

School of Information Technologies  
Faculty of Engineering & IT

## ASSIGNMENT/PROJECT COVERSHEET - GROUP ASSESSMENT

Unit of Study: ISYS5050 Knowledge Management Systems

Assignment name: Group Assignment

Tutorial time: 6:00 p.m. Thursday Tutor name: Pouya Salehpour

### DECLARATION

We the undersigned declare that we have read and understood the [University of Sydney Academic Dishonesty and Plagiarism in Coursework Policy](#), and, except where specifically acknowledged, the work contained in this assignment/project is our own work, and has not been copied from other sources or been previously submitted for award or assessment.

We understand that failure to comply with the *Academic Dishonesty and Plagiarism in Coursework Policy* can lead to severe penalties as outlined under Chapter 8 of the *University of Sydney By-Law 1999* (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.

We realise that we may be asked to identify those portions of the work contributed by each of us and required to demonstrate our individual knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

Project team members				
Student name	Student ID	Participated	Agree to share	Signature
1. Shengyue Qian	500098518	Yes / No	Yes / No	Shengyue Qian
2. Sicheng Guo	510191768	Yes / No	Yes / No	Sicheng Guo
3. Yixuan Liu	500269853	Yes / No	Yes / No	Yixuan Liu
4. Lerong Ouyang	500663486	Yes / No	Yes / No	Lerong Ouyang
5. Yushan Yan	490136906	Yes / No	Yes / No	Yushan Yan



# **The Analysis of Flood Hazard and Impact**

**ISYS5050 Group 18**

**Shengyue Qian    500098518**

**Yixuan Liu    500269853**

**Sicheng Guo    510191768**

**Lerong Ouyang    500663486**

**Yushan Yan    490136906**

**Group Project  
May 25<sup>th</sup>, 2022**

## **Abstract**

Flood has been considered as a major catastrophe around the global along with the negative impacts from several aspects including economy, culture and society. The main objective of this project is to take the analysis of flood frequency, flood magnitude, pattern of flood in terms of seasonal change, the damage and destruction the floods impact around the world overtime and other factors that might be relevant to triggering the occurrence of the flood. The study would utilize “FloodArchive” dataset provided by Dartmouth Flood Observatory, which includes details on all major flood events recorded globally between 1985 and 2021 with over 5000 such cases. The data has been pre-processed by Python. All cleaned data has been input into the Tableau to perform several visualization graphs in support of the analysis.

**Keywords:** *Flood, analysis, Tableau, visualization*

## Table of Contents

<b>Abstract .....</b>	<b>1</b>
<b>Introduction.....</b>	<b>4</b>
<b>1. Data Preparation &amp; Pre-processing.....</b>	<b>4</b>
1.1 Data Quality Issues .....	4
1.1.1 Irregularities in country names (formatting and spelling errors) .....	4
1.1.2 Missing values .....	4
1.1.3 Fact Check .....	5
1.1.4 Irregular MainCause written.....	6
1.2 ETL process and data cleaning steps. ....	6
1.3 Screenshot of important changes, before and after ETL.....	7
1.4 Specifying ETL, BI, and scripting languages.....	7
<b>2. Flood Magnitude Methodology .....</b>	<b>7</b>
2.1 Data Processing .....	7
2.2 Data Visualization .....	8
2.2.1 Top 5 countries in terms of number of floods.....	9
2.2.2 Top 5 countries in terms of average magnitude .....	9
2.2.3 Top 5 countries in terms of magnitude level.....	10
<b>3. Seasonal patterns over time.....</b>	<b>11</b>
3.1 Data pre-processing .....	11
3.2 Pattern analysis of flood data .....	11
3.2.1 Analysis on the trend of flood frequency with seasonal variation .....	11
3.2.2 Analysis of flood data of different severity class .....	3
3.2.3 Analysis of the trend of Main Cause data over seasons .....	4
3.2.4 Comprehensive analysis of section 3.2 .....	6
<b>4. Analysis of the damage and impact of floods.....</b>	<b>7</b>
4.1 Impact of flood change over time .....	7

4.2 Countries and areas most affected .....	8
<b>5. Further analysis of floods driving factors.....</b>	<b>11</b>
5.1 Related work and assumptions .....	11
5.2 Relevant data obtaining and cleaning up .....	12
5.2.1 Forest area and CO2 emission data from Worldbank.org.....	12
5.2.2 Data pre-processing using Python.....	13
5.3 Analysis of different cases.....	13
<b>Conclusions.....</b>	<b>15</b>
<b>References (APA 7<sup>th</sup> Edition).....</b>	<b>16</b>

# Introduction

A flood occurs when a big volume of water exceeds its usual limitations. It can also refer to the influx of the wave or the backflow of the river at the point where the rivers meet. The trigger of the flood varies in multiple ways. Most common and general floods caused by nature are heavy rain, detrimental hurricanes, severe storms, snow melt in the mountains. Human-caused mishaps, such as a dam break, over-releasing stream for electricity generation, excessive emission of carbon dioxide are all prone to triggering the flood. Some of floods might happen in causal while others in a regular pattern. In this case, it is vital to have a thorough analysis of the flood to mitigate the impact of flooding to a certain extent.

## 1. Data Preparation & Pre-processing

### 1.1 Data Quality Issues

#### 1.1.1 Irregularities in country names (formatting and spelling errors)

From the original dataset provided, there are some syntactical errors such as typos, and wrong spelling of the words. We have categorized them into several detailed classes as follows.

The first standard class of typos is that some country names contain unnecessary whitespaces. For example, as shown in figure 1.1(a), there exist ‘ Peru’, ‘Peru’ and ‘Peru ’ respectively from id 1494, 1509, and 1708. Although it may not be disturbing for human readers, the analytical data software may treat them as 3 totally different entities, affecting the accuracy of our result.

ID	GlideNum	Country
1494	0	Peru
1509	0	Peru
1708	0	Peru

a. Unnecessary whitespace

ID	Country
2469	Bosnia-Herzegovina
3507	Bosnia-Herzegovenia

b. Wrong word spelling

Figure 1.1. An example of a typo

The second common typo is the wrong spelling inside the file. It is particularly common for country names. figure1(b) gives the example of the country name ‘Bosnia-Herzegovina’. Id 3507 uses an incorrect English spelling of the name in the original file. Similar cases appear such as ‘El Salvador/ El Savalor’, ‘Cote d'Ivoire/Cote D'Ivoir’ and so on.

#### 1.1.2 Missing values

There are issues with missing values in “GlideNumber” and “MainCause”. As shown in Figures 1.2. a & b.

1	ID	GlideNumber	Count	Other	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
1039	1038		Philippines		124.471	6.76939	10938.81	12/24/1995	12/27/1995	News	0	0		1
1072	1071		USA		-81.1983	38.4217	63664.15	5/13/1996	5/20/1996	News	0	0		1
1074	1073		Yemen		47.415	15.4695	190981.1	5/14/1996	5/18/1996	News	0	0		1
3093	3093		Oman		57.933	28.5367	372988.5	6/6/2007	6/12/2007	News	61	60000	Tropical cyclone	1

1	ID	GlideN	Count	Other	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
1039	1038		Philippines		124.471	6.76939	10938.81	12/24/1995	12/27/1995	News	0	0		1
1072	1071		USA		-81.1983	38.4217	63664.15	5/13/1996	5/20/1996	News	0	0		1
1074	1073		Yemen		47.415	15.4695	190981.1	5/14/1996	5/18/1996	News	0	0		1
3093	3093		Oman		57.933	28.5367	372988.5	6/6/2007	6/12/2007	News	61	60000	Tropical cyclone	1

Figures 1.2.a&b Example of missing values in “GlideNumber” & “MainCause”

### 1.1.3 Fact Check

The values of the data “Displaced” and “Area” deviate from the real situation. Take “Displaced” as an example shown in Figure 1.3, the number of people “Displaced” in 2017 in the data provided for the Philippines is 140,000, compared to the actual situation of 2,529,000, which is ten times less. (data. world bank, 2022)



Figure 1.3 Number of people displaced in 2017 in the Philippines.

### 1.1.4 Irregular MainCause written

From the original dataset provided, there are many causes of flooding, and the dataset does not have a common separation symbol, which can lead to the same type of cause not being analysed together. The different causes in the dataset are separated by other symbols such as '/', 'and', 'or', etc. This is shown in Figures 1.4.a, 1.4.b, 1.4.c, 1.4.d.

1	ID	Glides	Count	Other	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
125	124	0	Ecuador	0	-77.4907	-0.131721	31908.09	3/5/1987	3/7/1987	News	300	15000	Dam/Levy, break or release	:
126	125	0	Peru	0	-76.7345	-11.8797	4672.27	3/9/1987	3/11/1987	News	12	1000	Dam/Levy, break or release	:
258	257	0	Nigeria	0	7.10119	6.37902	8884.3	9/21/1988	10/5/1988	News	0	70000	Dam/Levy, break or release	:

a. Use 'or' to separate causes

1	ID	Glides	Count	Other	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
4082	4083		China	Russia	130.272	50.5323	1430155.	8/7/2013	10/14/2013	News	100	260000	Heavy Rain, began with #4079	2

b. The presence of special symbols.

1	ID	Glides	Count	Other	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
5075	5075		India		85.5114	21.6184	345642.7	5/25/2021	5/28/2021	FloodList	1	1500000	Tropical Storm Yaas and Storm Surg	1.5

c. Use 'and' to separate the reasons.

