in table 3.5, appear to have tail dependence. Those variables (indicate in bold) are TWI (Topography Wetness Index), Subsurface Contact Time Index, SPI-01, SPI-02 and SPI-03. For instance, the semi-correlation for quadrant NE ($\rho_{NE} = -0.1186$) for variable TWI, is larger than the overall correlation ($\rho = -0.1143$). All those

variables will be eliminated as the relationship fails to be well represent by Gaussian Copula. Another variable, such as surface temperature, will also be discarded as it behaves the same as air temperature. Thus, this statistical analysis left us with only 11 variables as the input for building up the conceptual model. Those are air temperature, atmosphere water vapour content, evaporation, snow melt flux, precipitation flux, snowfall flux, soil moisture content, SOF coefficient, base flow index, stream density, and runoff coefficient.

3.2.3. Conceptual Approach

In this subsection, the process of introducing the assumptions in the model will take place (see subsection 3.1.1 regarding the model assumption). This is done by connecting the arcs between variables that resemble the causality. This step will be done in an iterative manner, and will be validated using the R^2 and NSE .

Removing Arcs from the Saturated BN

Before we start to build the model based on the concept of water balance. Another statistical check will be performed, in order to test whether the current variables from the previous stage are completely true. The motivation of this step is to check whether the variables are directly informative to the estimation of q_{MM} or helpful to give more information towards other variables (conditionally dependent), as well as removing the arcs that found to be meaningless in the BN. We are going to proceed this section using Uninet as the tool. At first, variables obtained from previous stage will be constructed in saturated graph of BN. Saturated is a condition where in the BN there is an arc between every pair of variables (see figure 3.7 (a)). We are going to remove several arcs from the saturated graph of BN that has correlation less than 0.2 (see figure 3.7 (b)). After removing few arcs from the saturated BN, we need to again compute the correlation coefficient matrix

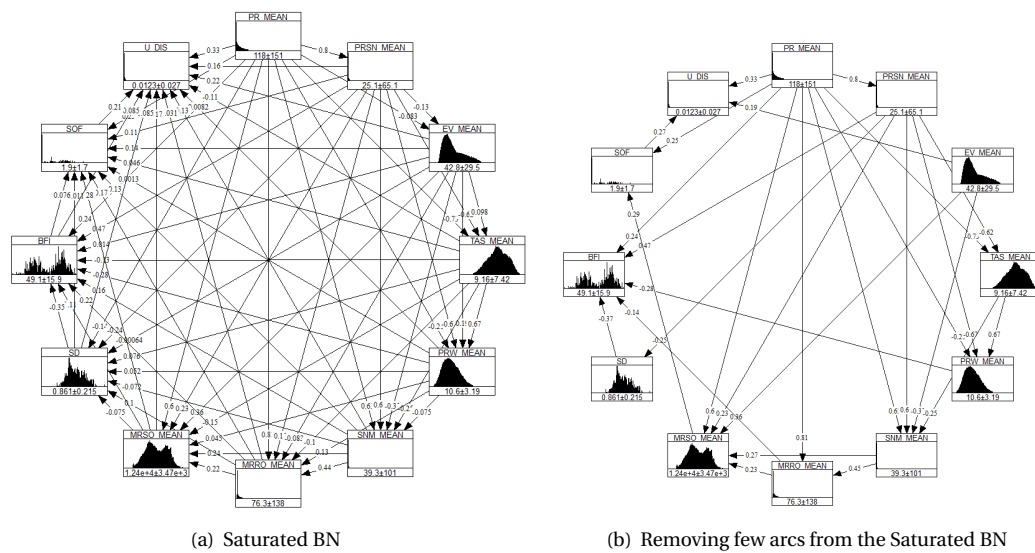


Figure 3.7: Saturated Check on the BN

to check whether there is any variables that is the least correlated towards each other in the system. This process will be repeated until there is no correlation coefficient that less than the minimum threshold (0.2). This value is set because the correlation below than 0.2 is found to be extremely insignificant in the system, although value of 0.2 is also insignificant. Then, another one last check on the BN correlation coefficient will be again performed to see how the variables correlated towards the set of parent variables. Figure 3.8 shows the final removal of BN rank correlation matrix obtained from UNINET. This matrix summarises that after removing some arcs of the system, all variables are still informative (or at least correlated toward each other) to estimate the q_{MM} . From this check we could conclude that the eleven variables will be considered in the conceptual model.

Building the Network Conceptually

As explained in section 2.1 regarding low-flow, we are going to assume the rainfall-runoff concept in the water balance equation in a catchment scale as the starting point of building-up the graph. At first, we are going to classify the variables based on its function in the hydrological processes. The input will be categorised into

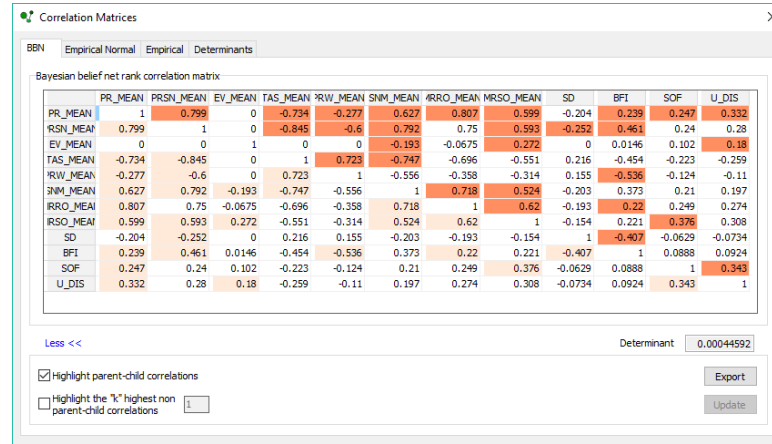


Figure 3.8: BN rank correlation matrix

three classes, viz. climatological characteristics, water fluxes, and the catchment characteristics. By having this class, it easier to connect the arcs based on the causality of the behaviour in nature. In general, the climatological characteristics, will be the trigger of gain and loss in water fluxes in the hydrological process, and later the water fluxes will adapt the runoff controlled by the characteristics within the catchments. Figure 3.9 portrays the variables for each category.

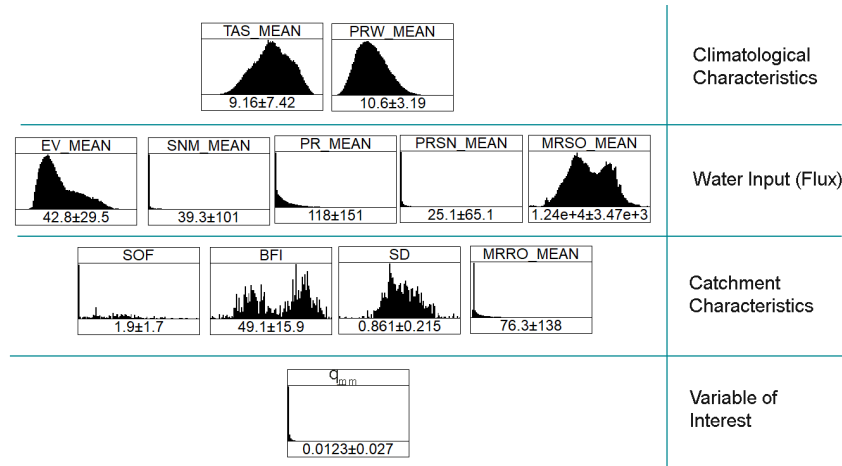


Figure 3.9: The categories for the input variables

THE RAW CONCEPTUAL MODEL – In this trial, we are going to include only climate characteristics and water fluxes in order to estimate the low flow in California. This will be the very basic water balance model that involve only the water fluxes. The validation of the model takes part after the BN has been built. As explained in section 2.4, the validation measures will be done using K-fold cross validation using the whole datasets as the training, and the 10-folds of testing datasets. The result of 10 times validating the model will be presented in the box plot in figure 3.11. The constructed conceptual model shows the mean NSE of 0.104 and mean R^2 of 0.108

THE FINAL BN GRAPH OF CONCEPTUAL MODEL – Building up the final BN is done by introducing the variable one by one into the system. The process is performed in a stepwise way, by also checking the correlation coefficient matrix and at the same time cohering the model with the concept of rainfall-runoff. The final outcome of the hydrological drought BN is shown in 3.12 (a). Performing the same setting of validation method to the final BN, the model shows a slightly better performance (NSE = 0.165 and R^2 = 0.143). The box plot of the validation measures of final BN depicts in figure 3.12

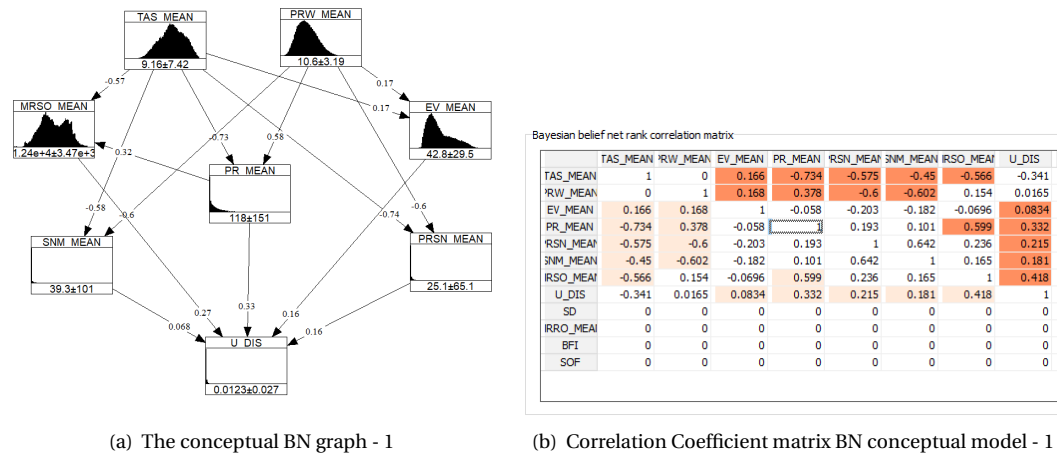


Figure 3.10: The result of BN graph-1 using UNINET

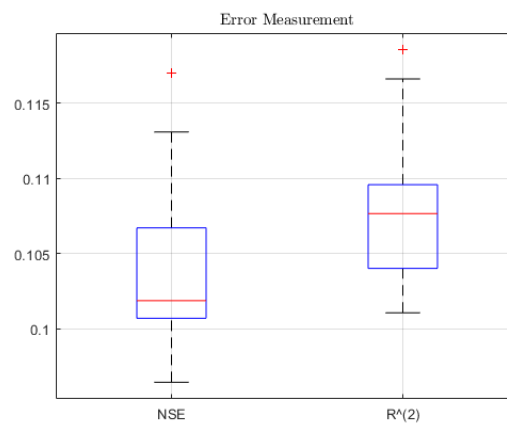


Figure 3.11: Box plot of validation measures of BN conceptual model - 1

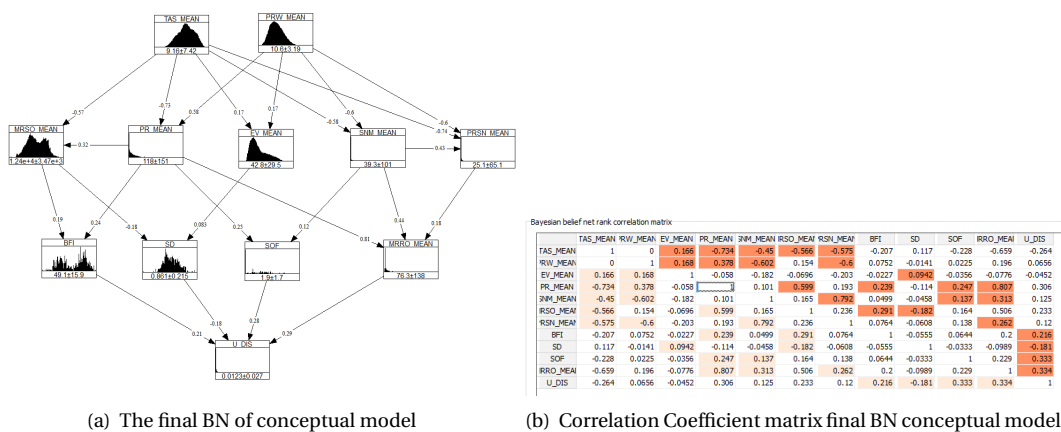


Figure 3.12: The result using UNINET

3.2.4. The Graphical Result of BN for Hydrological Drought

After went through all the mentioned processes, the result of the BN for Hydrological drought is depicted in the following figure 3.14. It has eleven variables involved in foresee the mean monthly discharge per area in California. This BN also demonstrates the part of the hydrological processes in nature. For instance, the climatological characteristics (air temperature and atmosphere water vapour content) are considered as the

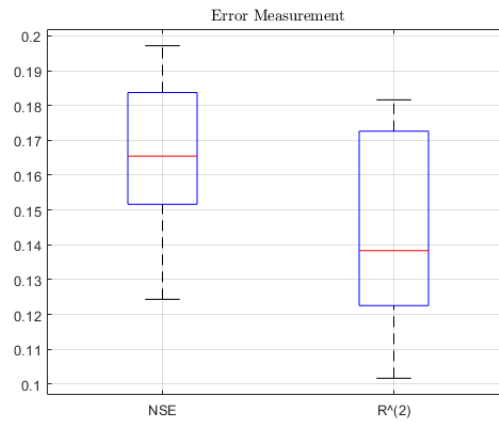


Figure 3.13: Box plot of validation measures of final conceptual BN

trigger of the water fluxes. At the same all the water fluxes in the might also differ based on the catchment characteristic to estimate the specific mean monthly discharge (q_{MM}). Although, the final BN has been obtained, there is still unsatisfactory aspect in terms of the model performance. This will be explained in the following section.

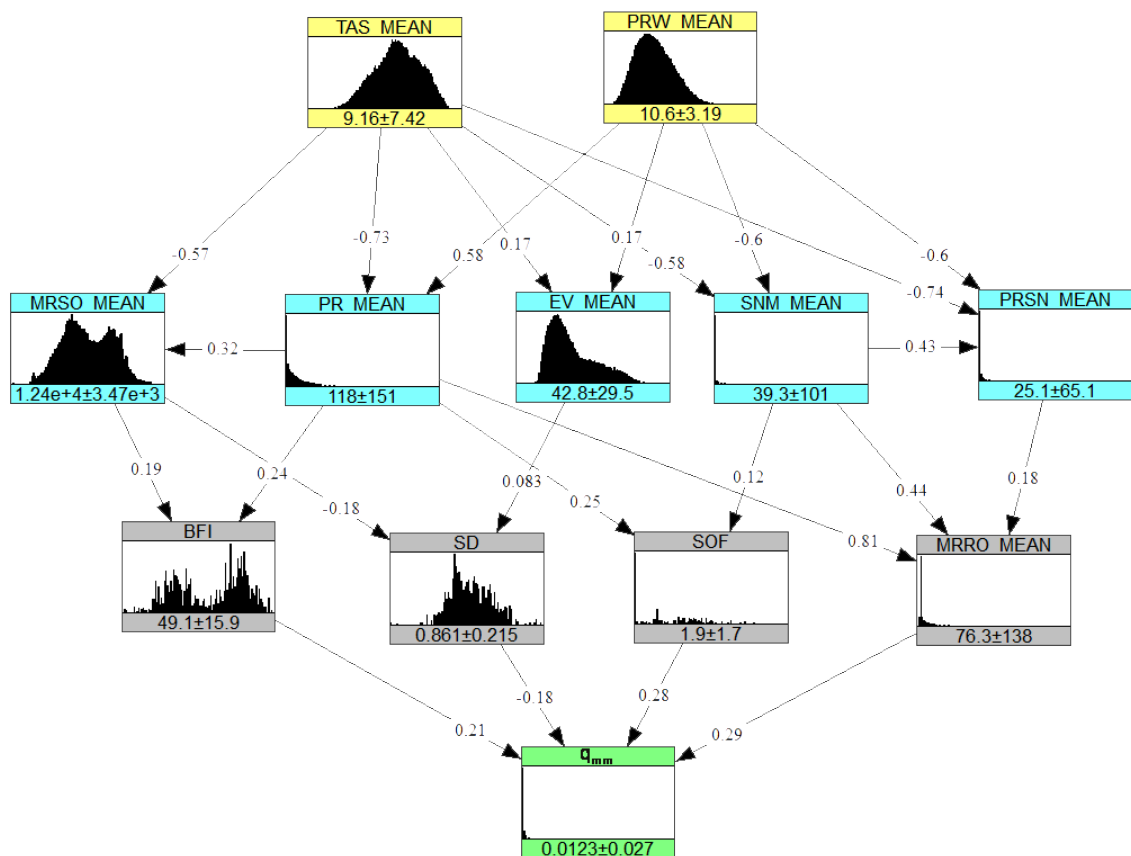


Figure 3.14: Final result of Bayesian Network for explaining hydrological drought using meteorological and hydrological variables

3.3. Results and Discussion

This section delivers scenarios to study the model behaviour of the BN Hydrological Drought, and discussion part which tells some recommendations in order to improve the model.

3.3.1. Inference Result: Comparing Different Data Input

Having unsatisfactory result of validation in the previous section, we would like to test the hydrological drought BN model under different types of data. It aims to study whether using different type of data will increase the performance of the model. The hydrological drought BN model in figure 3.14 was built upon the dataset in California. All variables were computed to such could be able to represent the value of low-flow in a catchment scale. However, a study conducted by Loon [40] demonstrates a different types of hydrological droughts. Different types of hydrological droughts were considered to have different ways of modelling. This is happening due to different governing processes, climates and catchments control.

Having that in mind, now we are going to test whether the response of the structure hydrological drought BN model could differ under different data sets. The current data set will be classified using two different independent factors, the season and the type of catchment. The seasonal period will be divided into 2 types, winter (December until May) and summer (June until November). Whereas the type of catchment is the full and non-full natural flow (see subsection 3.1.1). It is important to bear in mind that, when using the model, we are only using the structure of BN without inherit the correlation information from the existing model. This is aimed to test whether such structure could perform better towards which types of hydrological drought conditions. The test is carried out by taking 1000 numbers of testing data randomly for each category. Figure 3.15 shows the inference of hydrological drought BN under different types of datasets.