1	ID	Glides	Count	Other	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
18	17	0	China	0	127.897	48.0917	300325.0	4/21/1985	4/29/1985	News	0	200	Ice jam/break-up	1
71	70	0	USA	0	-116.954	44.0047	1742.13	1/1/1986	1/6/1986	News	0	100	Ice jam/break-up	1
75	74	0	USA	0	-87.8814	45.0248	4321.14	3/27/1986	4/5/1986	News	2	496	Ice jam/break-up	1
627	626	0	Canada	0	-117.494	56.4607	7996.65	2/27/1992	2/29/1992	News	0	4000	Ice jam/break-up	1
632	631	0	USA	0	-72.5285	44.2707	1311.44	3/12/1992	3/14/1992	News	0	0	Ice jam/break-up	2
644	643	0	Canada	0	-115.994	60.5419	1537.3	4/24/1992	4/26/1992	News	0	100	Ice jam/break-up	1
649	648	0	Canada	0	-78.1083	52.6169	88075.01	5/17/1992	5/18/1992	News	0	800	Ice jam/break-up	1
807	806	0	China	0	107.087	40.4443	12109.08	12/7/1993	12/15/1993	News	0	5200	Ice jam/break-up	1

d. Use '/' to separate the causes.

Figures 1.4 a&b&c&d Some special writing in MainCause.

## 1.2 ETL process and data cleaning steps.

The ETL process implemented for the dataset is as follows.

- The first step was to extract data from the Excel file and the Data. world bank website.
- Based on the problems we found, operations such as Validation, cleaning, formatting, aggregating, and enriching were then carried out. The details of the processing are as follows.
- For 1.1.1 Country name errors, we imported Python's Nominatim and CountryInfo packages and made comparative corrections.
- For 1.1.2 Missing value issue. We chose to make improvements by filling in the zeros and adding Nan.
- For 1.1.3 the problem of non-conformity with the truth. We chose to remove the incorrect data that was ten times different from the real data by comparing it with the real data.



- For 1.1.4 Maincause writing problem. We chose to write '/', 'and', 'or', ',' and other symbols uniformly with '/' separating them.
- Finally, the processed data was loaded, and the processed data was stored as a CSV file and exported for later application in Tableau BI.

### 1.3 Screenshot of important changes, before and after ETL.

We are using the Python language for the data cleaning transformation operation of the ETL. The specific screenshots would be shown below.

1	ID	Glidel	Count	Otheri	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
1039	1038		Philippines		124.471	6.76939	10938.81	12/24/1995	12/27/1995	News	0	0		1

Figure 1.5 Dealing with missing values in Gildenumber and MainCause (Before processing)

1	ID	Glidel	Count	Otheri	long	lat	Area	Began	Ended	Valida	Dead	Displa	MainCause	Severi
1039	1038	0	Philippines		124.471	6.76939	10938.81	12/24/1995	12/27/1995	News	0	0	nan	1

Figure 1.6 After processing

Dam/Levy/ break or release
Dam/Levy/ break or release

Figure 1.7 After processing

### 1.4 Specifying ETL, BI, and scripting languages.

In our data processing, our ETL process was done by using both python and NLP approaches. Editing was done using Anaconda coding software. Correction transforms were made by comparing the data in the official toolkit and Data. world bank. Finally, we will use Tableau BI for our analysis.

## 2. Flood Magnitude Methodology

It is vital to have a methodology to assess the severity of flood situation, otherwise the raw data or single combination of data is not able to effectively make the analysis of the impact of flood. According to the definition of Flood Magnitude severity calculation formula provided by the climate scientists, it can be noticed that three attributes are involved, which are Severity Class, Affected Area and Duration respectively. Multiple three of them correspondingly and then using a log function. That is the formula of Flood Magnitude.

### 2.1 Data Processing

- ***Flood Magnitude Calculation***

In terms of the “FloodArchive” dataset itself, it has already been cleaned and pre-processed in the former section. In this case, it could be utilized and deemed as a solid source. However, the dataset only contains two columns that are “Severity

Class” and “Affected Area” while “Duration” is not included in. In this respect, one new column “Duration” has been added. Duration, namely the length of the time, is relevant to “Began” and “Ended”. Under the circumstances, the Duration is calculated by the “Ended” to minus “Began” respectively, with days as the unit of it. After “Duration” has been added, three indispensable components consist of Flood Magnitude are all set up. The log function has been deployed after multiplication of three of them. The Flood magnitude has been calculated respectively and could be observed in the following [Figure 2.1.1].

=LOG(O113\*N113\*G113)

I	J	K	L	M	N	O	P	Q
ended	Validation	Dead	Displaced	MainCause	Severity	Duration	Flood Magnitude	Level of magnitude
1995/8/15 0:00 News		1530	7000000	Heavy rain		2	92	8.4871 Level 3
2004/10/7 0:00 News		3000	40000000	Monsoonal rain		2	109	8.4869 Level 3
1992/7/2 0:00 News		37	305992	Heavy rain		2	92	8.3944 Level 3

Figure 2.1.1 Calculation of Flood Magnitude

- **Definition of Magnitude Levels**

According to the question above, different levels of magnitude might have separate impacts to the affected countries. In this case, three levels of magnitudes (Level 1, Level 2, Level 3) have been set in accordance with the number in the magnitude. The magnitudes range from 1 to 9, in precise with four decimal places. Level 1 represents the number from 1 to 4, 4 to 7 has been allocated to Level 2 while the rest of them are classified in Level 3 (From 7 to 9) just like the figure shown in the [Figure 2.1.2]. The higher the level, the more severe the magnitude of impact to the corresponding country.

J	K	L	M	N	O	P	Q	R
Validation	Dead	Displaced	MainCause	Severity	Duration	Flood Magnitude	Level of magnitude	
129 News	38	600 Heavy Rain		2	55	7.9705 Level 3		
148 News	18	1000 Rain ; snowmelt		1.5	46	7.8067 Level 3		
156 News	48	31000 Heavy rain		2	60	7.7275 Level 3		
167 News	31	4000 Heavy Rain		2	45	7.6759 Level 3		
179 News	0	38000 Heavy rain		2	28	7.6289 Level 3		
183 News	5	3600 Heavy Rain ; Snowmelt		2	418	7.5869 Level 3		
187 News	1	0 Heavy Rain ; Snow		1.5	21	7.5641 Level 3		
206 News	0	600 Heavy Rain		2	34	7.4540 Level 3		
210 FloodList	2	0 Heavy Rain		1.5	41	7.4220 Level 3		
231 News	39	408000 Heavy Rain		2	22	7.3290 Level 3		
238 News	8	5000 Heavy rain		1.5	27	7.3131 Level 3		
244 News	26	2800 Heavy rain		1.5	11	7.3026 Level 3		
245 News	13	4668 Heavy rain		2	36	7.3012 Level 3		
249 FloodList	3	0 Snowmelt ; Heavy Rain ;		2	20	7.2947 Level 3		
257 News	0	50000 Heavy rain		2	34	7.2835 Level 3		
262 News	1	25000 Heavy Rain		2	39	7.2713 Level 3		
265 News	1	0 Heavy rain		1	23	7.2569 Level 3		
290 News	3	4000 Heavy rain		2	27	7.1886 Level 3		
305	0	0 Heavy Rain Snowmelt Dam		1	14	7.1527 Level 3		
367 News	3	5000 Heavy rain		1	22	7.0165 Level 3		
370 News	1	1100 Heavy Rain ; Snowmelt		2	30	7.0125 Level 3		
388 News	20	100000 Heavy rain		2	31	6.9874 Level 2		
392 News	100	20000 Tropical cyclone		2	13	6.9755 Level 2		
404 News	68	35000 Tropical cyclone		2	26	6.9596 Level 2		

Figure 2.1.2 Example of Labelling Magnitude Level

## 2.2 Data Visualization

The processed data has been input into the Tableau, one of the BI tools for data visualization in order to give the audience a more intuitive approach to interpret the information of the flood magnitude. The following subsections would focus on the top five countries with different approaches including number of floods, average magnitude,

level of magnitude respectively. With different approaches, results vary from each other.

### 2.2.1 Top 5 countries in terms of number of floods

The severity of flood could be interpreted from the aspect of total number of floods. As can be seen in the bar chart [Figure 2.2.1], the horizontal axis corresponds to the quantity of flood across the entire timeline from 1985 to 2021, while the vertical axis represents the country where the flood happened.

In accordance with the following bar chart, it is clear that the United States lists in top of the rank, up to 480 floods around the whole country since records began, followed by China, with 363 floods, approximately 120 less than the former. The lowest ranked country out of the five is the Philippines, with only 182 floods in total, more or less one-third of the amount in that of the United States. India and Indonesia placed the third and fourth, with the quantity of 282 and 224 separately.

**Comparison for Top 5 countries on Flood Magnitude (Based on Number of Floods)**

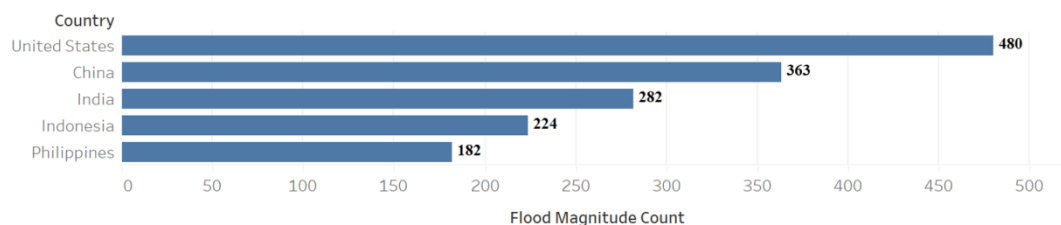


Figure 2.2.1 Number of floods comparison

### 2.2.2 Top 5 countries in terms of average magnitude

As has been mentioned before, flood severity is evaluated by the magnitude just being calculated in the former section. While each flood has a corresponding magnitude value, it is not easy to make the evaluation when they are separate. In this case, the average function has been deployed on the magnitude values by the same country.

The bar chart [Figure 2.2.2] illustrates the comparison among five countries with average magnitude of floods. The horizontal axis represents the magnitude ranging from 1 to 9 combined with the country as the vertical axis. It is clear that Mali suffers the most with the highest value of magnitude 7.3648. Chad and South Sudan ranked second and third, with 6.5841 and 6.5196 respectively. The average magnitude in Sudan and Kazakhstan are all less than 6.5, ranking in the fourth and fifth place with 6.4473 and 6.3843 correspondingly.

It cannot be ignored that only the level of magnitude in Mali is classified into the Level 3 while the other four are all categorized in Level 2.

### Comparison for Top 5 countries on Flood Magnitude (Based on Average Magnitude)

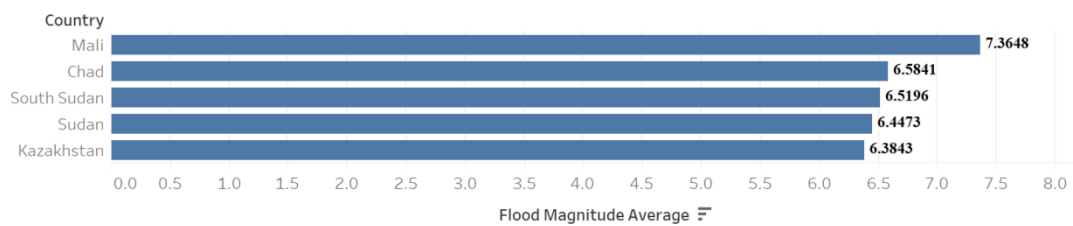


Figure 2.2.2 Average magnitude of floods comparison

### 2.2.3 Top 5 countries in terms of magnitude level

The third method is based on different levels of magnitude. There are three levels being mentioned in the former section corresponding to the magnitude value. The detailed information for all top 5 countries in separate levels has been presented in the table [Table 2.2.3] below. In addition, data in the table has been visualized in the following bar charts.

Table 2.2.3 Summarization of all Levels of Magnitude

Level of Magnitude	Top 5 Countries	AVG Magnitude
Level 3	Angola	7.9315
	Mauritania	7.7126
	Democratic Republic of the Congo	7.6945
	Tanzania	7.6390
	Sudan	7.6125
Level 2	South Sudan	6.4107
	Kazakhstan	6.3843
	Suriname	6.3056
	Turkmenistan	6.2287
	Niger	6.2231
Level 1	Burkina Faso	3.9886
	Morocco	3.9755
	Guyana	3.9439
	Mongolia	3.9415
	Sweden	3.9292

- Level 1**

#### Comparison for Top 5 countries on Flood Magnitude (Level 1)

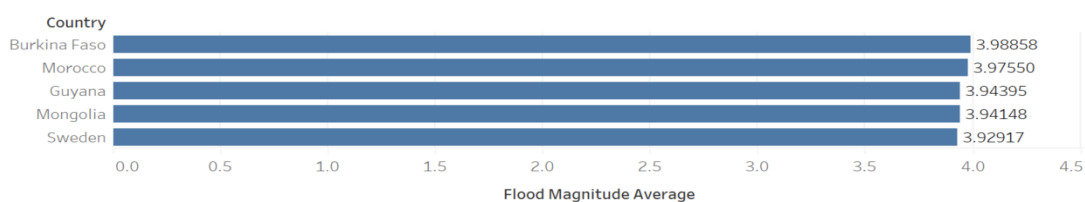


Figure 2.2.3.1 Level 1\_Average magnitude of floods comparison

- **Level 2**

### Comparison for Top 5 countries on Flood Magnitude (Level 2)

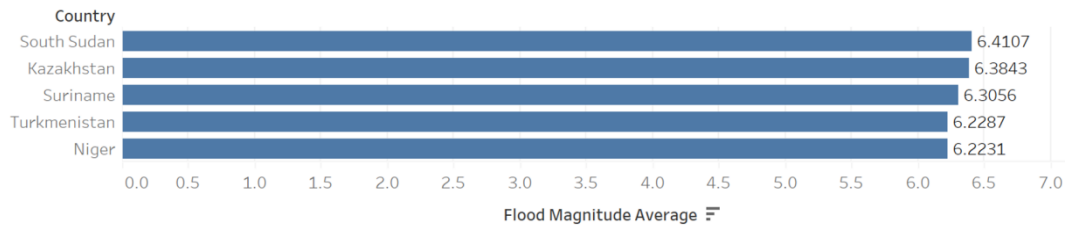


Figure 2.2.3.2 Level 2\_Average magnitude of floods comparison

- **Level 3**

### Comparison for Top 5 countries on Flood Magnitude (Level 3)

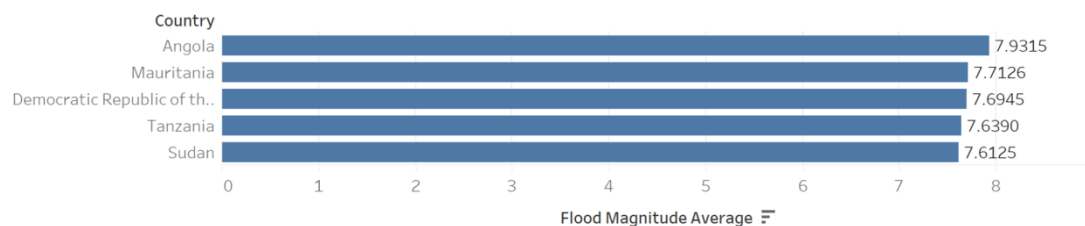


Figure 2.2.3.3 Level 3\_Average magnitude of floods comparison

## 3. Seasonal patterns over time

### 3.1 Data pre-processing

Considering that the geographical locations of different countries may be in different seasons at the same time, for example, the southern hemisphere is in winter in July while the northern hemisphere is in summer. We created the 'geographical location' list to divide the country data into two categories to help find the relationship between flood data and seasons [Figure 3.1.1].

Geographical location									
	A	B	C	D	E	F	G	H	I
1	ID	GlideNumber	Country	OtherCountry	Geographical location	long	lat	Area	Began
2	2	0	Brazil		0 Southern Hemisphere	-45.3489	-18.7111	678498.82	1985/1/15
3	3	0	Philippines		0 Northern Hemisphere	122.974	10.0207	12846.03	1985/1/20
4	4	0	Indonesia		0 Southern Hemisphere	124.606	1.01489	16542.12	1985/2/4
5	5	0	Mozambique		0 Southern Hemisphere	32.3491	-25.8693	20082.21	1985/2/9
6	6	0	Comoros		0 Southern Hemisphere	43.36	-11.6516	1035.61	1985/2/16
7	7	0	New Zealand		0 Southern Hemisphere	175.734	-37.2305	7871.37	1985/2/17
8	8	0	Indonesia		0 Southern Hemisphere	108.14	-7.04008	77091.11	1985/2/19
9	9	0	United States		0 Northern Hemisphere	-85.1742	40.6691	210527.96	1985/2/22
10	10	0	Bolivia		0 Southern Hemisphere	-63.2887	-21.2244	69706.89	1985/2/25
11	11	0	United States		0 Northern Hemisphere	-89.5537	40.6814	26266.14	1985/3/3
12	12	0	United States		0 Northern Hemisphere	-108.093	35.3824	26527.13	1985/3/13
13	13	0	United States		0 Northern Hemisphere	-96.7845	29.6044	141508	1985/3/14
14	14	0	United States		0 Northern Hemisphere	-83.5377	42.0122	16883.54	1985/3/30
15	15	0	Brazil		0 Southern Hemisphere	-48.0987	-4.13519	1970402.21	1985/4/2

Figure 3.1.1. Create "Geographical Location" column

### 3.2 Pattern analysis of flood data

#### 3.2.1 Analysis on the trend of flood frequency with seasonal variation

According to common definitions, December, January, and February in Northern

Hemisphere countries would be classified as winter, March to May as spring, June to August as summer, and September to November as winter. Correspondingly, December to February for Southern Hemisphere countries would be summer, March to May would be autumn, June to August would be winter, and September to November would be spring ("Australian weather and the seasons - australia.gov.au", 2008).

(For all the visualizations in question 3 below, the month numbers represent the total data for the next three months. For example, 12/1995 represents December 1995 to February 1996, and 3/1996 represents March to May 1996. In the following visual analysis, the conclusion is drawn without considering the values after September 2021 because the data are incomplete.)

According to this division, the variation of flood data with time and season is shown in [Figure 3.2.1.1]. It is clear that the total number of floods in the northern hemisphere is much higher than in the southern hemisphere, and the data vary considerably from time to time. The overall period from 1985 to 2002 shows a small increase and peaks in the June-August period of 2002 with a total number of floods of 108, followed by a fluctuating decrease in value.

As shown in [Figure 3.2.1.2] and Figure [3.2.1.3] sorted by the flood data of the northern and southern hemispheres, it can be found that the high frequency of floods in the northern hemisphere is usually the period June to August, and in the southern hemisphere, December to February. The corresponding time periods with low frequency of flooding are December to February and March to May for the Northern Hemisphere, and June to August and September to November for the Southern Hemisphere.