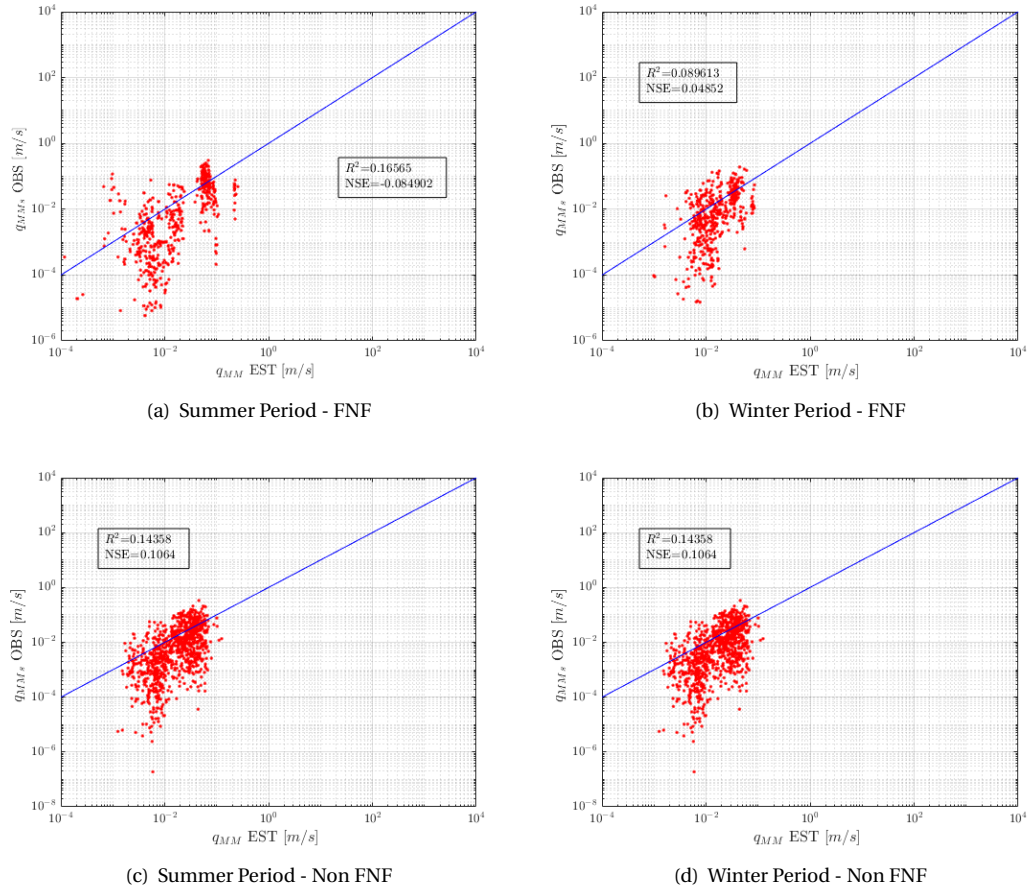


Figure 3.15: Inference result for different types of data classification in logarithmic scale. The plot shows between the observed of q_{MM} in y-axis and the estimated value of q_{MM} in x-axis

Moreover, table 3.6 summarises the result of validation measures for four types of data. It could be concluded from the table that the model foresees better the non-FNF catchment. The model also captures better the summer season rather than winter.

Table 3.6: Validation measures of different types of data

Types of classification	R^2	NSE
Summer Season and FNF	0.1656	-0.0849
Winter Season and FNF	0.0896	0.0485
Summer Season and Non-FNF	0.1436	0.1064
Winter Season and Non-FNF	0.1308	0.1284

3.3.2. Discussion

Seeing the poor estimation of the result of mean monthly discharge in California using BN, raises many question to be explored. There are many possibilities that are worth mentioning for future recommendations.

Modelling Assumption

Defining clear assumptions is the most critical aspect when it comes to model drought. These assumptions should be made thoughtfully to resemble the characteristics of hydrological drought. In a catchment scale, determining different types of hydrological drought on the top of understanding the mechanism would be a very critical stage. In a drought management, different types of droughts lead to different types of preventing measures and coping mechanisms [40]. This becomes evident based on the response of the model shows in

figure 3.15, which was varying under different types of input data. Furthermore, different types of drought leads to a different types of input variables. Therefore, a clear assumption of the type of hydrological drought should be made prior to building the model. In conjunction with the study case in California, that in fact has a high variation of climate virtues, makes it challenging to classifying the types of drought in a regional-scale model. Also in a bigger scale, this is still one of the major scientific challenges which is related to the diversity of causal mechanisms of hydrological drought [40, 43].

Missing Input Variable

When it comes to building a model, an important stage is to see how the data is used in general, and how each contribute to the mechanism. Drought processes is a complex mechanism. Even though the general processes could be explained using drought propagation, there are still variables that are not included in the model. Such variables are really hard to be measured and usually controlled by the complex system of the catchment. For example, the time lag of which the meteorological drought were translated to hydrological drought might play a big role of quantifying the mean monthly discharge. A very simple method to measure this variable is to analyse the time difference between meteorological and hydrological drought.

The uncertainty of input datasets

Regional climate dataset, like CORDEX, is a handy spatial dataset that are openly accessible. This makes the process to obtain data becomes easier, especially to places where ground measurements are scarce. However, database like CORDEX are also a product of modelling that has certain limitation. For instance, the uncertainty of the product of the modelling; and the other most common issues working with spatial datasets is the resolution that leads to the inaccuracy of the model. Therefore, some corrections on the data set might be a possible solution in having a good way to start with.

3.4. Conclusion

The purpose of this part of analysis is to build a BN to build a model to estimate low-flow using hydro-meteorological variables. However, the final hydrological drought BN model has shown a very poor estimated result, that is demonstrated by the validation measures ($NSE = 0.165$ and $R^2 = 0.143$). Several hypotheses were proposed to explain the result of the model BN hydrological drought in California and to improve the performance of the model. For instance, the time-lag of the drought propagation or the assumption of hydrological drought. Therefore, a conclusion to be drawn from this chapter that the hydrological drought BN (depicted in figure 3.14) is not an adequate model to be used to estimate low-flow, particularly in California.

II

Hydrological Drought Duration

Hydrological Drought using Generalised Linear Model

This chapter presents the analysis result of hydrological drought using Generalised Linear Model. Many studies had been conducted applying GLM as the tool to gain as much as possible information regarding climate condition. However, these studies are usually limited to the generation of synthetic rainfall and its duration using other climate predictors. For instance, a study by Leith [39] makes use of GLM to simulate the daily rainfall under different climate scenarios, including climate change. Later in 2013, a study by Chun et al. [6], used the same GLM model to simulate the rainfall, that was later used to characterise meteorological drought severity in UK. Another study regarding the application of GLM, which is more closely related to investigate drought characterisation, is performed by Tosunoğlu and Onof [70]. In his study, GLM is to be used to obtain the synthetic rainfall magnitude and its duration, and in combination with SPI¹² tried to understand the characteristic of drought severity and duration in Northeastern Turkey. Those state-of-the-art of the application of GLM in climate modelling raise an interesting idea on applying GLM to gain more revealing information regarding drought. In this research, GLM will be used as the tool to evaluate the influence of explanatory variables with respect to the rate of days without discharge within California. The motivation of this analysis is to be able to find the relationship between duration of the drought as a function of either meteorological or hydrological variables; and to study the current state of hydrological drought duration in California using the historical datasets. The output will be log-linear equation model with the relevant variables to measure the risk ratio¹³ (See section section 4.2.1 for further explanation of risk ratio) of drought duration. This chapter will be elaborated into three sections. The first section - preliminary stages on constructing linear model - includes the datasets that are used in the model, both response and explanatory variables; statistical pre-analysis in the form of correlation coefficient; and the scenarios on which to execute the final result. In the following section, the result will be elaborated based on the planned scenarios and its interpretation. The author will also include an additional subsection explaining one example of how to interpret the result in a simple model. Lastly, this chapter finishes by disclosing main findings in the conclusion.

4.1. Preliminary Stages on Constructing Linear Model

This stage is a step prior to the construction of the linear model. This section will discuss about input data for the model; the scenarios on the execution of the model; and the statistical pre-analysis of the datasets.

4.1.1. Variable in Datasets: Response and Explanatory Variable

As some terminologies regarding the type of variables have been introduced in section 2.5, the feeding data to the model will be classified into two types of variable. The response variable is the number of days without discharge (or with zero discharge) in a month that represents hydrological drought duration, whereas the explanatory variables are the possible relevant hydrological or meteorological time series. There are 810 basins within California delineated by GAGES II. The discharge data were available from USGS. Meteorological time

¹²SPI or Standardised Precipitation Index (see appendix A, section A.1 for further explanation)

¹³Risk ratio and multiplicative effect will be used interchangeably.

series were extracted from CORDEX, while SPI were extracted from NADM (North American Drought Monitor). From the raw data, each explanatory variable was processed such that the values are able to embody the drought processes. Table 4.1 shows the type of data for the final explanatory variables. As the characteristic

Table 4.1: Description of explanatory variables as input for the model

No.	Explanatory Variables	Notation	Unit	Type of Variables
1.	Precipitation Flux	PR	mm/day	monthly average
2.	Air Temperature	TAS	°C	monthly maximum
3.	Atmosphere Water Vapour Content	VC	mm	monthly minimum
4.	Snow melt Flux	SNM	mm/day	monthly average
5.	SPI-01	SPI_{01}	N/A	monthly

of drought are varying throughout space and time (see section 2.1.2), the model should be able to cope with spatial and temporal variability. In order to accommodate the spatial variability, the explanatory variable should correspond to the response variable under the same drainage basin. Therefore, one data configuration has different representative values for each variables. In addition to that, the explanatory time series should have the same time frame with the response variable. By means that each data configuration exhibits the drought event which occurred in an arbitrary month. On contrary, drainage basins that do not have any number of days without discharge in a month within the time frame were discarded. Thus, applying all the aforementioned criterion generates 410 basins that fits the interest of GLM analysis varying from year 1950 to 2005. A clear picture of the time series variation of all (explanatory and response) variables are presented in figure 4.1.

This figure shows the data for Station Tahquitz C NR PALM California (11258000) where its main basin has area of 43.691 km^2 . It has the longest time series among the 410 basins considered (from the 1951 to 2005) and is located in the Riverside County, South-East California, under Colorado River Region. Under the same station, the response variable (number of days in a month) is summarized in table 4.2. There were 4653 days with zero discharge in 55 years.

Table 4.2: Statistical summary of the response variable in Station Fresno R BL Hidden Dam

Station	Mean [days]	Max [days]	Min [days]	Sum [days]
10258000	7.18	31	0	4653

4.1.2. Model Structures: Scenarios for the Different Variables of Interest

Several scenarios will also be made to define different variables of interests to gain more insights of the response variable under specific circumstances. There will be three classes of scenarios on the model using GLM. The basis of the scenario is classifying the datasets into different criterion.

First Scenario

The first class scenario is dividing the dataset into different hydrological regions in California. This scenario is aimed studying the impact of drought duration under different nuances of hydrological function. There are ten hydrological regions in California, and tabulated with the details in the following table:

These regions are categorising the catchment areas based on the prevailing hydrological function. GLM will be applied to fit the data based on the classification of the hydrological function. Figure 4.2 illustrates the division of hydrological regions in the state of California. It shows that California has a diverse hydrological regions, four regions that are coastal dominated, three regions that are river dominated, one region which is water bodies (lake) dominated, and the rest are mountain dominated, that share the region with neighbouring state, Nevada.

Second Scenario

The second scenario is grouping the dataset based on the size of the catchment area. The purpose of this scenario is to assess the influence of the size of the catchment towards the hydrological drought duration within California. There are five different classes of catchment area, and the details are tabulated in table 4.4. Similarly to the first scenario, GLM will fit the data based on the considered classes. The drainage basins are mostly in the size between 100 and 1.000 km^2 with almost 200 basins.

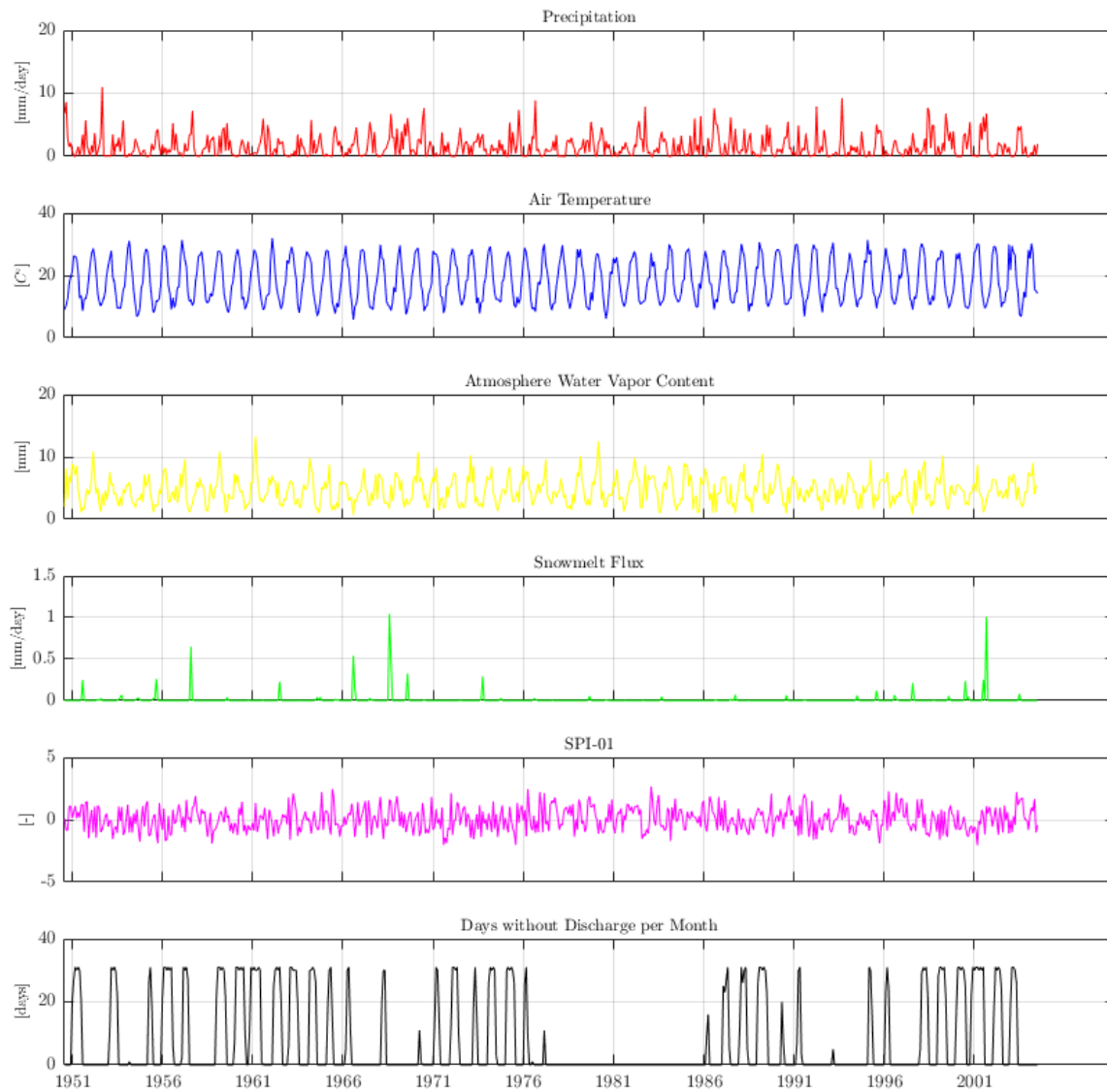


Figure 4.1: Example of time series of explanatory variables in station 10258000

Table 4.3: Description of hydrological regions in California

Region	Area (km^2)	Number of Contributed Stations
North Coast	50.292,56	12
San Francisco Bay	11.684,70	38
Central Coast	29.254,51	78
South Coast	28.340,49	116
Sacramento River	70.471,15	54
San Joaquin River	39.593,07	44
Tulare Lake	43.614,94	25
North Lahontan	15.852,36	9
South Lahontan	69.108,32	17
Colorado River	51.441,01	17

Third Scenario

The third scenario aims to study the development of hydrological drought duration influenced by precipitation under different model in California. The significance of the influence will be expressed in risk ratio

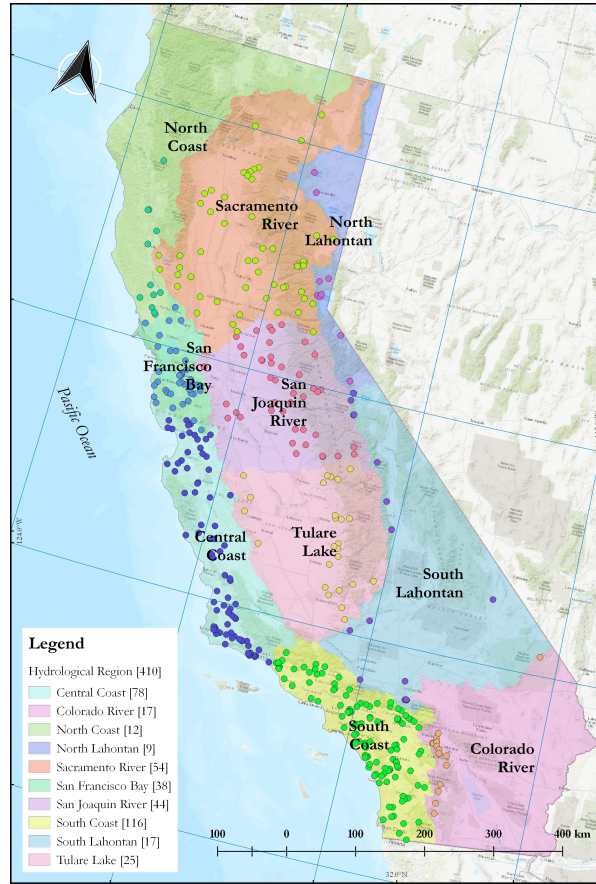


Figure 4.2: Ten hydrological regions in California that delineated based on the main hydrological functions