<Statistics on the total number of floods per season>

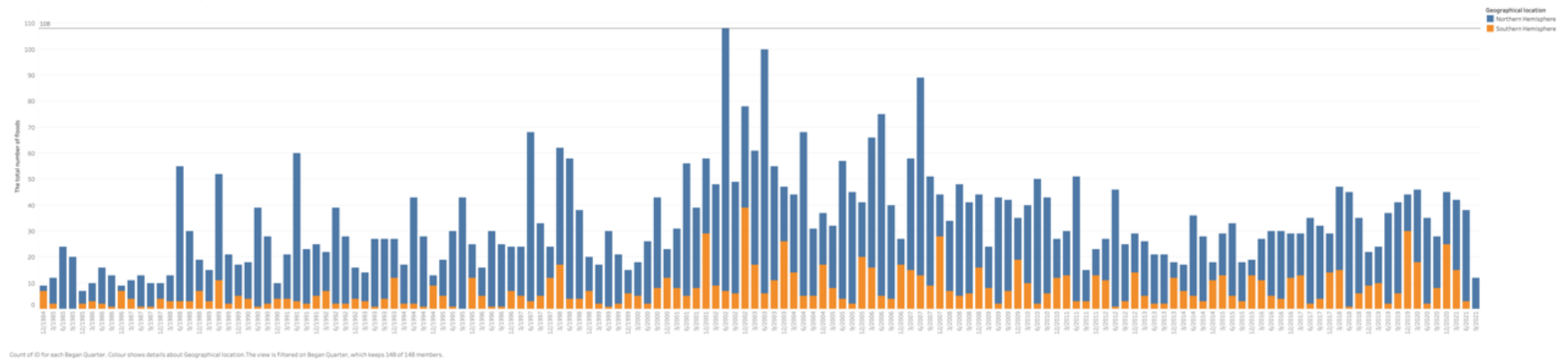


Figure 3.2.1.1. Statistics on the total number of floods per season

<Statistics on the total number of floods in northern hemisphere per season>

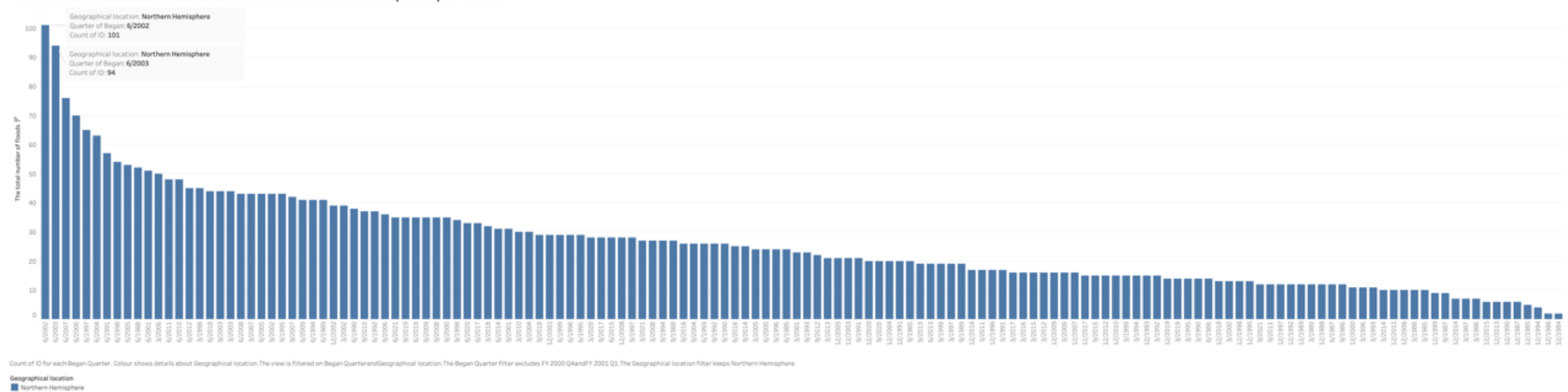


Figure 3.2.1.2. Statistics on the total number of floods in northern hemisphere per season

<Statistics on the total number of floods in southern hemisphere per season>

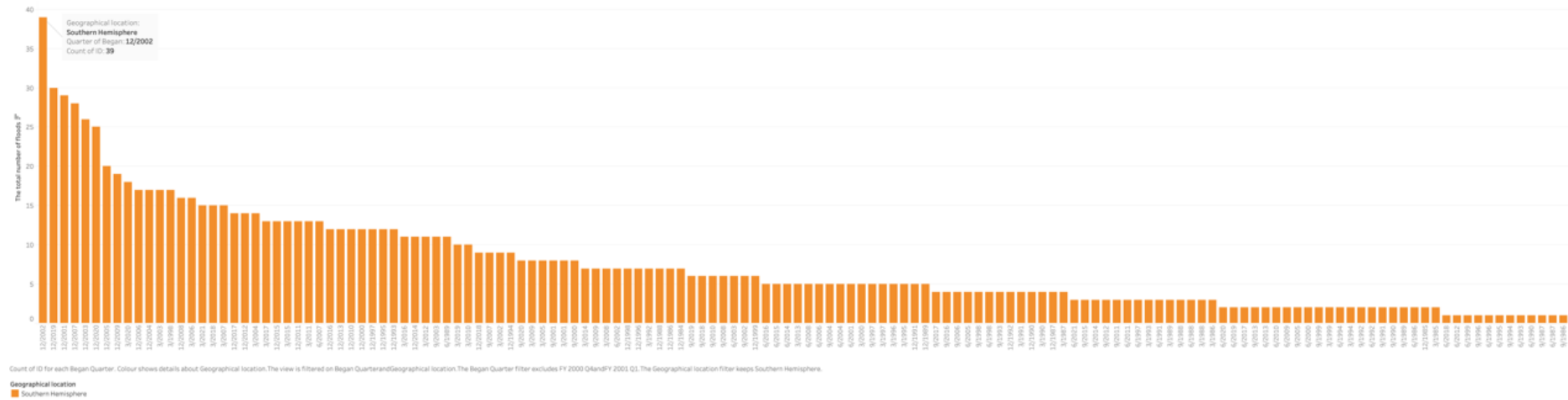


Figure 3.2.1.3. Statistics on the total number of floods in southern hemisphere per season

As a result, the number of global floods is strongly influenced by the seasons. Similarly, the high period of flooding in both northern and southern hemispheres is summer, and the low period is winter and spring. We can roughly conclude that the occurrence of floods is more closely related to the seasons and to geographical location less.



### 3.2.2 Analysis of flood data of different severity class

Based on the conclusions of last sub-section, the following analysis is based on data from the Northern Hemisphere for greater clarity.

The area graph [Figure 3.2.2.1] shows the trend of the total number of floods occurring in the Northern Hemisphere for different severity classes over time. The light blue area represents class 1, which maintains a continuous fluctuation from 1985 to 1998, then rapid decline before growing sharply. It peaks in the period June to August 2002 with a total of 86. Then it starts to decrease significantly from 2007 and keeps fluctuating in the range of 5 to 20 between 2010 and 2021.

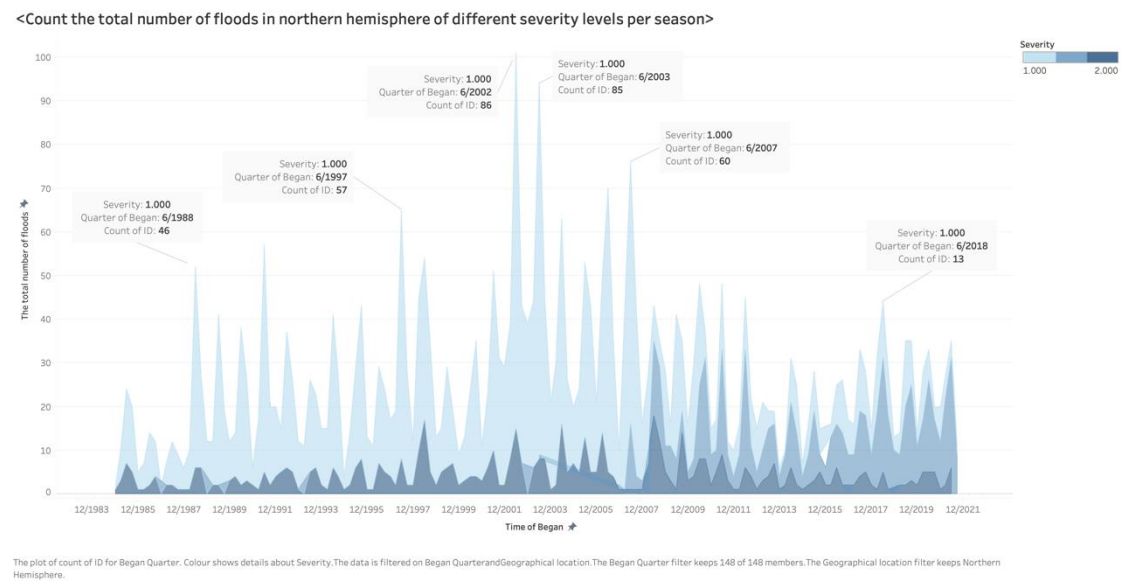


Figure 3.2.2.1. The total number of floods in Northern Hemisphere of different severity levels per season

The data of class 2 which is shown as the dark blue line is fluctuating steadily in the 0-18 range, reaching a maximum of 18 in June to August 2008 [Figure 3.2.2.2].

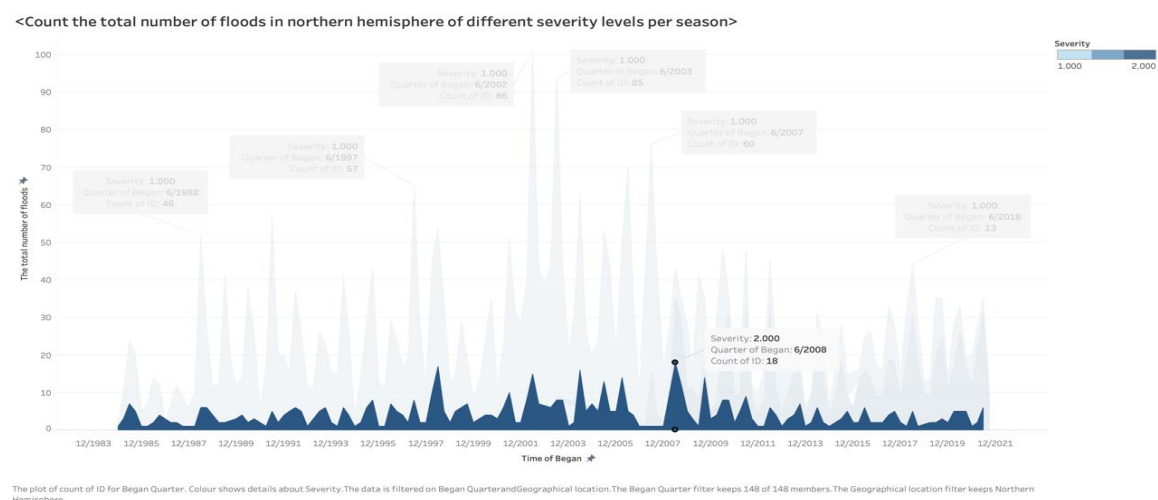


Figure 3.2.2.2. The total number of floods in Northern Hemisphere of severity class 1 per season

The severity class 1.5 is first shown in December 2003, then it maintains a fluctuating rise and generally exceeds class 1 after June to August 2008, reaching a peak of 27 in June to August 2012 [Figure 3.2.2.3].