Table 4.4: Description of class catchment area in California

Class Catchment Area	Number of Contributed Stations
$\leq 10 \text{ km}^2$	9
$10 - 100 \text{ km}^2$	132
$100 - 1,000 \text{ km}^2$	199
$1,000 - 10,000 \text{ km}^2$	67
$> 10,000 \text{ km}^2$	3

maps. The development is translated into four different subsets with different periods of time, (i) 1950-1965; (ii) 1950-1970; (iii) 1950-1990; and (iv) 1950-2005. Figure 4.3 portrays the contributed stations in different time periods. The reason beyond the division is that drought events are perceived as a sequential chain of events. By updating the datasets into longer period, rather dividing the time series into different time set, we are not treating the historical data as an independent data set. In other words, the model assumes the impact of prevalent drought event in a way affected by the previous event. In this scenario, the focus variable is precipitation as this variable is the main input source in the hydrological cycle. Furthermore, there will be different types of models of GLM, that will be build upon different explanatory variables. The first model focuses on the impact of precipitation to the hydrological drought duration given we only know precipitation. In contrast in the second model, GLM will be updated by feeding more explanatory variables. Alternatively saying that, we will be still focusing on the risk ratio of days with zero discharge induced by precipitation given we know more information. The second model involves the variables from table 4.1. Adopting the above conditions, the data will be treated in the following ways (note that these procedures only hold for scenario 3):

1. The fitting dataset with GLM will be done per station.
2. Each stations will be divided into four subsets based on the periods of interest. Note that not every

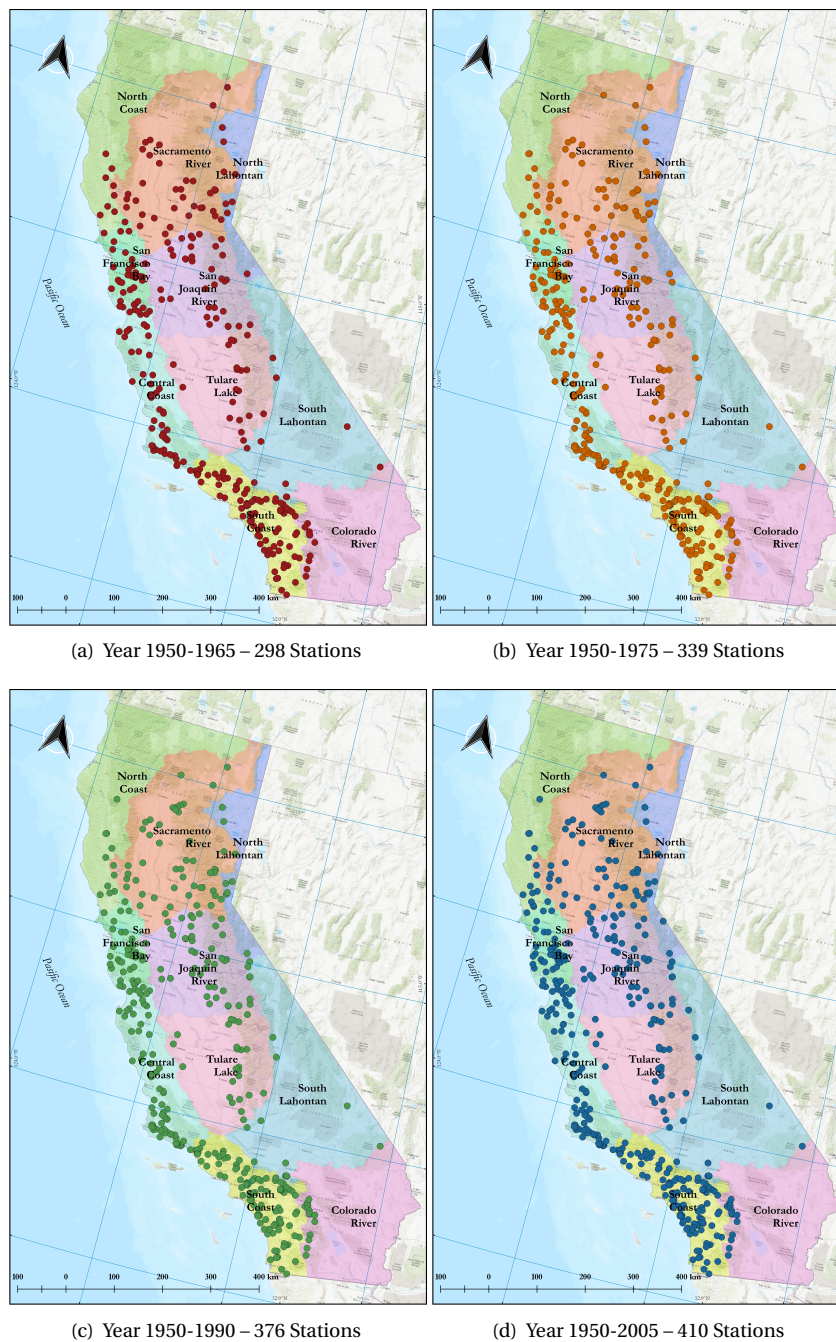


Figure 4.3: Station distribution on time period sub-datasets

station is contributing in every period of interest.

3. These sub-datasets will be analysed under two different models, yielding 8 model results of risk ratio distribution in California.
4. The risk ratio of each stations will be extended throughout California using spatial interpolation: interpolated distance weight.
5. The result will be displayed in a map with different time sub-datasets and model scenarios

It is important to bear in mind that projecting the data using spatial interpolation might have an uncertainty. In this study, the maps will be treated in a preliminary manner in order to give a broad idea in how the hydro-

logical drought duration distributed spatially. Of course in the further study, such map should be able to give idea of the trend of the spatial distribution of hydrological drought duration. However, this part would not be treated in this thesis.

To give the reader an idea the difference between scenarios, the following table tabulates the overview of the scenarios.

Table 4.5: Table summary of different scenarios

Scenario	Objectives	Class Data Filter	Fitting model per
Scenario 1	To study the impact of hydrological drought duration in different hydrological regions	Hydrological region	Regions
Scenario 2	To study the impact of hydrological drought duration under different catchment size	Catchment size area	Class Area
Scenario 3	To study the development of hydrological drought duration under different models	Four time periods	Station

4.1.3. Preliminary Steps: Calculating the Correlation Coefficient of the Datasets

Before applying the model into our dataset, a preliminary dependence analysis is conducted. This step will give an insight on how the dataset correspond between each other, precisely the response variable and explanatory variables. The relationship between these variables should also be comparable to its behaviour in nature, and its significance could be quantified using Spearman's Correlation Coefficient (see equation 2.7). Table 4.6 presents the correlation coefficient between all time series dataset. The bold number shows the correlation coefficient between the response variable (DWQ) and explanatory variables. The magnitude of

Table 4.6: Correlation coefficient of each variables

ρ_s	DWQ	PR	TAS	VC	SNM	SPI01
DWQ	1	-0.2166	0.2520	0.1127	-0.2139	-0.0343
PR		1	-0.7445	-0.4276	0.5361	-0.0372
TAS			1	0.6013	-0.6074	0.0224
CV				1	-0.5466	0.0238
SNM					1	-0.0084
SPI01						1

the correlation reveals the strength of dependence between the explanatory variables and response variable, statistically. The attributes of the magnitude tells us about the behaviour of the correlation; positive means two variables are moving in the same direction, meaning low/high quantile of one variable corresponds to the low/high quantile of the other variable; whereas negative means two variable are correlated in the opposite way. For instance, the spearman's correlation coefficient between DWQ (day without discharge) and PR (precipitation) is -0.2166 . The negative sign expresses that the variables changes in different directions. Otherwise stated that, having less precipitation leads to more days without discharge and vice versa. Having this statistical information will give another context, which is related to the physical mechanism, during the evaluation of the result in the model.

4.2. Results and Discussion

In this section, an example of using a simple mode is presented to ease the explanation of the concept of risk ratio used in drought duration analysis. This section also describes the final result of GLM under different scenarios. Followed by the the discussion of the result using risk ratio, in the following subsection.

4.2.1. Fitting Example: Analysing Precipitation to the Response Variable in California

In this subsection, an example on further interpretation of covariates on response variable will be elaborated. The following example makes use of merely precipitation to explain the influence of days with zero discharge per month, precisely in California. In hydrological cycle, precipitation is the main input source in the water balance. Lack of precipitation in a certain extent of time might be seen as an indirect caused of hydrological

drought (see subsection 2.1.2 regarding drought propagation). Therefore, interconnection between precipitation and drought duration should be stated in a negative feedback, meaning that having less precipitation will lead to have more days without discharge, and vice versa. This relationship has been quantified in section 4.1.3 by calculating the correlation coefficient. The concept between these variables would be the ground basis to evaluate whether the result of the regression model is comparable to its behaviour in nature. The GLM equation as a function of only precipitation in California, is formulated as:

$$\ln(E(X_i)) = \beta_0 + \beta_1 \cdot PR + \epsilon \quad (4.1)$$

As mentioned in section 2.5, the model is selected to follow a Poisson distribution, since we are dealing with count data in the response variable ($\ln(E(X_i))$), whilst the link function of the predictors ($\beta_1 \cdot PR$) is the logarithmic link function (see equation 2.13). Figure 4.4 demonstrates the result of GLM by fitting the precipitation data into the equation.

```
Generalized linear regression model:
log(Duration_Discharge) ~ 1 + Precipitation
Distribution = Poisson
```

Estimated Coefficients:		
	Estimate	SE
(Intercept)	2.2035	0.00113
Precipitation	-0.13631	0.00041977

Figure 4.4: Result analysis for GLM with precipitation as the explanatory variable

Referring to figure 4.4, the result shows the estimated coefficient for precipitation and the standard error (SE). Using the estimated coefficient, we are able to say something towards the increase\reduction of the risk of response variable using risk ratio. RISK RATIO - Risk ratio is a intuitive way to define a ratio between exposed incidences and unexposed incidences in a particular time frame. Risk ratio is a well known way to correlate a causal relationship between two variables in epidemiology studies. Risk ratio for log link function, mathematically, is formulated as follows:

$$RR = \frac{E(Y_i|Present)}{E(Y_i|Absent)} = \frac{e^{\beta_j(x_j+1)}}{e^{\beta_j(x_j)}} = e^{\beta_j} \quad (4.2)$$

This equation holds when the explanatory variable with factor x_j is present. Since the link function is logarithm, hence the inverse of the log of the coefficient is the actual value of the risk ratio. Given the above example (figure 4.4), the exponential of the estimated coefficient is the risk ratio, which happens to be 0.8726. It enables to explain the influence of precipitation flux towards the number of days without discharge. If the above regression is accepted then there will be a multiplicative effect of 0.8726 on the rate of day without discharge for 1 unit (mm/day) increase in mean monthly of precipitation flux. It is also possible to express the multiplicative effect in the percentage risk reduction (for $RR < 1$) or increase (for $RR > 1$). This could be done using the following formula:

$$\begin{aligned} \text{For risk reduction (RR < 1)} & : (1 - RR) \cdot 100\% \\ \text{For risk increase (RR > 1)} & : (RR - 1) \cdot 100\% \end{aligned}$$

Therefore, using the same example, the risk ratio could also be stated in this way. If the mean monthly precipitation is increasing 1 mm/day in California, there will be $((1 - 0.8726) \cdot 100\%)$ 12.74 % reduction in the number of days without discharge. Notice that these two variables is in-line with the rank correlation found in section 4.1.3.

In many cases, including this study, the variable of interest is expressed in a function of several covariates. If this is the case, risk ratio would be only explain one covariate with respect to the variable of interest, given the other covariates remain constant. Let $X_1, X_2, \dots, X_j$ be the predictors in the equation with β as its coefficient. This condition is formulated mathematically as follows:

$$RR = \frac{E(Y_i|Present)}{E(Y_i|Absent)} = \frac{e^{\beta_0 + \beta_1(X_1+1) + \beta_2(X_2) + \dots + \beta_j(X_j) + \epsilon}}{e^{\beta_0 + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_j(X_j) + \epsilon}} = e^{\beta_1}$$

Therefore, in the case of having more than one predictor, explaining a particular covariate using risk ratio would be done once we control for other covariates. Later in the application, it also shown that risk ratio is one of many methodologies to evaluate the sensitivity of particular covariate with respect to the response variable (see section 4.2.3).

4.2.2. Linear Model Equation: Applying GLM in California

This section reveals the mathematical function and the result of the regression model that fits into the function. The selection of the variables to estimate GLM's coefficients were done in two main stages. (i) Firstly, the fitting processes were done using stepwise GLM regression. In this stage, each possible variable was included in the model, one by one. The significance of the explanatory variables in the regression were evaluated. These steps were repeated by adding or removing the variable in the regression model until the regression reached the optimum result of covariates combination. This repetition is the so-called stepwise GLM regression. The significance of the explanatory variables will be checked statistically using p-value retrieved from the result of MATLAB. The p-value less than 0.05, shows that the covariate is statistically relevant to the variate of interest. As shown in table 4.7, the p-value for each covariate is less than 0.05, meaning that it is statistically significant to the response variable. (ii) Secondly, using the output of the first stage, the regression were again

Table 4.7: The GLM regression result retrieved from MATLAB using the whole datasets

	<i>(Intercept)</i>	<i>PR</i>	<i>TAS</i>	<i>VC</i>	<i>SNM</i>	<i>SPI₀₁</i>
Estimate	1.1355	-0.0447	0.0538	-0.0305	-0.0838	-0.1019
SE	0.0043	0.0005	0.0002	0.0005	0.0012	0.0010
p-value	0	0	0	0	0	0
162851 observations						
p-value system :	0					

checked by comparing the behaviour of each explanatory variables using the Spearman's Correlation Coefficient with the result of estimated coefficient of the model (look at the example in subsection 4.2.1). Referring to the table 4.7, the estimated coefficient shows the same attribute (or associated to the same behaviour) to the rank correlation coefficient in table 4.6. This phase helps to evaluate whether or not the outcome of the model resembles valid or logical physical representation of drought mechanism.

Table 4.7 shows the information of GLM regression using five covariates taken from MATLAB result. This regression result was performed using the whole datasets in California with 162.851 number of observations. The table consists of the estimated coefficient, SE or standard error of the estimated coefficient, and the p-value, all for each covariate. An extra information regarding the p-value of the system was also summarized in the table. The p-value of the system is showing less than 0.05, meaning that this GLM differs statistically from a constant model¹⁴. Alternatively saying that if the regression is accepted, the model could be used to forecast the response variable, given we know the covariates.

Having all the above information, the response variable as a function of the combination of five predictors using generalised linear model is mathematically expressed in equation 4.2. This linear equation was fitted using the whole dataset in California.

$$\ln(E(X_i)) = \beta_0 + \beta_1 \cdot PR + \beta_2 \cdot TAS + \beta_3 \cdot VC + \beta_4 \cdot SNM + \beta_5 \cdot SPI_{01} + \epsilon \quad (4.3)$$

where X_i is the count days without discharge in a month under particular basin- i which acts as the response variable. β exemplify the coefficient of the corresponding variable in the model. The details of all variables in the equation are mentioned in table 4.1.

If we accept the above regression, equation 4.2 is the linear model representation of the expected number of days with zero discharge per month. The duration of this hydrological droughts is explained better using these climate predictors in equation 4.2 in California. Thereafter, this equation will also be further used for all scenarios. The following table (table 4.8) shows the result of the estimated coefficient and its standard error using equation 4.2 for the whole datasets, class of catchment area, and for each hydrological regions.

¹⁴A further information regarding constant model could be referred to [1]

In general, the estimation of coefficients using equation 4.2 results in moderately consistent result. The estimated coefficient for the whole datasets appears to have a really small deviation, signified by the standard error. Furthermore, estimated coefficient of precipitation (PR) in several regions, for instance Colorado River, San Joaquin River, North and South Lahontan, are indicated in positive values. On the contrary, the other regions shows in negative values, which is to be the expected attribute. Another alteration in attributes also took place in predicting water vapour content, particularly in Region of Colorado River, San Joaquin River and South Lahontan. Those mentioned inconsistencies might happen due to many factors, such as the impact of basin characteristics or diverse climate virtue in California. Further analysis on the physical interpretation using risk ratio will be presented in the following subsection.

Table 4.8: Estimated value and standard error (S.E.) of the coefficient for the whole dataset, scenario 1 and scenario 2

Scenarios	β_0		PR		TAS		VC		SNM		SPI-01	
	Value	S.E.	Value	S.E.	Value	S.E.	Value	S.E.	Value	S.E.	Value	S.E.
Whole Dataset												
–	1.1355	0.0043	-0.0447	0.0005	0.0538	0.0002	-0.0305	0.0005	-0.0838	0.0012	-0.1019	0.0010
Class of catchment area												
$\leq 10 \text{ km}^2$	2.5017	0.0244	-0.0667	0.0035	0.0425	0.0012	-0.1274	0.0033	-0.6705	0.0397	-0.1127	0.0066
10 - 100 km^2	0.9454	0.0077	-0.0480	0.0009	0.0689	0.0004	-0.0439	0.0008	0.0244	0.0013	-0.1156	0.0019
100 - 1,000 km^2	1.0977	0.0063	-0.0439	0.0007	0.0570	0.0003	-0.0437	0.0007	-0.0746	0.0018	-0.0895	0.0015
1,000 - 10,000 km^2	1.3216	0.0104	-0.0235	0.0011	0.0394	0.0005	-0.0029	0.0012	-0.5840	0.0063	-0.1050	0.0024
$> 10,000 \text{ km}^2$	1.2823	0.0644	-0.0060	0.0077	0.0328	0.0029	-0.0829	0.0082	-2.0107	0.8726	-0.1639	0.0130
Hydrological region												
Central Coast	1.2796	0.0101	-0.0088	0.0011	0.0659	0.0004	-0.0697	0.0009	-0.3169	0.0716	-0.0976	0.0021
North Coast	1.4254	0.0154	-0.0141	0.0026	0.0558	0.0008	-0.0631	0.0026	0.3420	0.0305	-0.1189	0.0043
South Coast	0.0297	0.0842	-0.1465	0.0072	0.0456	0.0037	-0.0412	0.0074	-0.9261	0.2044	-0.0472	0.0156
San Francisco Bay	0.4226	0.0410	0.0204	0.0030	0.0232	0.0029	-0.0080	0.0094	-0.0513	0.0040	0.0163	0.0126
Colorado River	-0.3845	0.0196	-0.0026	0.0014	0.0416	0.0010	0.0931	0.0025	0.0067	0.0023	-0.1267	0.0050
Sacramento River	0.3368	0.0229	-0.0524	0.0020	0.0829	0.0011	-0.0427	0.0017	-2.9732	0.3361	-0.1176	0.0043
San Joaquin River	-0.1603	0.0170	0.0311	0.0014	0.0550	0.0007	0.1642	0.0016	-0.4117	0.0054	-0.0646	0.0037
Tulare Lake	1.6545	0.0068	-0.0291	0.0011	0.0378	0.0003	-0.0244	0.0009	0.0519	0.0192	-0.1133	0.0016
North Lahontan	1.2430	0.0165	0.0253	0.0013	0.0432	0.0009	-0.0330	0.0040	0.0236	0.0016	-0.0599	0.0056
South Lahontan	0.1783	0.0191	0.0007	0.0021	0.0490	0.0008	0.1415	0.0021	-0.0417	0.0038	-0.0356	0.0047

4.2.3. Hazard Estimation: Covariate Contribution to Drought Hazard Duration in California

In this section, the contribution of the estimated coefficients translated into risk ratio will be analysed based on the mentioned class scenarios. A short explanation using simpler model has been introduced in section 4.2.1 on how to translate the coefficient into risk ratio. Risk ratio resembles the multiplicative effect on a certain explanatory variable towards the response variable given that other explanatory variables remain the same. It means that only univariate interpretation could be done. Nonetheless, in nature, drought is a complex mechanism that involves many aspects rather only be explained by one particular variable. Thence, to reduce the bias of the interpretation, the precondition of the variable of interest should be clearly defined. An easy method to do so is by dissecting the discussion into different categories of the regions (for scenario 1). This will enable to identify the pattern of risk ratio easier based on the main function of the hydrology. Later in the discussion, the risk ratio will be presented by gradually including other parameters and focusing on the sensitivity of Precipitation Flux (*PR*). The class scenario 2 will also be displayed in the same way.