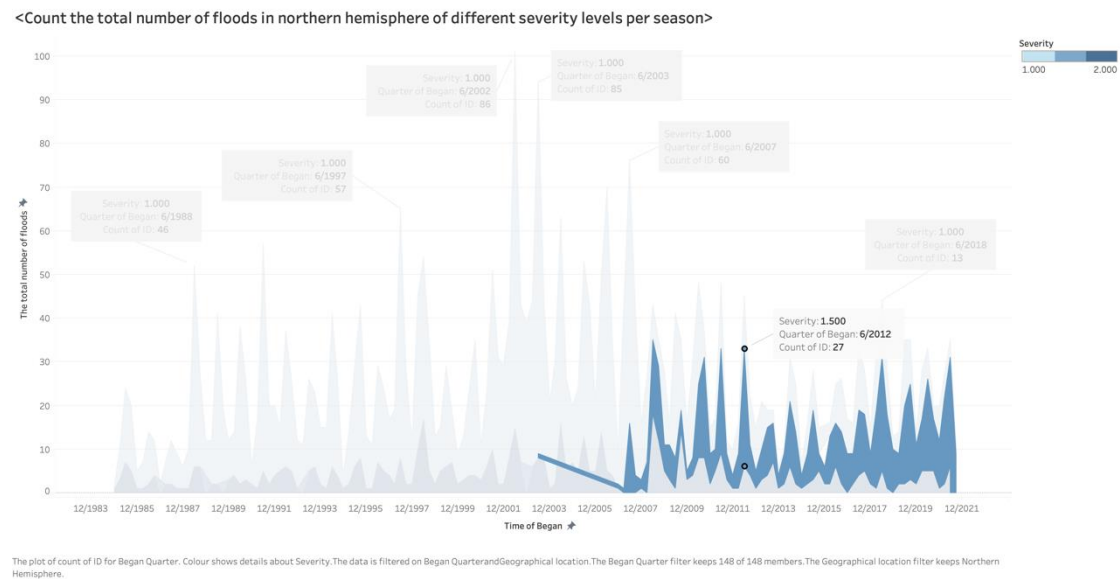


Figure 3.2.2.3. The total number of floods in Northern Hemisphere of severity class 1 per season

In conclusion, the severity class 1 is much higher than the other two classes until 2007, but slightly lower than class 1.5 in 2008 until 2021. The severity class 2 has been fluctuating steadily and remains in the lowest rank.

### 3.2.3 Analysis of the trend of Main Cause data over seasons

The following line chart [Figure 3.2.3.1] shows data on the six most common major causes of flooding in the Northern Hemisphere. The Heavy Rain line is consistently higher than the other causes, reaching a peak of 72 in the period June to August 2002 and remaining well above the other causes until 2007. Then it starts to decline significantly, being similar to the other lines but slightly above them overall between 2008 and 2021.

In addition, a very regular pattern can be found in the timing of Heavy rain, with peaks generally occurring in the June to August time frame, which is the northern hemisphere summer.

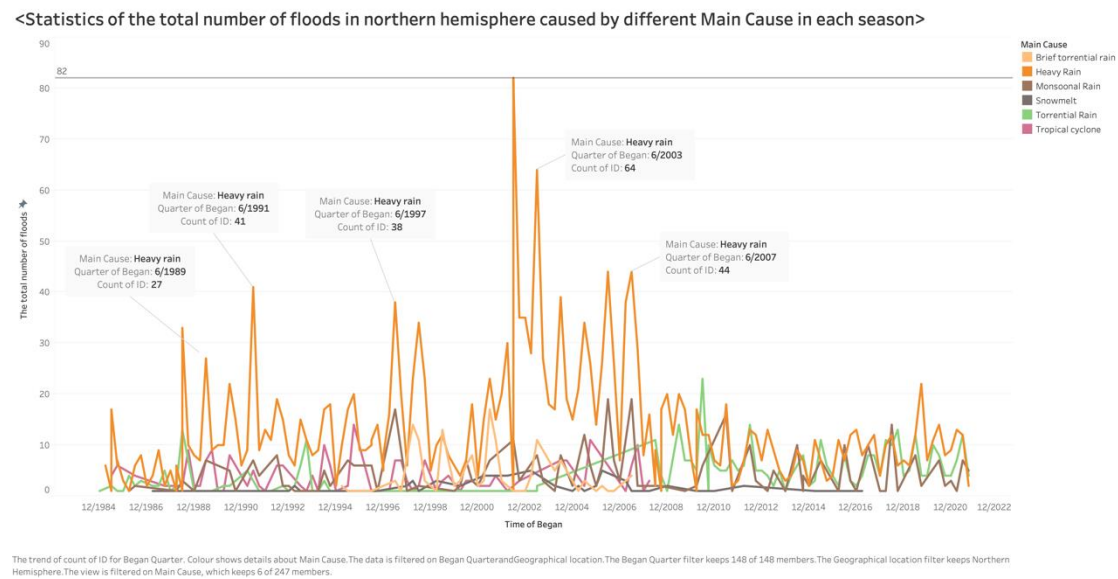


Figure 3.2.3.1. Statistics of the total number of floods in Northern Hemisphere caused by different Main Cause in each season

In addition to heavy rain, monsoonal rain and torrential rain also cause multiple floods individually. The trend of monsoonal rain fluctuations is regular and continuous, peaking in the summer of 2006 and 2007 with data of 19 shown in [Figure 3.2.3.2]. In contrast, torrential rain data remained in a low range after the initial fluctuations during 1995-2008, after which it began to rise rapidly and remained fluctuating for the next decade. In the summer of 2010, the number of torrential rains surpassed heavy rain for the first time and reached its highest point shown in [Figure 3.2.3.3].

Other causes generally remained in the lower numerical range, occurring more sporadically across time.

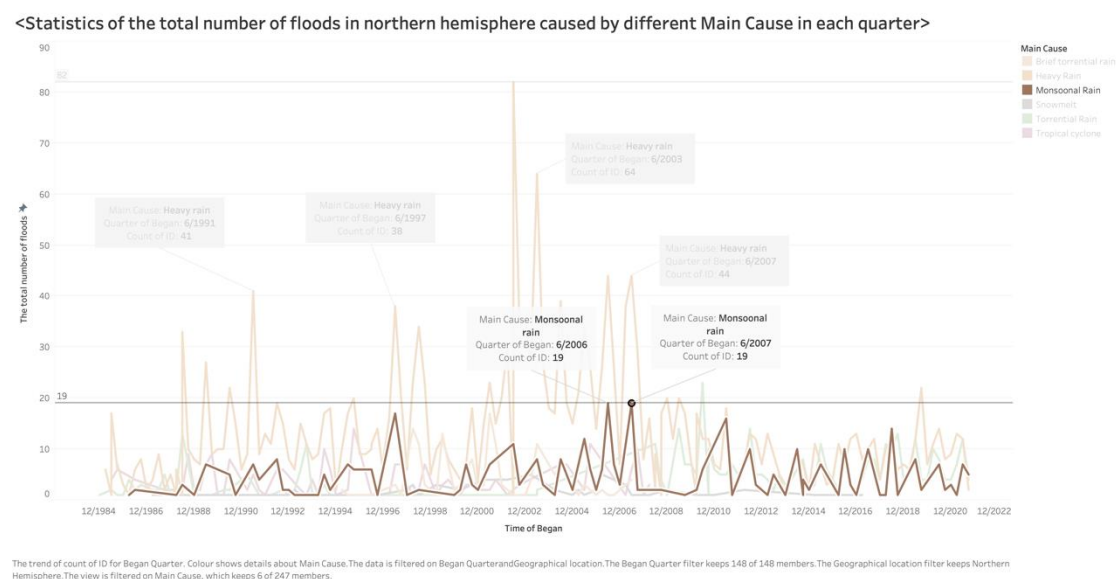


Figure 3.2.3.2. The trend of data of Monsoonal Rain in Northern hemisphere over season

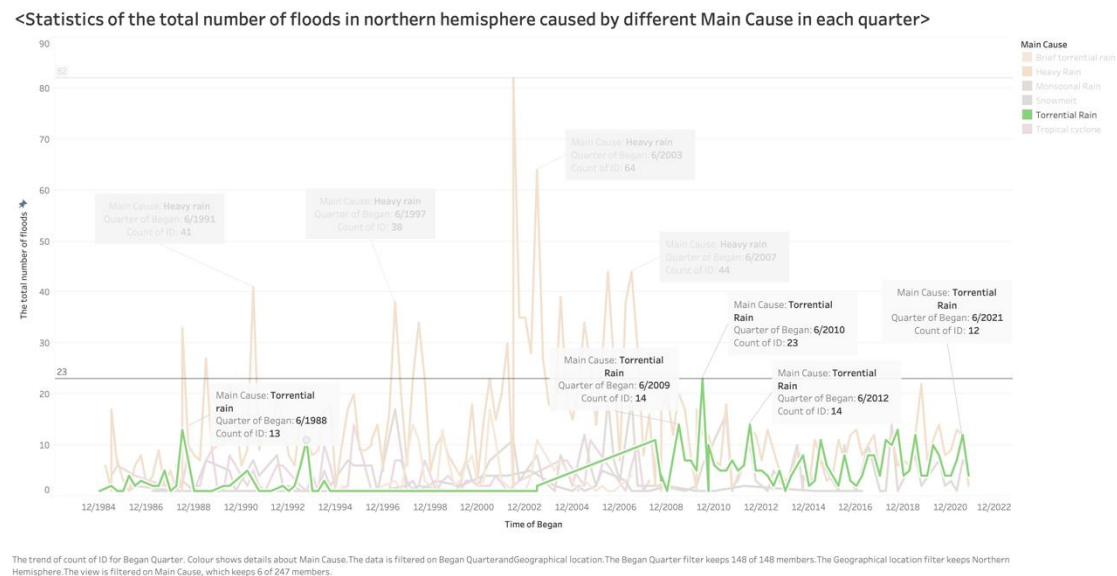


Figure 3.2.3.3. The trend of data of Torrential Rain in Northern hemisphere over season

Overall, the most common cause of flooding is Heavy Rain, which occurs much more often than other items. It is followed by Monsoonal Rain and Torrential Rain, especially Torrential rain has a very significant increase in the 2009 to 2021 period. And the high occurrence period of these common causes is all summer.