Scenario 1 - Risk Ratio based on Hydrological Region

RISK RATIO FOR COASTAL-DOMINATED REGION IN CALIFORNIA – Table 4.9 shows the result of risk ratio for coastal-dominated region in California. If the regression of the models are not rejected, several conclusions could be derived from the table. (i) In the coastal dominated areas in California, if the mean monthly precipitation is increasing for 1 mm/day, there will be about 0.88 - 13.63 % of risk reduction in days without discharge in a month. The percentage is yielded from the minimum and maximum value of the risk ratio of four regions with all inclusion of parameters. (ii) In addition to that, having an increase of 1 unit degree Celcius of air temperature in California's delta areas, consequently induces to an escalation of 3.85 - 8.64 % longer of days without discharge in a month. (iii) Region with less human activity (see figure 2.8), e.g. North Coast, has lower risk ratio of hydrological drought duration induced by Precipitation. In other words, it has higher reduction in percent risk. A worth mentioning consideration towards this argument is to include variable(s) that control the discharge within the catchment, for instance reservoir or hydraulic structures.

Table 4.9: Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for coastal-dominated region

Hydrological region and parameters included in the model		Risk Ratio				
		PR	TAS	VC	SNM	SPI01
Central Coast	PR	0.9119				
	PR + TAS	0.9995	1.0560			
	PR + TAS + VC	0.9930	1.0686	0.9310		
	PR + TAS + VC + SNM	0.9935	1.0685	0.9305	0.6587	
	PR + TAS + VC + SNM +SPI01	0.9912	1.0681	0.9327	0.7284	0.9070
North Coast	PR	0.8157				
	PR + TAS	0.8583	1.0389			
	PR + TAS + VC	0.8550	1.0465	0.9638		
	PR + TAS + VC + SNM	0.8645	1.0464	0.9611	0.3925	
	PR + TAS + VC + SNM +SPI01	0.8637	1.0467	0.9597	0.3961	0.9539
South Coast	PR	0.9020				
	PR + TAS	0.9701	1.0328			
	PR + TAS + VC	0.9735	1.0388	0.9756		
	PR + TAS + VC + SNM	0.9734	1.0388	0.9756	1.0143	
	PR + TAS + VC + SNM +SPI01	0.9714	1.0385	0.9759	1.0532	0.8929
San Francisco Bay	PR	0.8678				
	PR + TAS	0.9536	1.0740			
	PR + TAS + VC	0.9498	1.0859	0.9586		
	PR + TAS + VC + SNM	0.9515	1.0859	0.9569	0.0454	
	PR + TAS + VC + SNM +SPI01	0.9490	1.0864	0.9582	0.0511	0.8891

RISK RATIO FOR RIVER-DOMINATED REGION IN CALIFORNIA – For river-dominated regions, the result of risk ratio is presented in table 4.10. Generally, in comparison with coastal areas, river-dominated has lower risk ratio value due to precipitation flux and air temperature. Unsatisfactory model result occurred in San Joaquin

River and Tulare Lake, as one introduced snow melt flux into the equation. This inconsistency may occur because the covariates combination are not able to represent drought duration. Since the primary land use is occupied in agricultural production as human activities might play a big role causing drought. Therefore, a further investigation on drought duration induced by agricultural activity would be an alternative solution to gain more revealing information. Nevertheless, conclusions that could be inherited from the table, if one accept the regression of Colorado River and Sacramento River is (i) the rate of risk ratio of days without discharge ranges from 0.26 to 1.4 % influenced by precipitation. (ii) Furthermore, in average, the growth of 1 unit degree Celcius of air temperature results in 4.25 - 5.74 % increased risk with respect to the days without discharge in a month.

Table 4.10: Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for river-dominated region

Hydrological region and parameters included in the model		Risk Ratio				
		PR	TAS	VC	SNM	SPI01
Colorado River	PR	0.9000				
	PR + TAS	0.9816	1.0435			
	PR + TAS + VC	0.9932	1.0581	0.9360		
	PR + TAS + VC + SNM	0.9890	1.0583	0.9375	1.3621	
	PR + TAS + VC + SNM +SPI01	0.9860	1.0574	0.9389	1.4078	0.8879
Sacramento River	PR	0.9399				
	PR + TAS	1.0017	1.0617			
	PR + TAS + VC	0.9994	1.0411	1.1012		
	PR + TAS + VC + SNM	0.9980	1.0413	1.1015	1.0055	
	PR + TAS + VC + SNM +SPI01	0.9974	1.0425	1.0976	1.0067	0.8810
San Joaquin River	PR	0.8526				
	PR + TAS	0.9690	1.0912			
	PR + TAS + VC	0.9730	1.0562	1.2204		
	PR + TAS + VC + SNM	1.0318	1.0560	1.1791	0.6627	
	PR + TAS + VC + SNM +SPI01	1.0316	1.0565	1.1785	0.6625	0.9375
Tulare Lake	PR	0.8813				
	PR + TAS	0.9893	1.0738			
	PR + TAS + VC	0.9887	1.0498	1.1554		
	PR + TAS + VC + SNM	1.0007	1.0500	1.1516	0.9592	
	PR + TAS + VC + SNM +SPI01	1.0007	1.0502	1.1519	0.9592	0.9650

RISK RATIO FOR MOUNTAIN REGION IN CALIFORNIA – Unlike other region California, North and South Lahontan, are the regions that share most of its river basin with the neighbouring state, Nevada. This fact makes it difficult to conclude the physical interpretation of the result to its hydrological region. For instance, the precipitation risk ratio which is expected to have negative feedback towards hydrological drought, happened to occur in an opposite way. This poor estimation happened when other explanatory variables are included in the model, refer to table 4.11. This anomaly might be happening due to the fact that the point measurements in the North and South Lahontan are not able to well represent the entire region. Thus, more information about explanatory variables from neighbouring state Nevada might be relevant to represent the entire region.

All in all, if one accept the regression result of model in which all parameter are included (using equation 4.3), the reduced and increased ¹⁵ risk of hydrological drought duration due to precipitation and air temperature are depicted in figure 4.5, respectively. The blue hatch indicates the coastal-dominated regions in California, where as the green hatch indicates the river-dominated regions in California. In these maps, we excluded regions such as San Joaquin, Tulare Lake, North and South Lahontan, because the regression results show a risk ratio that is different from the expected (see previous explanation for each region). Having result of risk ratio per region gives another perspective on how we could proceed on the impact of hydrological drought duration due to certain parameter. In a preliminary manner, the changing behaviour of such parameter could be indirectly linked to the loss of economic. A worth mentioning example is utilising the risk ratio for agricultural strategy production. This computation could be done once the relationship between

¹⁵the concept of reduced/increased risk has been introduced in subsection 4.2.1

Table 4.11: Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for mountainous-dominated region

Hydrological region and parameters included in the model		Risk Ratio				
		PR	TAS	VC	SNM	SPI01
North Lahontan	PR	0.9778				
	PR + TAS	1.0014	1.0322			
	PR + TAS + VC	1.0018	1.0344	0.9877		
	PR + TAS + VC + SNM	1.0205	1.0236	0.9919	0.9501	
	PR + TAS + VC + SNM +SPI01	1.0206	1.0235	0.9921	0.9500	1.0164
South Lahontan	PR	0.9906				
	PR + TAS	1.0330	1.0364			
	PR + TAS + VC	1.0336	1.0391	0.9795		
	PR + TAS + VC + SNM	1.0259	1.0444	0.9670	1.0243	
	PR + TAS + VC + SNM +SPI01	1.0256	1.0442	0.9675	1.0239	0.9419

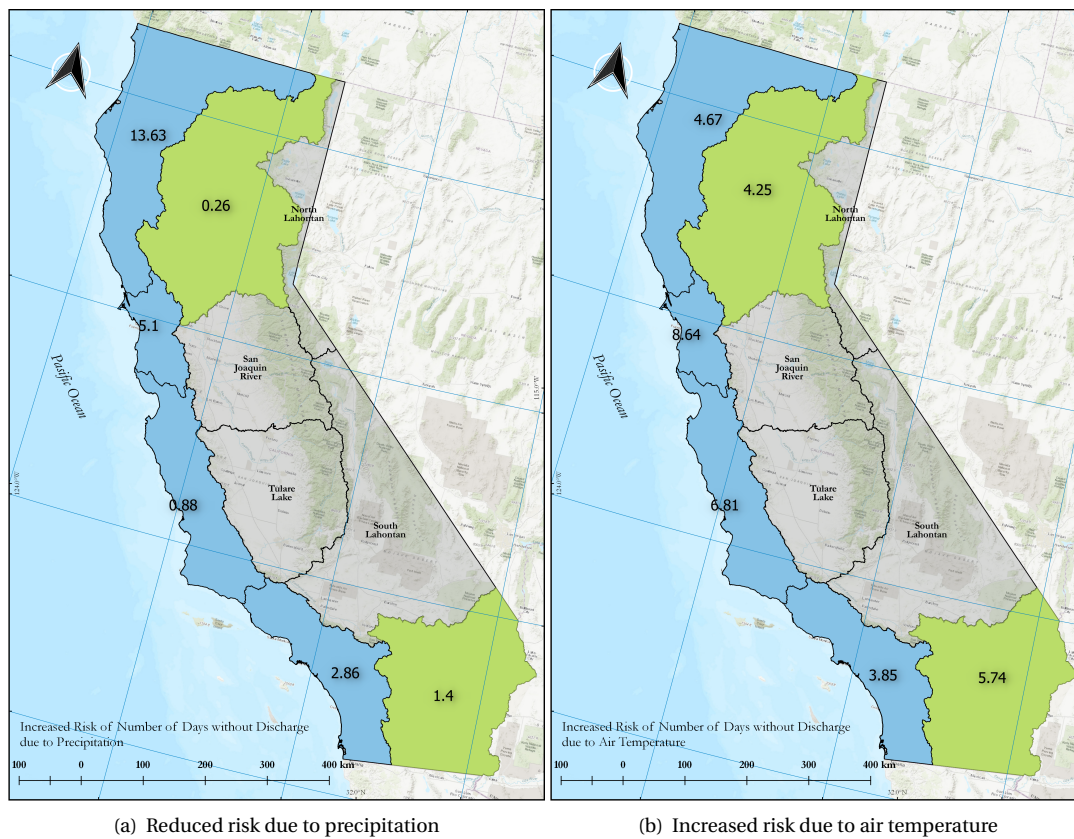


Figure 4.5: Maps of different risk of hydrological drought duration in the hydrological regions in California using model equation 4.3. The blue hatch indicates the coastal regions, while the green hatch indicates the river regions.

the response of the crop and hydrological drought duration is yielded. Therefore, the multiplication of risk ratio, crop production income, and the response factor of the crop could become the rough estimation of the production loss due considered parameter. For instance, in this case, if the mean monthly air temperature is increasing, hence triggering the risk of having longer duration. Using this rough calculation, the loss of agricultural income could be computed. One of the benefits of estimating the loss is that we could come up with a preliminary consequences in hydrological drought. Thus, in the later stage, the combination of the consequences and the hazard will be the quantification of risk of hydrological drought.

Scenario 2 - Risk Ratio based on Catchment Area

RISK RATIO FOR CATCHMENT SIZE AREA – In this section, the risk ratio of class scenario 2 (catchment size area) is presented in table 4.12. The result shows that in general, there is a increasing trend of multiplicative effect to the size of catchment. Meaning the bigger the size of the catchment, the reduced risk is expected to be lower. For illustration, if the precipitation flux is present in smaller catchment ($10\text{-}100\text{ km}^2$), there will be 13.68 % reduced risk of having days without discharge in a month. On contrary for bigger catchment such as class greater than $10,000\text{ km}^2$ within the same condition, there will be only 4.72 % reduced risk of having days without precipitation.

Table 4.12: Risk ratio of explanatory variables towards hydrological drought duration in different GLM model for class catchment area

Catchment area and parameters included in the model		Risk Ratio				
		PR	TAS	VC	SNM	SPI01
$\leq 10\text{ km}^2$	PR	0.8724				
	PR + TAS	0.9050	1.0228			
	PR + TAS + VC	0.9077	1.0457	0.8836		
	PR + TAS + VC + SNM	0.9372	1.0435	0.8791	0.5051	
	PR + TAS + VC + SNM +SPI01	0.9355	1.0434	0.8804	0.5114	0.8934
$10 - 100\text{ km}^2$	PR	0.8632				
	PR + TAS	0.9605	1.0617			
	PR + TAS + VC	0.9614	1.0715	0.9558		
	PR + TAS + VC + SNM	0.9553	1.0718	0.9556	1.0236	
	PR + TAS + VC + SNM +SPI01	0.9532	1.0713	0.9571	1.0247	0.8908
$100 - 1,000\text{ km}^2$	PR	0.8756				
	PR + TAS	0.9475	1.0507			
	PR + TAS + VC	0.9487	1.0593	0.9586		
	PR + TAS + VC + SNM	0.9584	1.0587	0.9569	0.9275	
	PR + TAS + VC + SNM +SPI01	0.9570	1.0586	0.9573	0.9281	0.9144
$1,000 - 10,000\text{ km}^2$	PR	0.8745				
	PR + TAS	0.9348	1.0434			
	PR + TAS + VC	0.9340	1.0381	1.0271		
	PR + TAS + VC + SNM	0.9781	1.0402	0.9968	0.5564	
	PR + TAS + VC + SNM +SPI01	0.9768	1.0402	0.9971	0.5577	0.9003
$> 10,000\text{ km}^2$	PR	0.9528				
	PR + TAS	0.9857	1.0162			
	PR + TAS + VC	0.9957	1.0339	0.9208		
	PR + TAS + VC + SNM	0.9987	1.0339	0.9176	0.0600	
	PR + TAS + VC + SNM +SPI01	0.9941	1.0333	0.9204	0.1339	0.8488

Another conclusion which could be derived from table 4.12 is that the more parameters are included in the model, the risk ratio is getting higher. Focusing on only one class of catchment area, class $100\text{-}1,000\text{ km}^2$, the multiplicative effect of precipitation tends to increase when another variables are included in the model, especially when precipitation is the only parameter. This makes sense, as we introduced more parameters, we are giving more significance to other parameters to control the response variable.

Scenario 3 - Development of Hydrological Drought Duration

Scenario 3 is specifically aimed to study the development rate of risk ratio of hydrological drought duration in California under different models. The explanation of the result will be divided into two parts. The first part will focus on the distribution of the data from time perspective, and the second part will elaborate on the spatial projection of the result. As mentioned in subsection 4.1.2, regarding scenario 3, the focus of the risk ratio will be only induced by precipitation. There are two different equations of GLM to be used under four subsets of time periods (see 4.1.2 - Scenario 3 for detail explanation). The model equations to be used in this scenario are equation 4.1 and equation 4.3, as will be assigned as model 1 and model 2, respectively.

RISK RATIO HISTOGRAM – In this part, the data will be shown in histogram in order to analyse the general

trend of the distribution. Figure 4.6 and figure 4.7 depict the histograms of risk ratio (induced by precipitation) for model 1 and model, respectively. Each histogram in the figure shows different results for four time periods. The x-axis shows the risk ratio data in California, whereas the y-axis is the frequency of its risk ratio. The red line shows the fitted distribution using kernel density – a non-parametric probability density function. Meanwhile, the green line shows the mean (or average) value of the risk ratio distribution. In a broader sense, the risk ratio shows a slight increased of frequency. Meaning that by updating the time series, the drought events with longer duration in a month has happened more frequent in California. It is indicated by the bulkiness of the kernel-PDF from the figure is slowly getting fatter (see red graph). However there are some differences when it comes to compare between two models. (i) The increased intensity of the risk ratio from time to time in model 1 is rather slower compared to model 2. It means that model 1 foresees more longer hydrological duration than model 2 (in the upper tail). (ii) Also, if we focus on the lower tail, which is shorter duration of hydrological drought, there is a modest increment in model 2. Whilst in model 1, the PDF shows almost constant trend over time.

In order to see the overall trend of the result, figure 4.8 depicts the mean value of risk ratio for each time

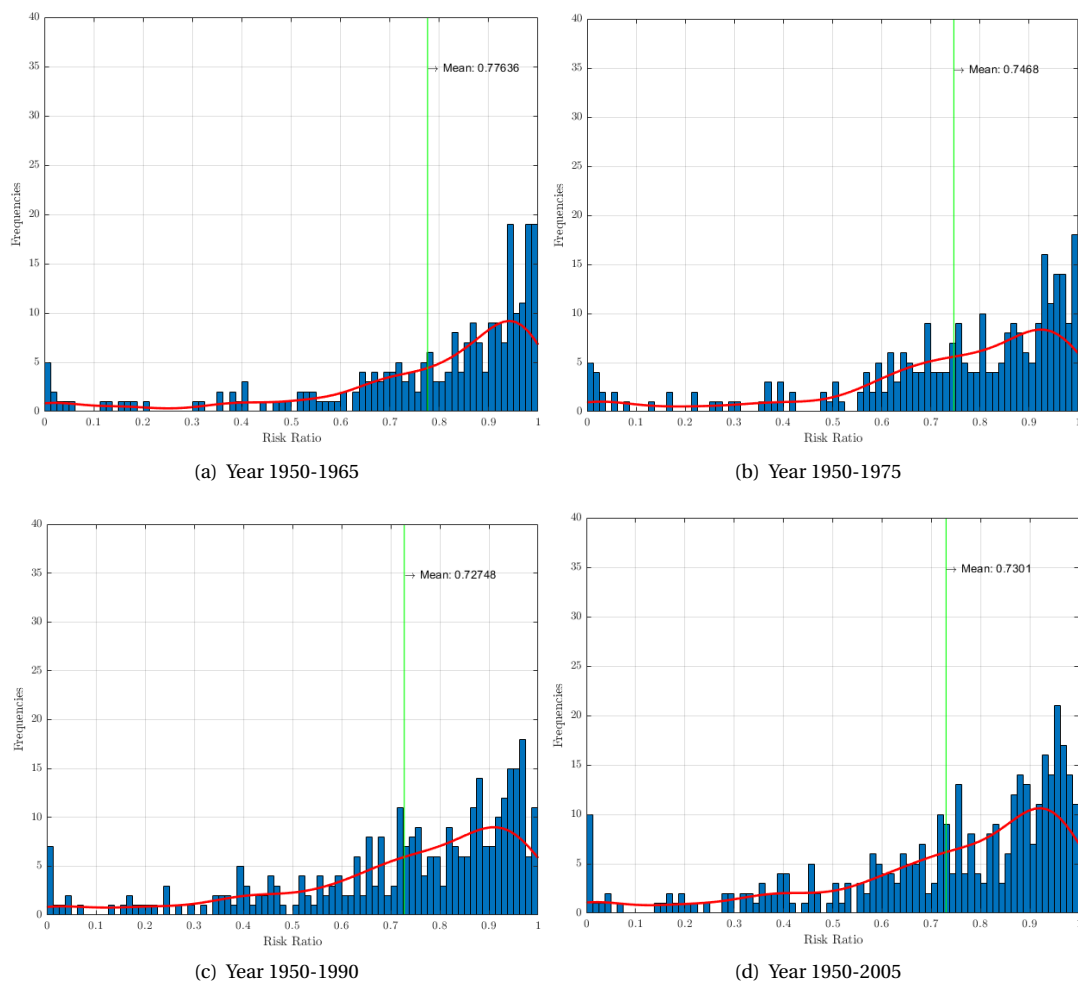


Figure 4.6: Histogram of risk ratio ($0 < RR < 1$) of days of without Discharge induced by precipitation model 1. The red graph shows the kernel PDF, the greenline indicates the mean value.

periods for both model. Although the risk ratio in the bigger tail happens more frequent, however the overall trend shows a decline of the average multiplicative effect over time period. Meaning that from 1950 until 2005, for any 1 mm/day increase of average monthly precipitation, it appears to reduce the risk ratio of days without discharge per month by updating the time series for both model. Alternatively stated that the number of days without discharge is decreasing. This will lead to have shorter duration of hydrological drought.

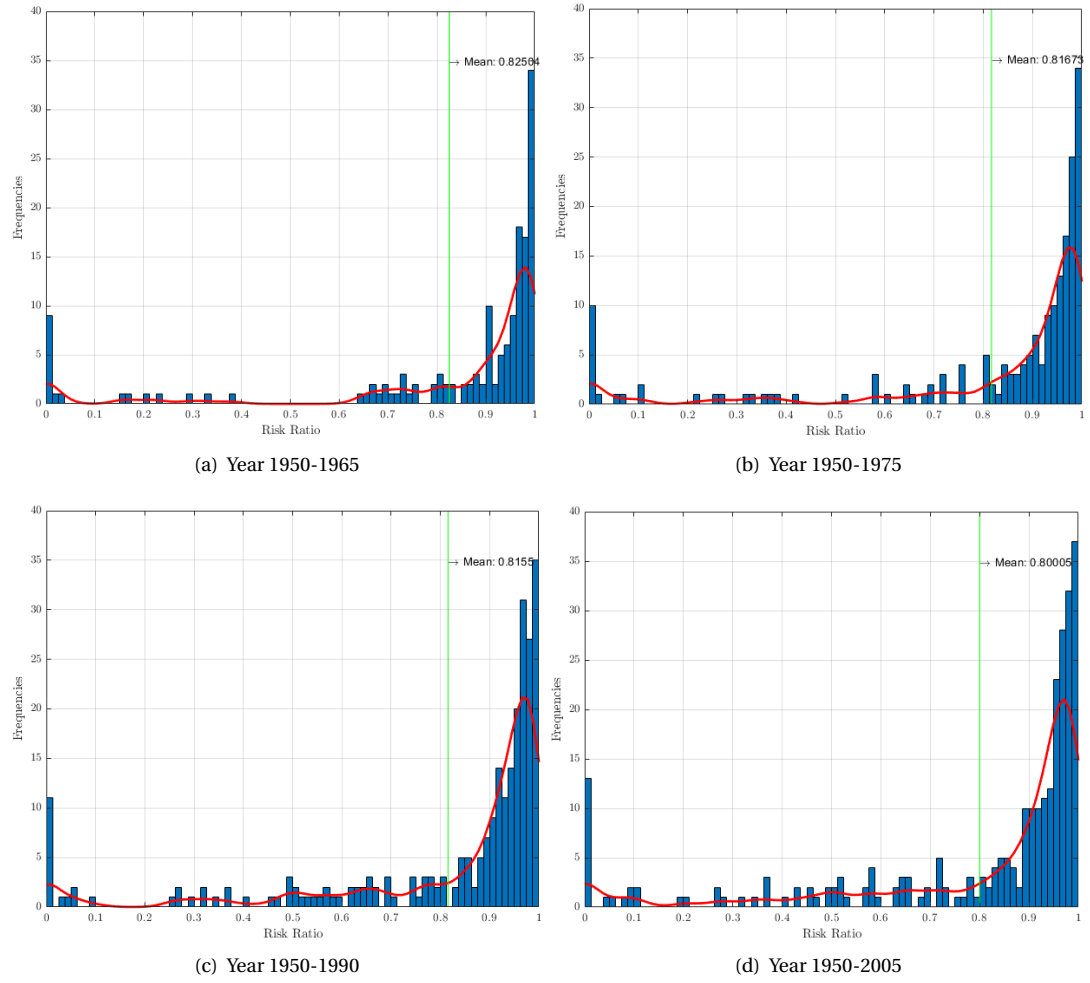


Figure 4.7: Histogram of risk ratio ($0 < RR < 1$) of days of without Discharge induced by precipitation model 2. The red graph shows the kernel PDF, the greenline indicates the mean value.

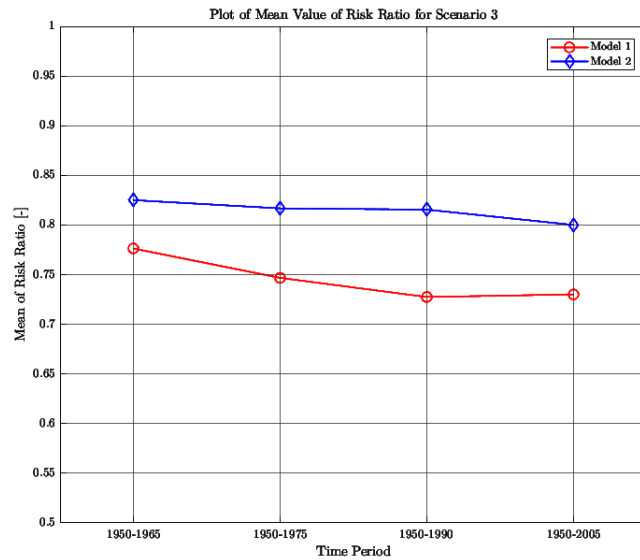


Figure 4.8: Mean value of the risk ratio (induced by precipitation) versus time periods of the data for both models.

SPATIAL DISTRIBUTION OF RISK RATIO IN CALIFORNIA – In the following part, the risk ratio yielded per station will be projected into maps. The purpose of the maps is to find whether there is any spatial distribution of the duration of hydrological drought changes over time. Nonetheless, it is important to bear in mind, that this part is still a preliminary stage of analysing the spatial distribution. There might be uncertainties, for instance, while interpolating the results. As also has been mentioned in subsection 4.1.2, the quantification of the uncertainty is not scope of the thesis. In general, the maps shows the spatial distribution of hydrological drought duration induced by precipitation in California. The maps are presented in figure 4.9 and figure 4.10, for model 1 and model 2, respectively. In each map, the legend shows in this way: (i) there are five quantiles of different colours. (ii) The red colours shows the first quantile (0 - 0.2) with shorter duration of hydrological drought, meanwhile, the yellow colours indicates the last quantile (0.8 - 1.0) with longer duration of hydrological drought.

In figure 4.9, conclusion can be drawn that in the northern part of California there is change of risk ratio from shorter into longer duration of hydrological drought (number of days with zero discharge) over time due to precipitation. The evidence is shown that the red spots were slowly disappeared, replaced by dominant yellow colour (longer duration). Furthermore, in the southern part (especially South Coast Region), in the beginning the risk ratio were dominated with yellow spots. However, as the time period is updated into the model, the risk ratio of longer days without discharge are replaced by having shorter period.

Moving towards figure 4.10 using model 2, a remarkable outcome depicts in North California. Spread of strong red dots indicating risk of having shorter drought duration drastically changed into yellow dots. For northern part, the risk ratio distribution of model 2 works in an opposite way compared to model 1. However, a similar trend occurs in the South California. Also in a broader term, maps of model 2 results in yellowish distribution indicating higher risk ratio, whereas maps of model 1 shows more orange distribution (lower risk ratio). This conclusion is also aligned with the mean value of risk ratio when comparing two models in figure 4.8.

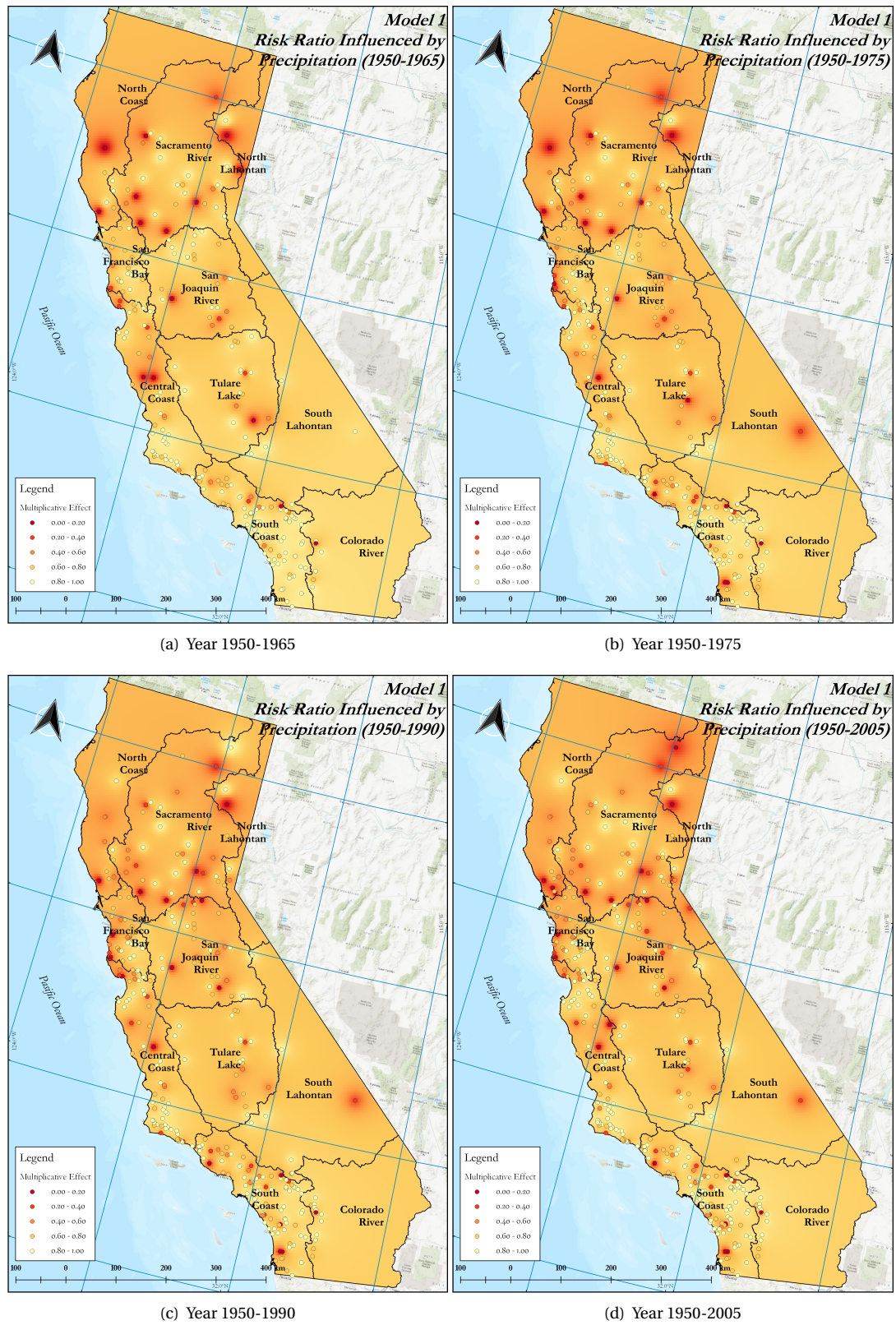


Figure 4.9: Spatial distribution map of risk ratio of number of days without discharge induced by precipitation using model 1

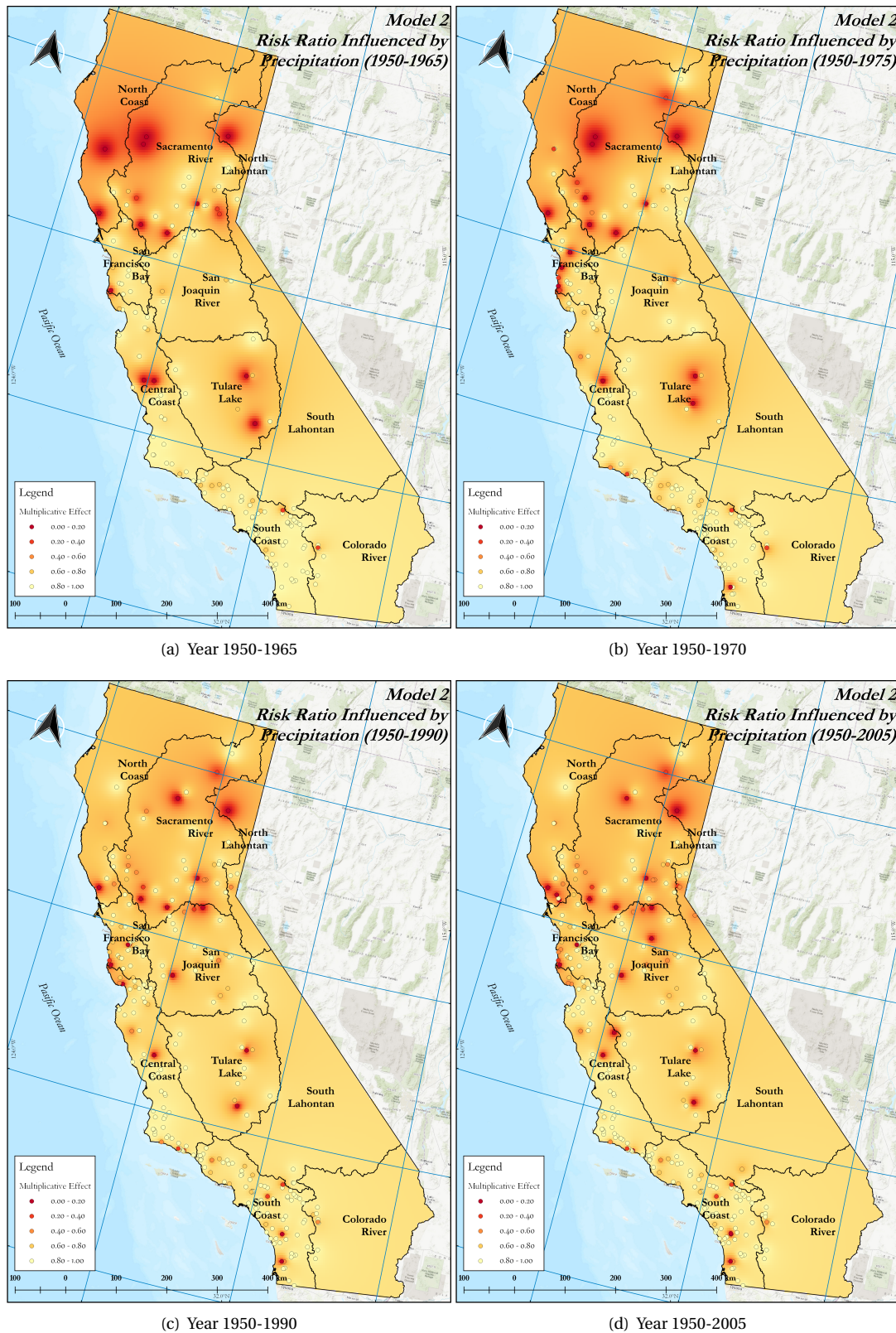


Figure 4.10: Spatial distribution map of risk ratio of number of days without discharge induced by precipitation using model 2

4.2.4. ANOVA: Comparing the Result of Scenario 3

In this subsection, a further analysis towards the risk ratio model on scenario 3 will be discussed. The outcome of risk ratio in scenario 3 shows a slightly difference mean value for both models, that is presented

in 4.13. Analysis of variance is selected because this technique is flexible and practical for comparing sev-

Table 4.13: The statistical summary on the risk ratio result on scenario 3

	Model 1				Model 2			
Time periods	1	2	3	4	1	2	3	4
Mean	0.7764	0.7468	0.7275	0.7301	0.8250	0.8167	0.8155	0.8000
Variance	0.0610	0.0609	0.0610	0.0643	0.0806	0.0803	0.0699	0.0777

eral means. It is not about analysing the population variance, rather the treatment means by identifying the source of variability of the data [56]. Therefore, ANOVA (see section 2.6 for theoretical background) enables us to test whether the slight difference of the mean happens due to chance variations or due to inherent differences among the population. There are several analysis that will be conducted in this section, those are as follows:

1. to test whether at least one of the means in table 4.13 is statistically different than the others (One-way ANOVA)
2. to test whether or not time period and model have interaction between each other (Two-way ANOVA)
3. If both above tests do not show a significant result, therefore, a multi-comparison test introduced by *Tukey-Kramer* will be performed.