### 3.2.4 Comprehensive analysis of section 3.2

In a comprehensive analysis, the number of flood occurrences increases and then decreases with time progression, with the most frequent occurrence in the northern hemisphere from June to August 2002 and in the southern hemisphere from December 2002 to February 2003, and then leveling off in the last decade.

The occurrence of global floods is strongly influenced by the seasons, with the most likely period in the Northern Hemisphere being from June to August and less likely to occur from December to May. In the Southern Hemisphere, the high occurrence period is from December to February, while June to November is infrequent. Heavy rain is the most common cause of flooding, and the corresponding flood level is usually Class 1.

However, it is also worth noting that after 2008, the number of floods caused by torrential rain increased significantly, and it became one of the three main causes of floods along with monsoonal rain and heavy rain. Correspondingly, floods of Class 1.5 started to occur frequently, similar to the number of Class 1. The number of extreme events has always been less. Thus, it can be observed that heavy rain usually leads to large flood events, while torrential rain may be the main cause of very large flood events.

#### The following patterns of flooding can be derived:

- The total number of flood occurrences increases first with time development peaking in the summer of 2002 and then decreasing, leveling off in the last decade.
- There is a clear correlation between the occurrence of floods and the seasons.

Summer has been the period of high flooding, while winter and spring have a lower probability of occurrence.

- Before 2008, the main cause of flooding was Heavy Rain, which usually led to large flood events. After 2008, Torrential Rain and Monsoonal Rain have also become common factors leading to flooding, and Torrential Rain may be associated with very large flood events.

## 4. Analysis of the damage and impact of floods

### 4.1 Impact of flood change over time

This section specifically looked into the damage and impact of floods worldwide. To evaluate the damage and impact of floods over time, we adopted the **dead number and displaced number** as the measure value because they are directly related to human lives.

Here is our key finding: Though the flood magnitude trend varies, the dead and the displaced number drop yearly except for several flood outliers. We assume this is due to the development of technology.

In order to get the general trend, we need to pre-process the data. From the original data, we can see several significant catastrophes where the number of dead far exceeds any other floods. For example, the Thailand flood in 2004 took the lives of 160,000 people. This number exceeds any other total number for a year from 1985 to 2021. The distribution of the dead value is shown in [Table 4.1]. Therefore, we define a major catastrophe as a flood with 10,000 deaths or more. We took the catastrophes out for a separate study in the next section and used the median of the remaining data to look into the general trends.

Table 4.1 Distribution of fatalities

<b>Fatalities</b>	<b>&lt; 10,000</b>	<b>10,000-50,000</b>	<b>50,000-100,000</b>	<b>Over 100,000</b>
<b>Occurrences</b>	4971	3	1	2
<b>Total</b>	4977			

As can be seen in the following line chart [Figure 4.1.1], the flood magnitude level remained stable from 1985 to 2021, the number of the dead and the displaced had an overall decreasing trend. With the rising number of the global population, there are less lives lost from the floods. It might sound counterintuitive at first, but we then find that it may be caused by the improvement of technology, especially in the field of flood forecasting.

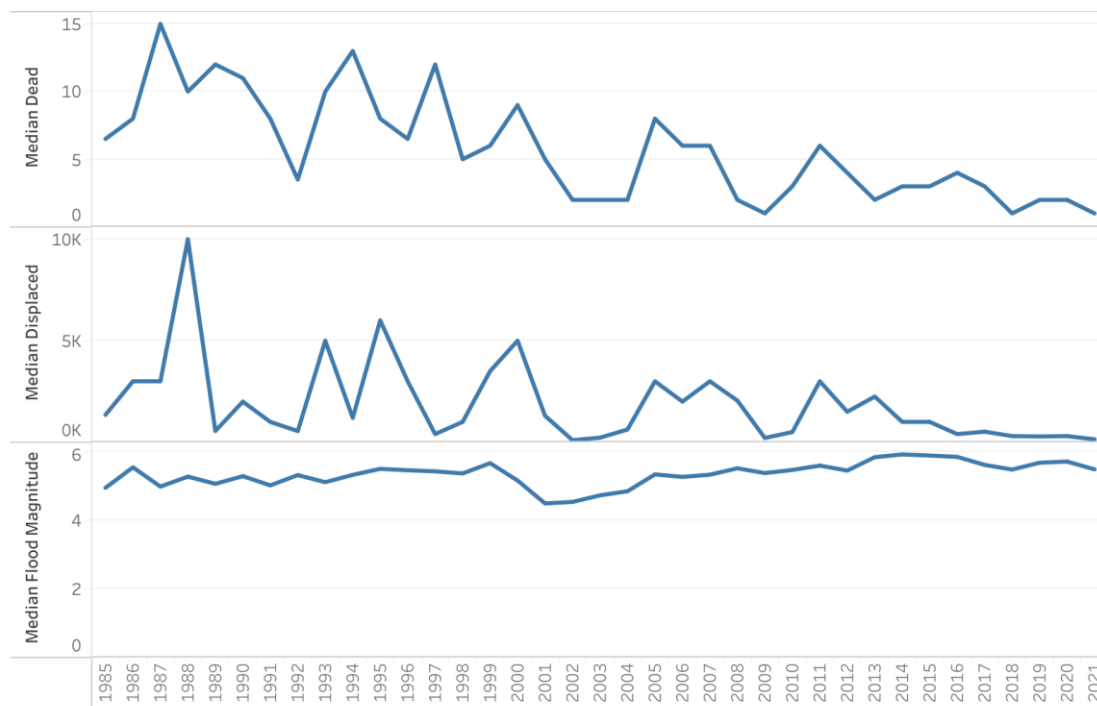


Figure 4.1 Median dead/displaced/flood magnitude sum between 1985-2021

With the speedy improvement of information technology, flood prediction now is equipped with many advanced tools for better accuracy. The study from Yu suggests that a high improvement of prediction accuracy could be achieved by improved radar prediction methods (Yu et al., 2015). And in the machine learning field, different flood prediction models have been developed to contribute to policy suggestions and minimization of the loss of human life (Mosavi et al., 2018). Therefore, though the flood magnitude level varies over the years, humans become more prepared and able to handle these natural disasters to lower their damage.

## 4.2 Countries and areas most affected

In this section, we will discuss two categories of countries and areas which had been affected by the floods. The first type is the historically significant catastrophe which took the lives of many in one disaster. The second is the general countries or areas which suffer the most floods per year on average.

- ***Countries and areas with most damage in one flood***

As shown in the previous section, we define a major catastrophe as a flood with 10,000 deaths or more. [Table 4.2.1] shows the major catastrophes from 1985 to 2021.

From this table, we can see that the majority of the major catastrophes took place in Southeast Asia. located right next to the Pacific Ocean. The main causes are torrential rain, tropical cyclones and tsunamis.

Table 4.2.1 Major Catastrophe

Country	Year of Began	Dead	Displaced	Area	Main Cause
Thailand	2004	160,000	5,000,000	Southeast Asia	Tidal surge
Bangladesh	1991	138,000	10,000,000	Southeast Asia	Tropical cyclone
Venezuela	1999	20,006	400,000	South America	Brief torrential rain
Japan	2011	10,000	200,000	East Asia	Tsunami
Myanma	2008	10,000	1,500,000	Southeast Asia	Tropical cyclone

- **Countries and areas with most accumulated damage**

As mentioned above, we used two metrics to evaluate the damage of flood to human society. The first is the dead number from the flood, which shows the life threat posed by the disaster. The second is the displaced number, which shows how much the flood impacted the local infrastructure.