One-way ANOVA Result

In this part, the result of the first analysis will be presented. The analysis will check the mean equality for both models that is presented in table 4.13 using one-way ANOVA. The motivation is to check if the slight difference of the mean for each model happened by chance or based on its distribution. Therefore, the null hypothesis for this analysis applied for both models is whether there is an equality between mean value of risk ratio over time. Table 4.14 and 4.15 display the one-way ANOVA result for model 1 and model 2, respectively.

Table 4.14: Table of result for one-way ANOVA model 1 (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the *F*-statistics)

Source	SS	df	MS	F	p-value
Groups	0.4099	3	0.1366	2.2059	0.0857
Error	73.9040	1193	0.0619		
Total	74.3139	1196			

Table 4.15: Table of result for one-way ANOVA model 2 (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the *F*-statistics)

Source	SS	df	MS	F	p-value
Groups	0.0659	3	0.0220	0.2870	0.8348
Error	66.3098	867	0.0765		
Total	66.3756	870			

These tables present the result of the formula demonstrated in table 2.1. The conversion of f-statistic to the p-value (retrieved from MATLAB result) will ease to evaluate the null hypothesis. If we look at the p-value of both models ($p - value_{model1}$: 0.0857 and $p - value_{model2}$: 0.8348), both p-values are below than 0,05. It means that null hypothesis cannot be rejected. In other words, we could come to a conclusion that we cannot neglect the fact that there is an equality between the means for both models.

Two-way ANOVA Result

After there is no significant evidence of different means for both models, we need to check also the interaction between factors of grouping the samples. In our case, the categorised factors would be the time periods and the types of the models (equations) (see subsection 4.1.1 about scenario 3). The motivation beyond this analysis is to find whether apart from precipitation (as the main input in hydrological cycle), all other variables

that involve in the models are significant throughout the time. This is also a important part before proceeding to multiple comparison, to ensure that there is no interaction between considered factors. Therefore the interaction of the *time* and *model* will be checked, and the following null hypotheses are formulated for two-way ANOVA:

1. The mean of all groups of time-period factor are equal.
2. The mean of all groups of the types of models are equal.
3. There is no interaction between these two factors.

Table 4.14 shows the result of Two-way ANOVA for two independent factors, *time* and *model* retrieved from MATLAB. The p-value for factor time-periods fails to reject the null hypothesis with 0.1893 ($> 5\%$), meaning

Table 4.16: Table Two-way ANOVA (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the *F*-statistics)

Source	SS	d.f.	MF	F	p-value
time	0.325	3	0.1084	1.5921	0.1893
model	2.314	1	2.3136	33.9933	6.40E-09
time*model	0.088	3	0.0294	0.4320	0.7301
Error	140.208	2060	0.0681		
Total	143.149	2067			

that there is no mean difference between groups. On contrary, the *model* factors p-value is less than 0.05, meaning that the null hypothesis is rejected, and there is evidence to suggest that the equality of the true mean. This implies that the outcome from different models happened due to the distribution of the data. Moreover, the result shows there is no interaction between two factors, since we cannot reject the hypothesis (3). This result also implies that there is no influence of different model factor throughout the time periods.

Multiple Comparison Model Result

Before proceeding to multiple comparison method, the information regarding ANOVA of the entire groups of data should be obtained, as the prerequisite of performing multiple comparison. It is expected that the result of the one-way ANOVA should not reject the null hypothesis. Therefore, another one-way ANOVA analysis was performed, this time, by collaborating all groups of data as the input of the analysis. The p-value in table

Table 4.17: Table one-way ANOVA of groups of data (where SS is the sum of squares; d.f is degree of freedom; MS is the mean squares; F is the *F*-statistics)

Source	SS	d.f.	MF	F	p-value
Groups	2.941	7	0.4202	6.173	3.61E-07
Error	140.208	2060	0.0681		
Total	143.149	2067			

4.17 shows the value is far less than 0.05, meaning that the null hypothesis is rejected. From this hypothesis, one could conclude that there is enough evidence to suggest the difference of true mean. Another worth mentioning prerequisite of performing multiple comparison is that there is no interaction between two independent grouping factors (*model* and *time*). Thus, now the Tukey's method is valid to be performed

As explained in section 2.6, Tukey's method will evaluate whether each group of data differs significantly towards the other. This will be done by testing all pairwise combination of groups. The null hypothesis would a mean value difference for the considered pairwise. The result mean estimate difference of all pairwise combination of groups using *Tukey-Kramer Methods* are tabulated in table 4.18.

Based on Tukey's Theorem stated in equation 2.14, the null hypothesis can rejected on each case of pairwise combination if zero is not contained between the difference of mean estimate and the high confidence interval. Therefore, if we look carefully for example the pairwise combination number 12 (between group Model 1 - 1970 and Model 2 -1990), there is no zero contained between the mean difference (-0.0687) and the higher confidence interval (-0.0006). Alternatively saying that we are going to reject the null hypothesis if the higher confidence interval is at least less than 0 (High C.I. < 0). Another method to test the hypothesis is by looking

Table 4.18: Table of mean estimate difference for all combinations of two groups (where Lower C.I. is the lower confidence interval; δ ME is the difference of mean estimate; High C.I. is the high confidence interval)

No.	Group	Group	Lower C.I.	Δ ME	High C.I.	p-value
1	Model 1 - 1965	Model 1 - 1970	-0.0396	0.0296	0.0987	0.9008
2	Model 1 - 1965	Model 1 - 1990	-0.0184	0.0489	0.1162	0.3511
3	Model 1 - 1965	Model 1 - 2005	-0.0197	0.0463	0.1122	0.3978
4	Model 1 - 1965	Model 2 - 1965	-0.1304	-0.0487	0.0331	0.6171
5	Model 1 - 1965	Model 2 - 1970	-0.1172	-0.0404	0.0364	0.7549
6	Model 1 - 1965	Model 2 - 1990	-0.1098	-0.0391	0.0316	0.7017
7	Model 1 - 1965	Model 2 - 2005	-0.0934	-0.0237	0.0461	0.9701
8	Model 1 - 1970	Model 1 - 1990	-0.0452	0.0193	0.0839	0.9855
9	Model 1 - 1970	Model 1 - 2005	-0.0464	0.0167	0.0798	0.9931
10	Model 1 - 1970	Model 2 - 1965	-0.1578	-0.0782	0.0013	0.0575
11	Model 1 - 1970	Model 2 - 1970	-0.1443	-0.0699	0.0045	0.0834
12	Model 1 - 1970	Model 2 - 1990	-0.1368	-0.0687	-0.0006	0.0460
13	Model 1 - 1970	Model 2 - 2005	-0.1203	-0.0532	0.0138	0.2382
14	Model 1 - 1990	Model 1 - 2005	-0.0637	-0.0026	0.0585	1.0000
15	Model 1 - 1990	Model 2 - 1965	-0.1755	-0.0976	-0.0196	0.0037
16	Model 1 - 1990	Model 2 - 1970	-0.1620	-0.0892	-0.0165	0.0049
17	Model 1 - 1990	Model 2 - 1990	-0.1542	-0.0880	-0.0218	0.0014
18	Model 1 - 1990	Model 2 - 2005	-0.1378	-0.0726	-0.0074	0.0170
19	Model 1 - 2005	Model 2 - 1965	-0.1717	-0.0949	-0.0182	0.0044
20	Model 1 - 2005	Model 2 - 1970	-0.1581	-0.0866	-0.0152	0.0058
21	Model 1 - 2005	Model 2 - 1990	-0.1502	-0.0854	-0.0206	0.0017
22	Model 1 - 2005	Model 2 - 2005	-0.1337	-0.0699	-0.0062	0.0200
23	Model 2 - 1965	Model 2 - 1970	-0.0779	0.0083	0.0946	1.0000
24	Model 2 - 1965	Model 2 - 1990	-0.0713	0.0095	0.0904	1.0000
25	Model 2 - 1965	Model 2 - 2005	-0.0550	0.0250	0.1050	0.9814
26	Model 2 - 1970	Model 2 - 1990	-0.0746	0.0012	0.0771	1.0000
27	Model 2 - 1970	Model 2 - 2005	-0.0583	0.0167	0.0916	0.9976
28	Model 2 - 1990	Model 2 - 2005	-0.0532	0.0154	0.0841	0.9975

at the p-value yielded from the model. The bold p-value (number 12, 15 - 22) shows the combinations of pairwise groups that have true mean differences (as the p-value is less than 0.05) and we cannot reject the null hypothesis for remaining combinations. These pairwise combinations also have high C.I. less than 0.

Now the focus of interest will be narrowed down to combinations that have true mean. These groups are Model 1 - 1990 and Model 1 - 2005. Figure 4.11 shows which groups have significant differences with respect to Model 1 - 1990 and Model 1 - 2005. The blue line in graphs illustrates the focussed group (or the fixed variable), whereas the red lines are the groups that has significant differences with respect to the considered group.

An interesting conclusion could be derived from figure 4.11. It exhibits that as the model 1 has been updated until 1990, there is a true mean difference compared to all the results in model 2. The differentiation between model 1 and model 2 is the explanatory variables that control the response variable. Therefore, given we only know precipitation, the time-updated model 1 is underestimating the risk ratio of number of days with zero discharge induced by precipitation. It is concluded that using model 1 is not recommended. Having said that this conclusion emphasizes that hydrological drought, particularly duration parameter, involves more variables rather just precipitation. This is also relevant, as in the concept of drought propagation (introduced in subsection 2.1.2), that hydrological drought as a final product of drought mechanism requires many variables and other mechanisms in it, not solely on precipitation.

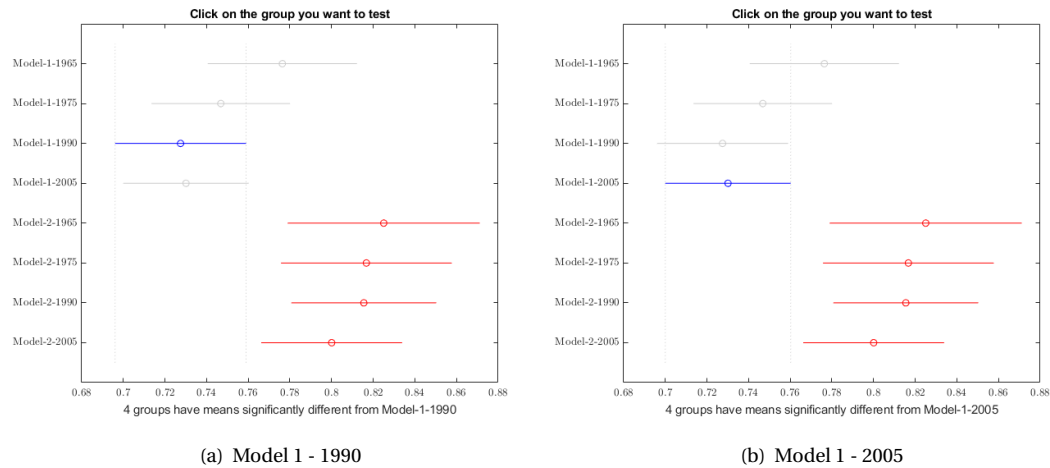


Figure 4.11: Multicomparison on Model 1 - 1990 and Model 1 - 2005

4.3. Conclusion

In this section, the main findings of the analysis will be highlighted in the following points:

1. As we introduce more parameters into the model, the risk ratio will also increase, in a sense of having less significance towards the response variable. Important to note the limitation of using risk ratio is that the influence of particular variable could be only expressed individually towards the variable of interest, which in this case the hydrological drought duration. This limitation does not provide a macro-scale perspective of hydrological drought duration, but rather a preliminary knowledge towards the response of variable of interest on a certain covariate (sensitivity analysis).
2. Based on the outcome per hydrological region in California, if the trend of mean monthly precipitation is increasing for 1 mm/day, the coastal area will experience 0.88 - 13.63 % reduced risk of number of days with zero discharge per month. Whereas, in river dominated area, it will have 0.26 - 1.4 % of reduction in the same risk. Moreover, increase mean air temperature of 1 degree Celcius will lead to 3.85 - 8.64 increased of having longer duration in coastal areas, while only 4.25 - 5.74 % in river dominated areas. The above conditions hold if the regression is accepted.
3. Risk ratio of hydrological drought duration expressed in number of days without discharge induced by precipitation is increasing with the size of catchment area.
4. As the time series is updated, the general trend shows a slightly declining trend of mean values of risk ratio for both types of models induced by precipitation. However, this argument is rejected statistically using one-way ANOVA. The result analysis shows that there is no significant true mean difference.
5. Statistically, there is no interaction between the time and model factors. However, there is enough evidence to say about the true mean difference between the considered models. This evidence is established when the time series for model 1 is updated until 1990 compared to the whole groups of model 2. Another evidence is that model 1 tends to underestimate the risk ratio, which makes model 2 performs better than model 1. This becomes also relevant to the fact that hydrological drought event happens not solely due to the effect of lack of precipitation.

Conclusion and Recommendation

This chapter delivers the conclusion which tries to answer the research questions that were formulated in section 1.3; and also points out some recommendations for the sake of future research.

5.1. Conclusion

In this section, the author will motivate the answer for the formulated research questions. The first research question:

What are the hydro-meteorological variables that have a strong influence in characterising hydrological drought parameters?

This question could be answered if one accept the performance of the model. The following table presents variables that have strong relation in explaining hydrological drought parameters (severity and duration) particularly in California.

Hydrological Drought Parameters	Severity	Duration
Variable of Interest	Low Stream-Flow	Number of Days without Discharge in a month
Hydro-Meteorological Variables	Air Temperature, Atmosphere Water Vapour Content, Evaporation, Snowmelt Flux, Precipitation Flux, Snowfall Flux, Soil Moisture, SOF Coefficient, BFI, Stream Density, Surface Runoff Coefficient	Precipitation Flux, Air Temperature, Athmosphere Water Vapour Content, Snowmelt Flux, SPI-01

The author will also try to answer the second formulated research question.

What are the advantages (and disadvantages) of hydrological drought models, particularly using statistical models?

As mentioned in section 1.2, that such mechanism of the drought event is complicated, as a consequence it is hard to explain its physical processes. To date, all the evidences that are closely related to characterise drought were taken into account to study its broader trend. Therefore, the lesson learn of this research of using statistical tools to model the hydrological drought will be elaborated based on the hydrological drought parameters.

HYDROLOGICAL DROUGHT SEVERITY USING BAYESIAN NETWORK

1. Lack of input variables

Rainfall-Runoff concept which believes to be a good way to measure the propagation of drought, involves many type of variables. There are many parameters that are not included in the this model and might possibly relevant to the drought processes. These variables either may be hardly or costly to be measured. One of the parameter that might play a big role in drought propagation is the time-lag transmission from meteorological to hydrological drought.

2. The type input variables and its uncertainty

As explained in 3.1.2 regarding the type of feeding data of the model, this study utilised most of data using spatial dataset, are accessibly open to everyone. This model ease the user to feed another data from another drought-prone areas, as long as such data matches the characteristics of type of hydrological drought. Thus, this benefit will give a better overview to study global drought condition, regardless the limitation of ground measurement data. In long run, if the quantification of drought hazard becomes more accessible to drought hotspot areas, the mitigation of disaster risk management will be much easier to perform. However, the trade-off using such dataset is the uncertainty of the dataset that usually is not well quantified.

HYDROLOGICAL DROUGHT DURATION USING GLM

1. Time series as an input variable for the model

In nature, the duration of hydrological drought does not solely depend on the climatological variables. One of the challenge using GLM is that type of feeding data to fit the model should be time series. Nevertheless, there are many variables related to the geometrical aspect of the watershed that also play a role in the drought duration.

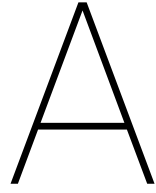
2. Risk Ratio as one of the measures to study the drought duration

Risk Ratio allows the user to intuitively interpret the influenced of only particular variable towards the variable of interest. This limitation does not provide a macro-scale perspective of hydrological drought duration, but rather a preliminary knowledge towards the response of variable of interest on a certain covariate (sensitivity analysis). In fact, the complexity of drought event should be explained by a particular condition that requires, sometimes more than one variable, in a way to demonstrate the drought event.

5.2. Recommendation

In this particular section, some recommendations in order to improve the performance of the statistical model for future research will be elaborated. Those are as follow:

1. To make a very clear distinction of the type of hydrological drought on the top of the general mechanism for the assumption of the model.
2. Hazard quantification is a preliminary stage in disaster risk management. The combination between hazard and consequences will be more informative in mitigating drought risk, however not the scope of the study.
3. An uncertainty often occurs in the option of validating method. Another type of validation methods would be worth to quantify also the uncertainty of validating the model.
4. This research does not combine the interaction between hydrological drought severity and duration. Its relationship is important toward each other, as the deficit accumulates over the duration of the drought event [74].



Input Variables: Possible Relevant Variables in Drought Mechanism

This appendix elaborates a brief explanation of the input variables for the possible variables in explaining drought mechanism. Features in the propagation of drought controlled by the catchment characteristic and climate condition[75].

A.1. Spatial Datasets

Most of the variables to be used in the model were derived from CORDEX in North American Region. The historical datasets were the outcome of the regional climate modelling of RCA4 under SMHI. Those variables are as follow:

1. Precipitation Flux
2. Snowfall Flux
3. Evaporation Flux
4. Snow Melt Flux
5. Air Temperature
6. Surface Temperature
7. Soil Moisture Content
8. Atmosphere Water Vapour Content
9. Surface Runoff Coefficient

All of the values of these variables were converted from mean daily into mean monthly. In order to get the representative value for each delineated drainage basin, mean weighted area were used.

Precipitation Index

The SPI is a statistical indicator that compares the cumulated precipitation during a period of n months with the long-term cumulated rainfall distribution for the same location and accumulated period [65]. Standardized precipitation is the ratio between the difference of mean precipitation for a specified time period and standard deviation, where both are calculated from the past records. The moving average period is to determine a set of time scales of period i months where i is 3,6,12,24, or 48 months [46]. The result of the moving average is used to fitted the probability versus precipitation. Using the information between two relationship, in combination with an estimate of the inverse normal to calculate the deviation for a normally distributed pdf ($N(\mu = 0, \sigma = 1)$), SPI could be computed at any precipitation data point.

A.2. Hydrological Variables: Drainage Basin Characteristics

In general, the characteristic of drainage basin resembles the geometry or physical representation of the catchment. Basin characteristic plays role on the controlling of the water flow path within the catchment.

1. **COMPACTNESS:** This parameter expresses the ratio between the perimeter of the basin and a circle of an equal area to the basin. The minimum value equals to unity for a circular area. Compactness could mathematically states as follow

$$C = \frac{P}{2\sqrt{\pi M}} \quad (\text{A.1})$$

where C is the compactness coefficient; P is the perimeter of the drainage basin; and M is the area of the drainage basin.

2. **STREAM DENSITY :** Drainage density which represented in stream density reflects the geological and petrographical conditions of a catchment and acts as an index representing the infiltration capacity and the transmissivity of the soil [68]. It is the ratio of number of stream/river networks per unit area within the catchment area. Mathematically, this parameter is formulated as follow

$$d_s = \frac{N_s}{M} \quad (\text{A.2})$$

where d_s is stream density; N_s is number of river networks; and M is the drainage basin area.

3. **BASE FLOW INDEX:** The BFI or base flow index is a ratio of base flow to total streamflow, expressed as a percentage and ranging from 0 to 100. Base flow is the sustained, slowly varying component of streamflow, usually attributed to ground-water discharge to a stream.
4. **SATURATED OVERLAND FLOW:** Dunne overland flow, also know as saturation overland flow, is generated in a basin when the water table "outcrops" on the land surface (due to the infiltration and redistribution of soil moisture within the basin), thereby producing temporary saturated areas. These saturated areas generate Dunne overland flow through ex-filtration of shallow ground water and by routing precipitation directly to the stream network. The time frame of SOF is in minutes to hours.
5. **HORTONIAN OVERLAND FLOW:** Horton overland flow, also known as infiltration-excess overland flow, is generated in a basin when infiltration rates are exceeded by precipitation rates.
6. **TOPOGRAPHY WETNESS INDEX (TWI):** It is an index developed by Beven and Kirby, 1979 to measure the topography control on hydrological process/flow paths. This index also helps to indicate the steady wetness index on a certain point of the catchment.

$$TWI = \ln \frac{a}{\tan \beta} \quad (\text{A.3})$$

where a is the up-slope area per unit contour length; and $\tan \beta$ is the slope gradient at the point measurement.

7. **SUBSURFACE-FLOW CONTACT TIME INDEX:** Subsurface flow contact time index. The subsurface contact time index estimates the number of days that infiltrated water resides in the saturated subsurface zone of the basin before discharging into the stream [source: GAGES II].

B

Result of Partial Correlation of All Input Variables

This appendix presents the figures yielded from the computation between the variable of interest (q_{MM}) and all possibly relevant variables. There are 21 graphs presented in figure [B.1](#), [B.2](#), and [B.3](#). To obtain this bivariate analysis, the marginal data were transformed into standard normal. Then the Pearson's product moment correlation were computed for four quadrants (NE , NW , SE , SW). On one hand to compute the positive correlation, semi-correlations in the upper right (NE) and lower left (SW) quadrants are:

$$\begin{aligned}\rho_{NE} &= \rho(Z_1, Z_2 | Z_1 > 0, Z_2 > 0) \\ \rho_{SW} &= \rho(Z_1, Z_2 | Z_1 < 0, Z_2 < 0)\end{aligned}$$

where Z_1 and Z_2 are the standard normal transforms of two random variables. On the other hand to compute the negative correlation, semi-correlation in the upper left (NW) and lower right (SE) quadrants are ρ_{NW} and ρ_{SE} are define similarly [\[34, 53\]](#). In a broader sense, the larger (absolute) values of the semi-correlations compared to the overall correlation (in the tittle of each graph) exemplify the tail dependence between two variables. This implies, in general, that Gaussian copula might not be a good representation, despite its fast computation. The information in figure [B.1](#), [B.2](#) and [B.3](#) will be used in table [3.5](#).

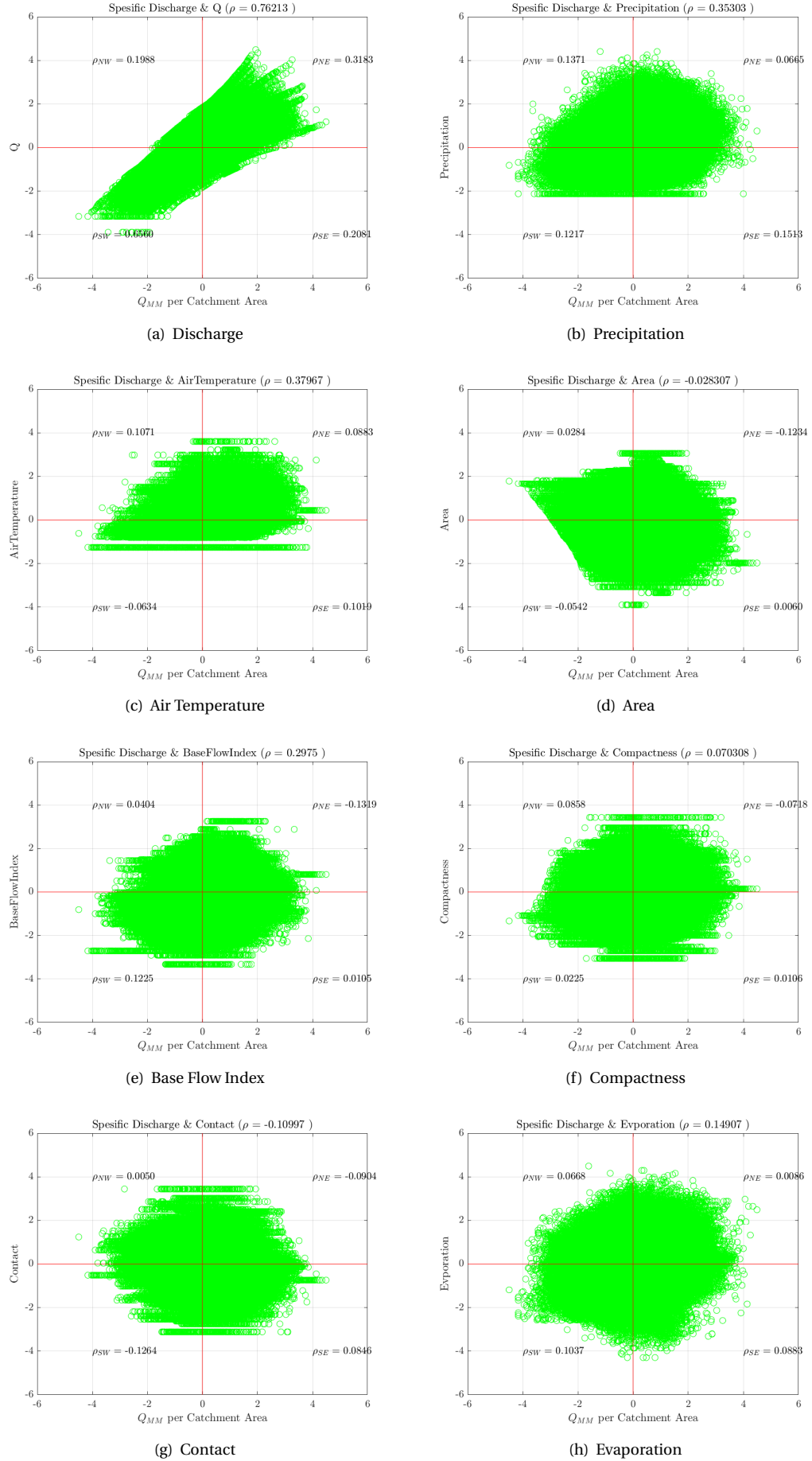
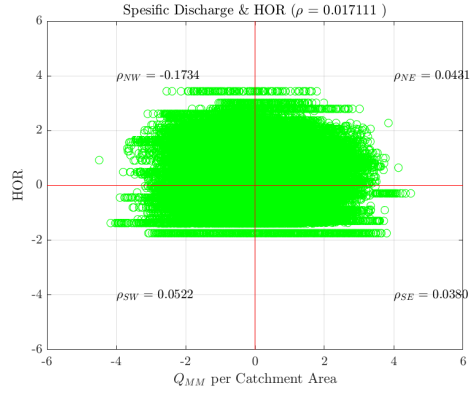
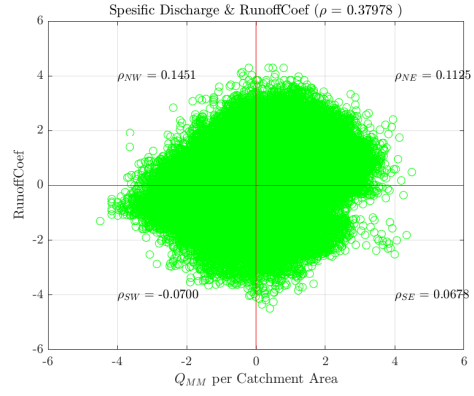


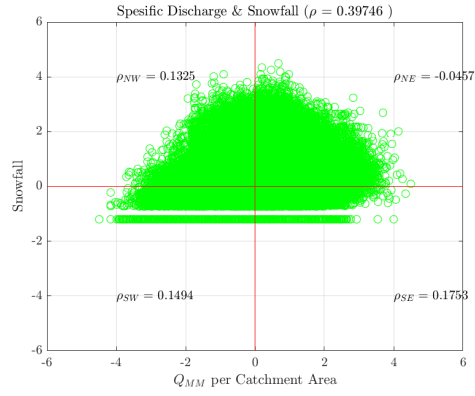
Figure B.1: Partial Correlation of mean monthly discharge per catchment area and input variables (1)



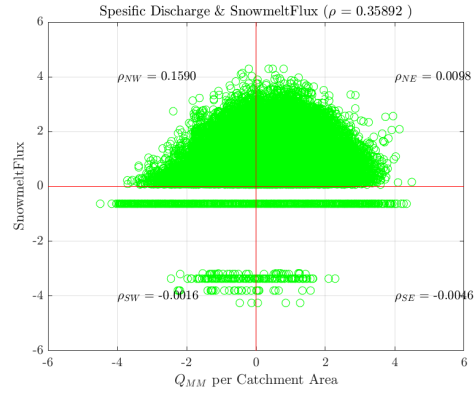
(a) HOF Coefficient



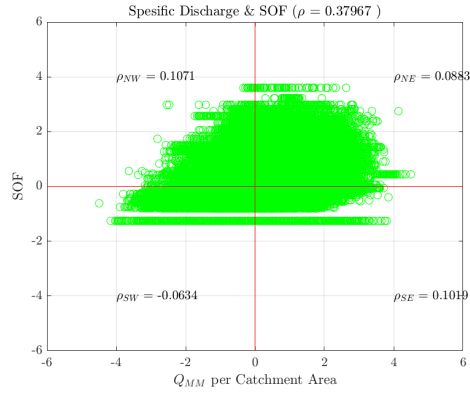
(b) Runoff Coefficient



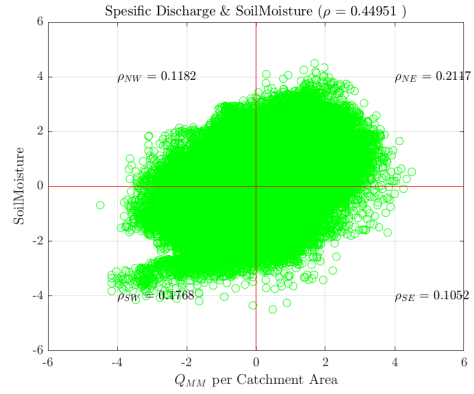
(c) Snofall Flux



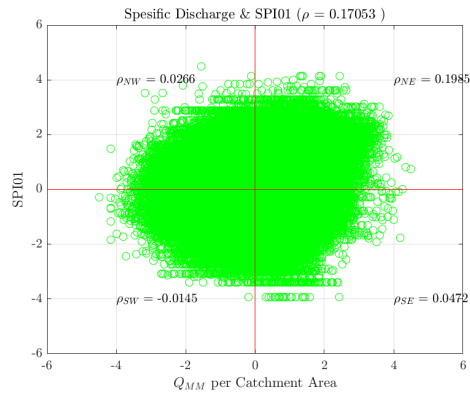
(d) Snow melt Flux



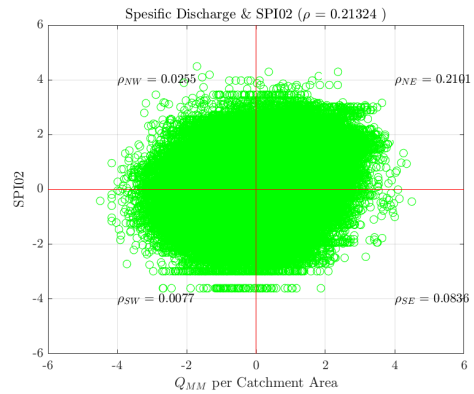
(e) SOF Coefficient



(f) Soil Moisture Content



(g) SPI-01



(h) SPI-02

Figure B.2: Partial Correlation of mean monthly dishcharge per catchment area and input variables (2)

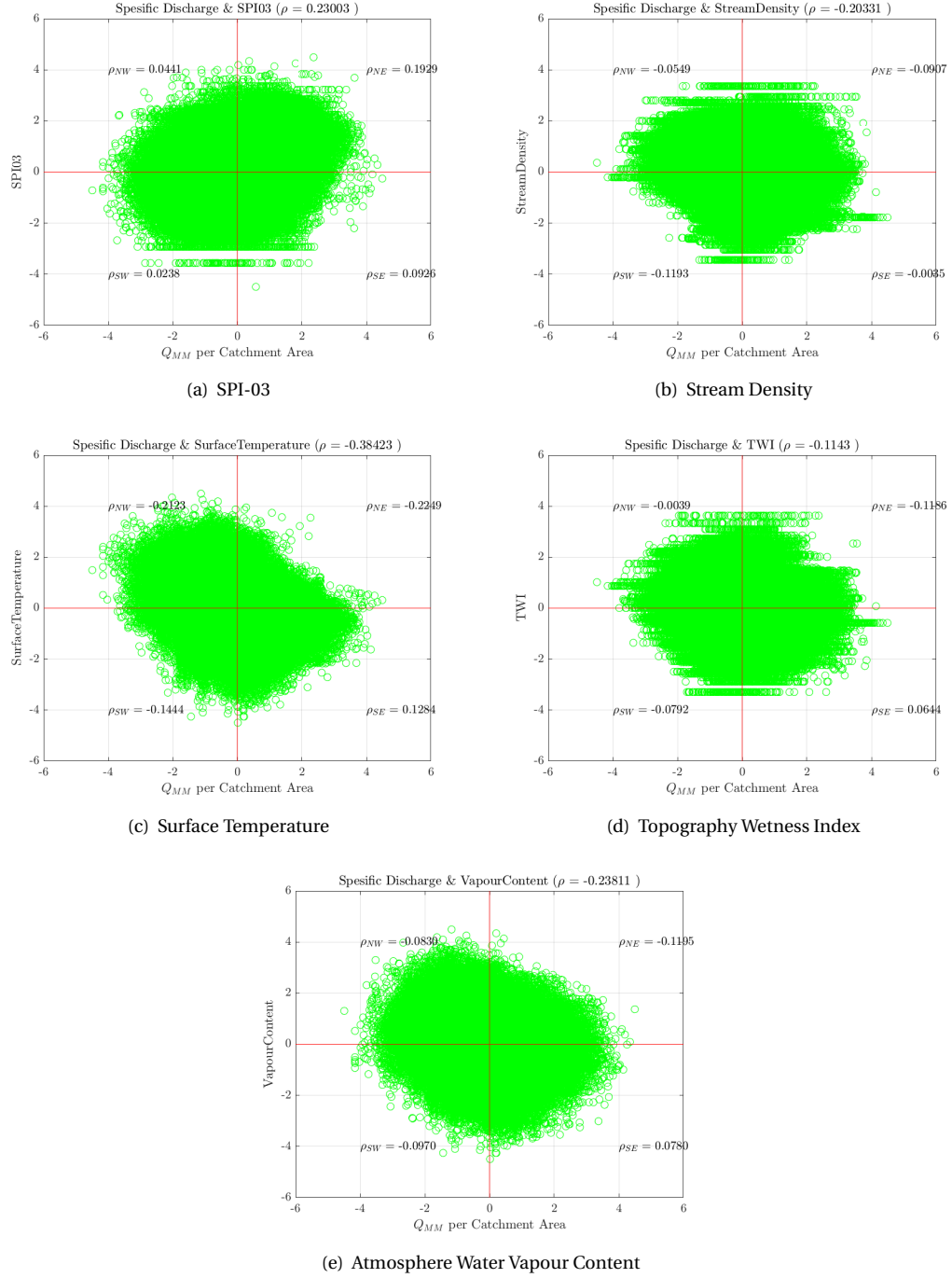


Figure B.3: Partial Correlation of mean monthly dishcharge per catchment area and input variables (3)