Sum of Death From 1985-2021

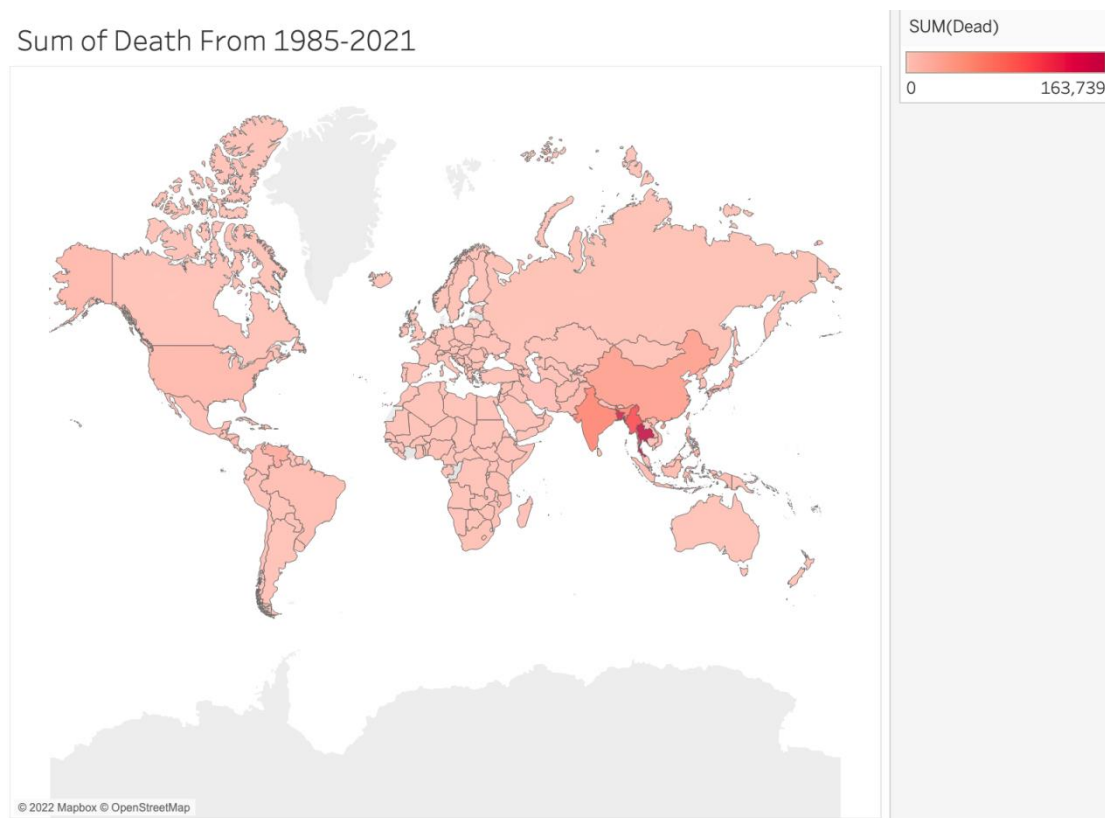


Fig 4.2.1 Sum of Death from 1985-2021

From the world map [Figure 4.2.1], it can be seen that the biggest life threat is located in South Asia. There are also visible areas such as the Caribbean area. The top five countries with most death number are also shown in the bar chart [Figure 4.2.2].



### Sum of Death From 1985-2021

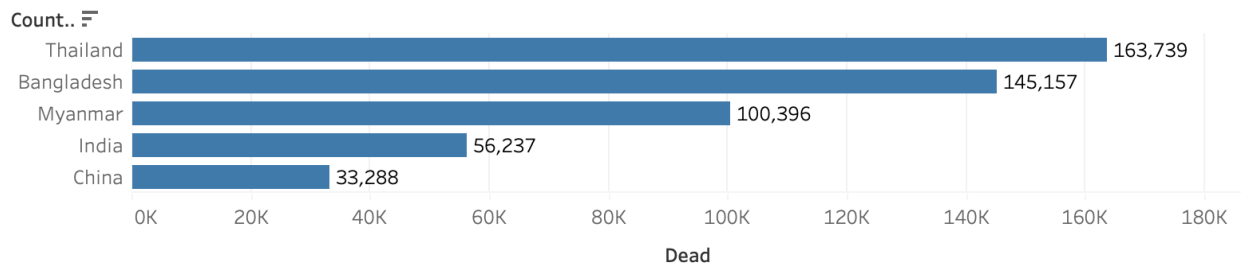


Figure 4.2.2 Top 5 countries for the sum of death from 1985-2021

From [Figure 4.2.3], we can see that the floods which affected the local infrastructure most are in Asia as well, especially in China and India. The top five countries with most displaced number are shown in [Figure 4.2.4].

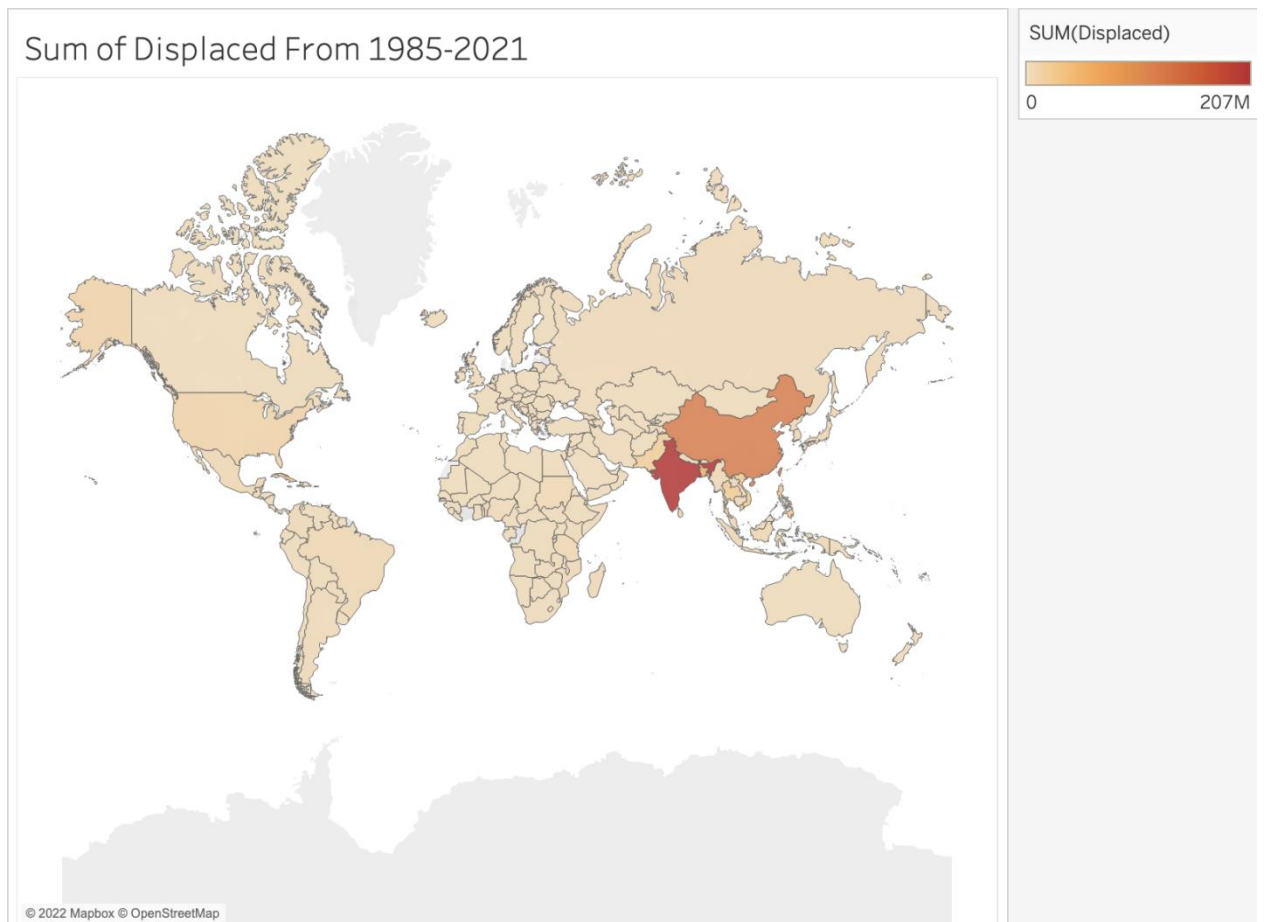


Figure 4.2.3 Sum of Displaced from 1985-2021



## Sum of Displaced From 1985-2021

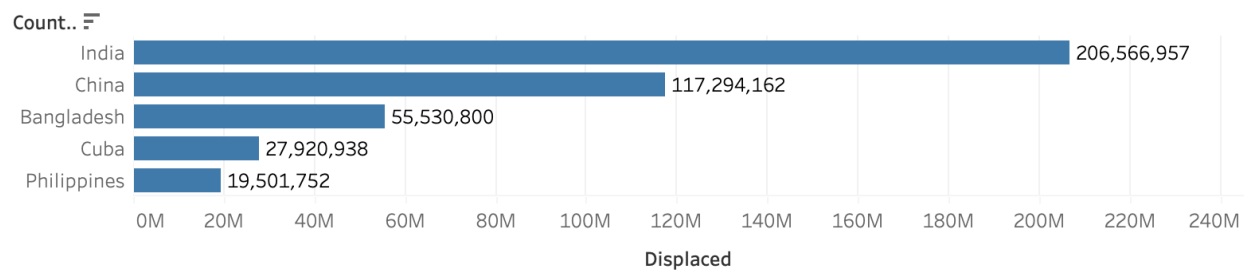


Figure 4.2.4 Top 5 countries for the sum of displaced from 1985-2021

## 5. Further analysis of floods driving factors

### 5.1 Related work and assumptions

It has been discussed the magnitude metric, patterns and trends and the influence of floods on humans worldwide according to previous analysis, which benefits subsequent study on driving factors of floods. To figure out the potential factors and their relationship with floods occurrence, literature review related to inducement of floods.

According to World Health Organization's classification for floods, it can be concluded as the following three types of floods:

1. Flash floods caused by rapid and excessive rainfall
2. River floods caused by consistent rain and snow melting
3. Coastal floods caused by storm surges (tropical cyclones and tsunami)

About 80% of floods happened in the recent ten years attributed to climate change and the frequency and intensity of extreme precipitation is expected to continue to increase. (WHO, 2021) What's more, it can be found that CO<sub>2</sub> emission is closely related to climate change. James M. Lenihan's simulation of temperatures in the western and eastern United States shows that CO<sub>2</sub> emissions have a significant impact on moisture and temperature in each region, which creates great uncertainty about the future trends of the American ecosystem (Lenihan et al., 2008). Therefore, we made an assumption that floods are indirectly associated with CO<sub>2</sub> emission.

Besides, according to Gregory J. Retallack's analysis on the relationship between floods occurrence and CO<sub>2</sub>, it can be inferred that forest areas also possibly impact the happening of floods because more water flows off in streams and floods when plants transpire less. He quantified this effect using high-resolution data of changing stomata density and size in a mesic tree, Ginkgo, from 1754. The measured decrease in maximum potential transpiration matches with rising water levels in the Mississippi River and reflects a 29 percent decrease in potential transpiration from 1829 to 2015, leading to rising atmospheric CO<sub>2</sub> levels and falling transpiration encouraging floods. (Retallack and Conde, n.d.) Thus, another assumption is about the relationship between floods and the variation of forest area of each country.

## 5.2 Relevant data obtaining and cleaning up

### 5.2.1 Forest area and CO2 emission data from Worldbank.org

To conduct quantitative analysis, we obtained forest area and CO2 emission data from [worldbank.org](https://worldbank.org). Parts of two datasets are shown below.