Bibliography

- [1] The mean (constant, intercept-only) model for forecasting. URL <https://people.duke.edu/~rnanu/411mean.htm>.
- [2] Jules J. Beersma and T. Adri Buishand. Joint probability of precipitation and discharge deficits in the Netherlands. *Water Resources Research*, 40(12):1–11, 2004. ISSN 00431397. doi: 10.1029/2004WR003265.
- [3] Tiao J Chang. Investigation of precipitation droughts by use of-kriging method. *Journal of Irrigation and Drainage Engineering (ASCE)*, 117(6):935–943, 1992. doi: [https://ascelibrary.org/doi/pdf/10.1061/\(ASCE\)0733-9437\(1991\)117%3A6\(935\)](https://ascelibrary.org/doi/pdf/10.1061/(ASCE)0733-9437(1991)117%3A6(935)).
- [4] Tiao J Chang and J. R. Stenson. IS IT REALISTIC TO DEFINE A 100-YEAR DROUGHT FOR WATER MANAGEMENT. *Water Resources Bulletin (AWRA)*, 26(5):823–829, 1990.
- [5] Jennifer Chaussee. California Drought Expected to Cost State \$2.2 Billion in Losses - Scientific American, 2014. URL <https://www.scientificamerican.com/article/california-drought-expected-to-cost-state-2-2-billion-in-losses/>.
- [6] Kwok Pan Chun, Howard Wheeler, and Christian Onof. Prediction of the impact of climate change on drought: an evaluation of six UK catchments using two stochastic approaches. *Hydrological Processes*, 27(11):1600–1614, 2013. ISSN 08856087. doi: 10.1002/hyp.9259.
- [7] B. Clausen and C.P. Pearson. Regional frequency analysis of annual maximum streamflow drought. *Journal of Hydrology*, 173(1-4):111–130, 1995. ISSN 00221694. doi: 10.1016/0022-1694(95)02713-Y. URL <http://www.sciencedirect.com/science/article/pii/002216949502713Y>
<http://linkinghub.elsevier.com/retrieve/pii/002216949502713Y>.
- [8] A A O Couasnon. Characterizing flood hazard at two spatial scales with the use of stochastic models. 2017.
- [9] CRED. Economic loss due to major droughts worldwide up to 2016 | Statista. URL <https://www.statista.com/statistics/267700/economic-loss-due-to-drought-worldwide/>.
- [10] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, 50(3):885–900, 2007. ISSN 2151-0040. doi: 10.13031/2013.23153. URL <http://elibrary.asabe.org/abstract.asp??JID=3&AID=23153&CID=t2007&v=50&i=3&T=1>.
- [11] H. A. R. De bruin and J. N. M. Stricker. Evaporation of grass under non-restricted soil moisture conditions. *Hydrological Sciences Journal*, 45(3):391–406, 2000. ISSN 0262-6667. doi: 10.1080/02626660009492337. URL <http://www.tandfonline.com/doi/abs/10.1080/02626660009492337>.
- [12] Noah S. Diffenbaugh, Daniel L. Swain, and Danielle Touma. Anthropogenic warming has increased drought risk in California. *Proceedings of the National Academy of Sciences*, 112(13):3931–3936, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1422385112. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1422385112>.
- [13] Maxx Dille, Robert S. Chen, Uwe Deichmann, Arthur L Lener-Lam, Margareth Arnold, Jonathan Agwe, Piet Buys, Oddvar Kjekstad, Bradfield Lyon, and Gregory Yetman. *Natural Disaster Hotspots A Global Risk Analysis*. Number 5. 2005. ISBN 9780821359303. doi: 10.1080/01944360902967228.
- [14] Annette J Dobson. *An introduction to generalized linear models*. 2002. ISBN 1-58488-165-8. doi: 10.1198/tech.2002.s91.
- [15] John A. Dracup, Kil Seong Lee, and Edwin C G. Paulson. On the statistical characteristics of drought event. *Water Resour Res*, 16(2):289–296, 1980. ISSN 0043-1397. doi: 10.1029/WR016i002p00289.

- [16] Elfatih A.B. Eltahir. Drought frequency analysis of annual rainfall series in central and western sudan. *Hydrological Sciences Journal*, 37(3):185–199, 1992. ISSN 21503435. doi: 10.1080/02626669209492581.
- [17] Zekâi Şen. *Applied Drought Modeling, Prediction, and Mitigation*. 2015. ISBN 9780128021767. doi: 10.1016/B978-0-12-802176-7.00006-7. URL <http://www.sciencedirect.com/science/article/pii/B9780128021767000067>.
- [18] Zekâi Şen. *Applied Drought Modeling, Prediction, and Mitigation*. 2015. ISBN 9780128021767. doi: 10.1016/B978-0-12-802176-7.00006-7. URL <http://www.sciencedirect.com/science/article/pii/B9780128021767000067>.
- [19] Encyclopædia Britannica. California, 2018. URL <https://www.britannica.com/place/California-state>.
- [20] Encyclopedia.com. California facts, information, pictures | Encyclopedia.com articles about California. URL <https://www.encyclopedia.com/places/united-states-and-canada/us-political-geography/california>.
- [21] J.A. Falcone. Geospatial Attributes of Gages of Evaluating Streamflow. Technical Report -, 2011.
- [22] M. Falkenmark, J. Rockström, and L. Karlberg. Present and future water requirements for feeding humanity. *Food Security*, 1(1):59–69, 2009. ISSN 1876-4517. doi: 10.1007/s12571-008-0003-x. URL <http://link.springer.com/10.1007/s12571-008-0003-x>.
- [23] FAO. Drought. *FAO Land and Water*, pages 1–4, 2014.
- [24] FEMA. National mitigation strategy: Partnerships for building safer communities, 1995. URL <http://hazardmitigation.calema.ca.gov/docs/1995{ }National{ }Mitigation{ }Strategy{ }web.pdf>.
- [25] David M. Frick, Dennis Bode, and Jose D. Salas. Effect of Drought on Urban Water Supplies. I: Drought Analysis. *Journal of Hydraulic Engineering*, 116(6):733–753, 1990.
- [26] Joshua C. Galster, Frank J. Pazzaglia, Bruce R. Hargreaves, Donald P. Morris, Stephen C. Peters, and Richard N. Weisman. Effects of urbanization on watershed hydrology: The scaling of discharge with drainage area. *Geology*, 34(9):713–716, 2006. ISSN 00917613. doi: 10.1130/G22633.1.
- [27] Christian Genest and Anne-Catherine Favre. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007. ISSN 1084-0699. doi: 10.1061/(ASCE)1084-0699(2007)12:4(347). URL <http://ascelibrary.org/doi/10.1061/{%}28ASCE{%}291084-0699{%}282007{%}2912{%}3A4{%}28347{%}29>.
- [28] Daniel Griffin and Kevin J Anchukaitis. How unusual is the 2012 – 2014 California drought ? *Geophysical Research Letters*, 41:9017–9023, 2014. ISSN 00948276. doi: 10.1002/2014GL062433.1.
- [29] Gunnar Hagman, Henrik Beer, M Bendz, and A Wijkman. Prevention better than cure. Report on human and environmental disasters in the Third World. 2. 1984.
- [30] Anca Hanea, Oswaldo Morales Napoles, and Dan Ababei. Non-parametric Bayesian networks: Improving theory and reviewing applications. *Reliability Engineering and System Safety*, 144:265–284, 2015. ISSN 09518320. doi: 10.1016/j.ress.2015.07.027. URL <http://dx.doi.org/10.1016/j.ress.2015.07.027>.
- [31] A. Jan Hendriks, Aafke M. Schipper, Marloes Caduff, and Mark A.J. Huijbregts. Size relationships of water inflow into lakes: Empirical regressions suggest geometric scaling. *Journal of Hydrology*, 414-415:482–490, 2012. ISSN 00221694. doi: 10.1016/j.jhydrol.2011.11.025. URL <http://dx.doi.org/10.1016/j.jhydrol.2011.11.025>.
- [32] Hege Hisdal, Lena M. Tallaksen, B. Clausen, Elisabeth Peters, and A Gustard. Hydrological Drought Characteristic. In Lena M. Tallaksen and H. A.J. Van Lanen, editors, *Development in Water Science*, chapter 5, pages 139–198. Elsevier B.V., Amsterdam, 1st edition, 2004.

- [33] J.W. Hopmans. Infiltration and Unsaturated Zone. *Treatise on Water Science*, pages 103–114, 2011. doi: 10.1016/B978-0-444-53199-5.00031-2. URL <http://www.sciencedirect.com/science/article/pii/B9780444531995000312>.
- [34] Harry Joe. Dependence Modeling with Copulas. *CRC Press*, page 479, 2015. doi: 10.1007/0-387-28678-0.
- [35] Dale Kalser and Phillip Reese. California drought impact pegged at \$2.7 billion | The Sacramento Bee, 2015. URL <http://www.sacbee.com/news/state/california/water-and-drought/article31396805.html>.
- [36] KNMI. KNMI - Precipitation Deficiency / Drought, . URL <https://www.knmi.nl/nederland-nu/klimatologie/geografische-overzichten/neerslagtekort{ }droogte>.
- [37] KNMI. KNMI - Continuous potential precipitation surplus, . URL <https://www.knmi.nl/nederland-nu/klimatologie/geografische-overzichten/neerslagoverschot>.
- [38] Z. W. Kundzewicz, L. J. Mata, N. W. Arnell, P. Döll, P. Kabat, B. Jiménez, K. Miller, T. Oki, Z. Şen, and I.A. Shiklomanov. Freshwater resources and their management. In *Climate Change 2007: Impact, Adaptation, and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 173–210. Cambridge University Press, Cambridge, 2007.
- [39] Nadja Leith. Using generalised linear models to simulate daily rainfall under scenarios of climate change. (December), 2005.
- [40] A F Van Loon. *On the Propagation of Drought: how climate and catchment characteristics influence hydrological drought development and recovery*. PhD thesis, Wageningen University, 2013.
- [41] Shahrbanou Madadgar and Hamid Moradkhani. Drought Analysis under Climate Change Using Copula. *Journal of Hydrologic Engineering*, 18(7):746–759, 2012. ISSN 1084-0699. doi: 10.1061/(ASCE)HE.1943-5584.
- [42] Shahrbanou Madadgar and Hamid Moradkhani. Spatio-temporal drought forecasting within Bayesian networks. *Journal of Hydrology*, 512:134–146, 2014. ISSN 00221694. doi: 10.1016/j.jhydrol.2014.02.039. URL <http://dx.doi.org/10.1016/j.jhydrol.2014.02.039>.
- [43] Terry Marsh, Gwyneth Cole, and Rob Wilby. Major droughts in England and Wales, 1800-2006. *Weather*, 62(4):87–93, 2007. ISSN 14778696. doi: 10.1002/wea.67.
- [44] Valérie Masson-delmotte, Wilfran Moufouma-okia, and Sonia Seneviratne. Drough in a changing climate: AR5 and recent scientific advances. Technical report, 2016.
- [45] P Mccullagh and Ja Nelder. Generalized linear models, 1972. ISSN <null>. URL <http://www.jstor.org/stable/10.2307/2344614>.
- [46] Thomas B Mckee, Nolan J Doesken, and John Kleist. The relationship of drought frequency and duration to time scales. *AMS 8th Conference on Applied Climatology*, (January):179–184, 1993. ISSN 0899-8418. doi: citeulike-article-id:10490403. URL <http://ccc.atmos.colostate.edu/relationshipofdroughtfrequency.pdf>.
- [47] Merriam-Webster. Decile | Definition of Decile by Merriam-Webster. URL <https://www.merriam-webster.com/dictionary/decile>.
- [48] Ashok K. Mishra and Vijay P. Singh. A review of drought concepts. *Journal of Hydrology*, 391(1-2):202–216, 2010. ISSN 00221694. doi: 10.1016/j.jhydrol.2010.07.012. URL <http://dx.doi.org/10.1016/j.jhydrol.2010.07.012>.
- [49] S. Mohan and N. C. V. Rangacharya. A modified method for drought identification. *Hydrological Sciences Journal*, 36(1):11–21, 1991. ISSN 0262-6667. doi: 10.1080/02626669109492481. URL <http://www.tandfonline.com/doi/abs/10.1080/02626669109492481>.
- [50] Newworldencyclopedia.org. California, 2016. URL <http://www.newworldencyclopedia.org/entry/California>.

- [51] J.R. Nimmo. Vadose Water. *Encyclopedia of Inland Waters*, 1:766–777, 2009. doi: 10.1016/B978-012370626-3.00014-4. URL <http://www.sciencedirect.com/science/article/pii/B9780123706263000144>.
- [52] Pacific Institute. California drought (Background), 2016. URL <http://www.californiadrought.org/drought/background/>.
- [53] Dominik Paprotny and Oswaldo Morales-Nápoles. Estimating extreme river discharges in Europe through a Bayesian network. *Hydrology and Earth System Sciences*, 21(6):2615–2636, 2017. ISSN 16077938. doi: 10.5194/hess-21-2615-2017.
- [54] Subin Pinkayan. Conditional Probabilities of Occurrence of Wet and Dry Years Over a Large Continental Area. *Hydrology Papers (Colorado State University)*, 12(April):47, 1966.
- [55] Deepthi Rajsekhar, Vijay P. Singh, and Ashok K. Mishra. Hydrologic Drought Atlas for Texas. *Journal of Hydrologic Engineering*, 20(7):05014023, 2015. ISSN 1084-0699. doi: 10.1061/(ASCE)HE.1943-5584.0001074. URL <http://ascelibrary.org/doi/10.1061/{%}28ASCE{%}29HE.1943-5584.0001074>.
- [56] Kandethody M. Ramachandran and Chris P. Tsokos. *Mathematical Statistics with Applications*. 2015. ISBN 9780124171138. doi: 10.1016/B978-0-12-417113-8.00009-6.
- [57] Renji Remesan and Jimson Mathew. *Hydrological Data Driven Modelling*. 2015. ISBN 978-3-319-09234-8. doi: 10.1007/978-3-319-09235-5. URL <http://link.springer.com/10.1007/978-3-319-09235-5>.
- [58] H. E. Robin Burgers, Aafke M. Schipper, and A. Jan Hendriks. Size relationships of water discharge in rivers: Scaling of discharge with catchment area, main-stem length and precipitation. *Hydrological Processes*, 28(23):5769–5775, 2014. ISSN 10991085. doi: 10.1002/hyp.10087.
- [59] Todd Rowland and Eric W. Weisstein. Canonical. URL <http://mathworld.wolfram.com/Canonical.html>.
- [60] Maria Alzira Santos. Regional Droughts: A Stochastic Characterization. *Journal of Hydrology*, 66:183–211, 1983. URL https://ac.els-cdn.com/0022169483901853/1-s2.0-0022169483901853-main.pdf?{_}tid=5e7dedfa-e445-40c4-97bb-0d7fb045b6ce{&}acdnat=1520951504{_{}}610c764f001fc0a6c2760dc18e5adf1e.
- [61] H.H.G Savenije. Hydrology of Catchments, Rivers and Deltas. Technical report, Delft University of Technology, Delft, 2006.
- [62] Justin Sheffield and Eric F Wood. Drought: Past Problems and Future Scenarios (Hardback) - Taylor & Francis, 2011. URL <http://www.taylorandfrancis.com/books/details/9781849710824/>.
- [63] Ji Yae Shin, Muhammad Ajmal, Jiyoung Yoo, and Tae Woong Kim. A Bayesian network-based probabilistic framework for drought forecasting and outlook. *Advances in Meteorology*, 2016:8–13, 2016. ISSN 16879317. doi: 10.1155/2016/9472605s.
- [64] V. U. Smakhtin. Low flow hydrology: A review. *Journal of Hydrology*, 240(3-4):147–186, 2001. doi: 10.1016/S0022-1694(00)00340-1.
- [65] Jonathan Spinoni, Gustavo Naumann, Hugo Carrao, Paulo Barbosa, and Jürgen Vogt. World drought frequency, duration, and severity for 1951-2010. *International Journal of Climatology*, 34(8):2792–2804, 2014. ISSN 10970088. doi: 10.1002/joc.3875.
- [66] Kerstin Stahl. *Hydrological Drought: A Study Across Europe*. Doctoral dissertation, Institut für Hydrologie der Universität, 2001. URL <https://freidok.uni-freiburg.de/fedora/objects/freidok:202/datastreams/FILE1/content>.
- [67] Kerstin Stahl and Hege Hisdal. Hydroclimatology. In L. M. Tallaksen and H. A.J. Van Lanen, editors, *Development in Water Science*, pages 19–51. Elsevier B.V., Amsterdam, 1 edition, 2004. ISBN 0-444-51688-3.
- [68] L. M. Tallaksen and H. A.J. Van Lanen. *Hydrological drought: processes and estimation methods for streamflow and groundwater*. Elsevier, 48 edition, 2004.

- [69] The National Drought Mitigation Center. Time Series Data of Drought in California, 2018. URL <http://droughtmonitor.unl.edu/Data/Timeseries.aspx>.
- [70] Fatih Tosunoğlu and Christian Onof. Joint modelling of drought characteristics derived from historical and synthetic rainfalls: Application of Generalized Linear Models and Copulas. *Journal of Hydrology: Regional Studies*, 14(October):167–181, 2017. ISSN 22145818. doi: 10.1016/j.ejrh.2017.11.001.
- [71] USGS. USGS Land Cover Institute, 2016. URL https://landcover.usgs.gov/states/california/california_{_}big.jpg.
- [72] USGS. California Droughts Compared | USGS California Water Science Center, 2017. URL <https://ca.water.usgs.gov/california-drought/california-drought-comparisons.html>.
- [73] H. A.J. Van Lanen, N. Wanders, L. M. Tallaksen, and A. F. Van Loon. Hydrological drought across the world: Impact of climate and physical catchment structure. *Hydrology and Earth System Sciences*, 17(5): 1715–1732, 2013. ISSN 10275606. doi: 10.5194/hess-17-1715-2013.
- [74] A. F. Van Loon and G. Laaha. Hydrological drought severity explained by climate and catchment characteristics. *Journal of Hydrology*, 526:3–14, 2015. ISSN 00221694. doi: 10.1016/j.jhydrol.2014.10.059. URL <http://dx.doi.org/10.1016/j.jhydrol.2014.10.059>.
- [75] Anne F. Van Loon. Hydrological drought explained. *Wiley Interdisciplinary Reviews: Water*, 2(4):359–392, 2015. ISSN 20491948. doi: 10.1002/wat2.1085. URL <http://doi.wiley.com/10.1002/wat2.1085>.
- [76] Richard T. Wetherald and Syukuro Manabe. Simulation of hydrologic changes associated with global warming. *Journal of Geophysical Research Atmospheres*, 107(19), 2002. ISSN 01480227. doi: 10.1029/2001JD001195.
- [77] Donald A. Wilhite. Drought as a natural hazard: Concepts and definitions. *Drought: A Global Assessment*, pages 3–18, 2000.
- [78] World Economic Forum. *The global risks report 2018, 13th edition*. 2018. ISBN 978-1-944835-15-6. doi: 978-1-944835-15-6. URL [{_}0Ahttps://www.weforum.org/reports/the-global-risks-report-2018](http://www3.weforum.org/docs/WEF_{_}GRR18_{_}Report.pdf).
- [79] Emir Zelenhasić and Atila Salvai. A method of streamflow drought analysis. *Water Resources Research*, 23(1):156–168, 1987. ISSN 19447973. doi: 10.1029/WR023i001p00156.