Country	Most Recent Year	Most Recent Value	
Afghanistan	2018	7,440	
Albania	2018	5,560	
Algeria	2018	151,670	
American Samoa			
Andorra	2018	460	
Angola	2018	27,340	
Antigua and Barbuda	2018	530	
Argentina	2018	177,410	
Armenia	2018	5,550	
Aruba			
Australia	2018	386,620	

Figure 5.2.1.1. CO2 emission data











Country	Most Recent Year	Most Recent Value	
Afghanistan	2020	1.9	
Albania	2020	28.8	
Algeria	2020	0.8	
American Samoa	2020	85.7	
Andorra	2020	34.0	
Angola	2020	53.4	
Antigua and Barbuda	2020	18.5	
Argentina	2020	10.4	
Armenia	2020	11.5	
Aruba	2020	2.3	
Australia	2020	17.4	

Figure 5.2.1.2. Forest area data

The CO2 emission data displays how CO2 emissions have changed over time for each country; The Forest area data shows how each country's forest area has changed over time as a percentage of its total land area.

## 5.2.2 Data pre-processing using Python

There are several issues with raw data, which will have an influence on the accuracy and precision of analysis. For example, the existence of several blank units or rows and data irrelevant with the FloodArchive.csv file need to be abandoned. The following image shows the irrelevant data and blank rows in the raw data.

West Bank and Gaza	PSE	CO2 emissions (kt)	EN.ATM.CO2E.KT						
Pacific island small states	PSS	CO2 emissions (kt)	EN.ATM.CO2E.KT						
Post-demographic dividend	PST	CO2 emissions (kt)	EN.ATM.CO2E.KT						
French Polynesia	PYF	CO2 emissions (kt)	EN.ATM.CO2E.KT						

Figure 5.2.2.1. Blank rows

Lower middle income	LMC	Forest area (% of land area)	AG.LND.FRST.ZS	29.3908837469578
Low & middle income	LMY	Forest area (% of land area)	AG.LND.FRST.ZS	34.3298170338299

Figure 5.2.2.2. Irrelevant data

Besides, the time dimensions of two datasets are different - records in forest area dataset are from 1990 to 2020 while in CO2 emission are from 1990 to 2018. Therefore, we decided to conduct analysis on specific duration between 1990 and 2018, ignoring 2-year extra data provided by forest area data.

What's more, to gain an intuitive overview of the trends and variation of different factors, we used normalization to scale data at different measures to the range from 0 to 1, mitigating the neglect of delicate changes caused by large values from other attributes.

## 5.3 Analysis of different cases

Because of the enormous capacity of flood case data, several instances of different countries with high frequency and intensity of floods (the results are obtained from previous analysis in section 2 to 4) are selected to represent our approximation for dependency analysis. We used Python to put data from three different sources together and generate several line charts, which respectively illustrate the relationship of two assumed driving factors and floods occurrence in each top five countries with the highest frequency and intensity. In each chart, blue line represents the total amount of floods happened in each year in each country; orange line means the variation of forest area percentage; green line shows the transition of the capacity(kt) of CO2 emission each year in each country.

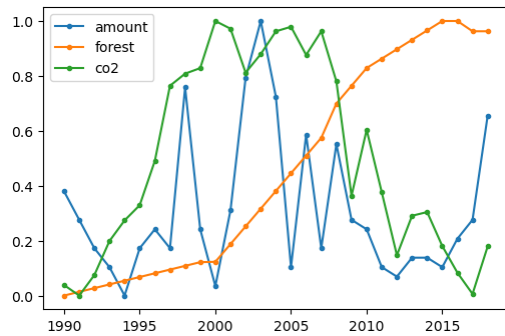


Figure 5.3.1. United States

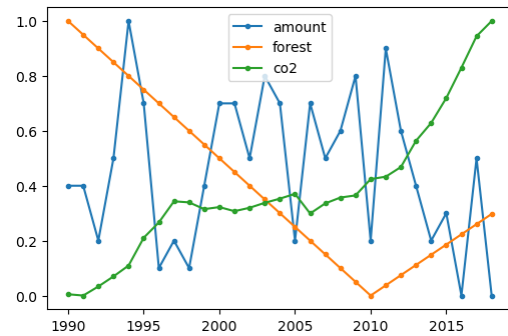


Figure 5.3.2. Philippines

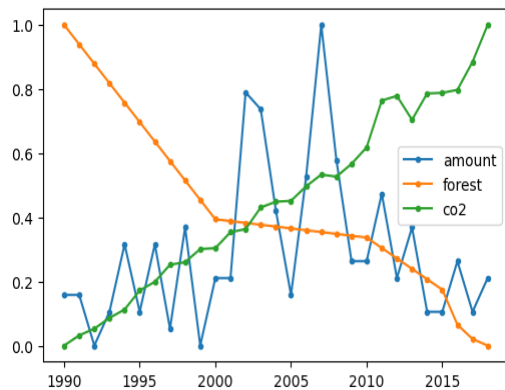


Figure 5.3.3. Indonesia

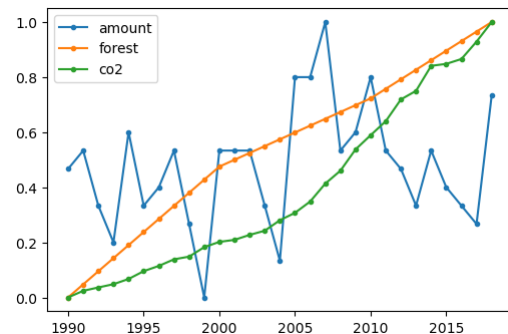


Figure 5.3.4. India

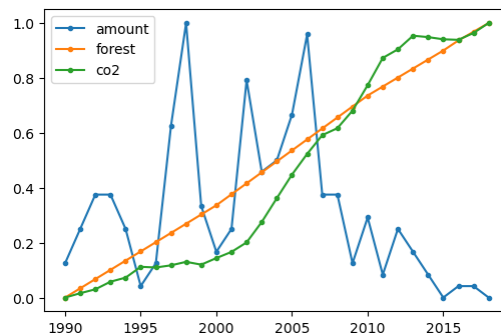


Figure 5.3.5. China

According to the line charts above, it can be found that the relationship between different factors and flood occurrence varies from country to country, but it shows a relatively slight positive correlation between CO<sub>2</sub> emission and the frequency of flood happening in the United States. The number of floods that happened in 2003 reached the peak while CO<sub>2</sub> emission was also at the summit in the United States. After 2003, the frequency of floods was descending, and CO<sub>2</sub> emission had also been controlled to a specific level.

The reason behind this correlation may be the temperature rising brought by increasing CO<sub>2</sub> emission. What's more, it can also be seen that there is no such a positive correlation in other countries. Several literatures have been found that various precautionary measures have been taken to mitigate or avoid floods damage. For example, Three Gorges Dam has been built since 1994 to control water level in the

middle of China where the Yangtze River flows through. Besides, we also found that floods happened between 2001-2018 in Indonesia were mainly caused by forest area loss. The analysis of Global Forest Watch (GFW) indicates the loss of 887 hectares of tree cover on Cyclops Mountains, Papua, between 2001 and 2018, resulting in the flooding of Waibu, Sentani and East Sentani Districts. (wri-indonesia.org, n.d.)

To sum up, plenty of reports have illustrated the tacit relationship between CO<sub>2</sub> emission and forest area. CO<sub>2</sub> emission usually acts as a stimulated factor for flood, while more forest areas often protect humans from suffering from floods.

## **Conclusions**

This report has thoroughly introduced and analysed the impact of flood from different aspects in the above sections including the flood magnitude, the pattern of flooding consistent with seasons, dead and displaced number of people and other driving factors that trigger the flood.

In terms of section 2, the flood magnitude has been calculated in proportion to the definition of it. After that, level of magnitude has been defined for the further analysis, which involves number of floods, average magnitude and magnitude levels. Visualization have been made to each evaluation relevant to the top 5 countries.

In terms of section 3, it can be found that the pattern of flooding is closely related to the seasons. Floods are more frequent in summer and reversed in winter and spring. The number of floods peaked in 2002 and then declined and levelled off in the last decade. Before 2008, the main cause of flooding was Heavy rain, while after 2008, Torrential rain and Monsoonal rain have also become common causes. Comparing the trends of Main Cause and Severity, Heavy rain usually leads to large flood events, while very large flood events may be related to Torrential rain.

In section 4, we adopted the dead number and the displaced number to evaluate the damage and effect of one flood. Then we found that though the flood magnitude varies over time, the damage and impact of the flood on human society decreased from 1985 to 2021 year by year. We then identified the countries which suffered the most damage both at one time and accumulatively.

In terms of section 5, several analyses related to the relationship between different factors and flood occurrence have been conducted. According to quantitative analysis and literature review, it can be found out that the frequency and magnitude of flood occurrence has an indirect correlation with climate change on account of different CO<sub>2</sub> emission and forest area.

## References (APA 7<sup>th</sup> Edition)

*Australian weather and the seasons* - *australia.gov.au*. Web.archive.org. (2008). Retrieved from <https://web.archive.org/web/20121021091448/http://australia.gov.au/about-australia/australian-story/austn-weather-and-the-seasons>.

*Internally displaced persons, new displacement associated with disasters (number of cases) - India*. Data. (n.d.). Retrieved from <https://data.worldbank.org/indicator/VC.IDP.NWDS?locations=IN>

Lenihan, J.M., Bachelet, D., Neilson, R.P. and Drapek, R. (2008). *Simulated response of conterminous United States ecosystems to climate change at different levels of fire suppression, CO2 emission rate, and growth response to CO2*. *Global and Planetary Change*, 64(1-2), pp.16–25. doi:10.1016/j.gloplacha.2008.01.006.

Mosavi, A., Ozturk, P., & Chau, K.-wing. (2018). *Flood prediction using Machine Learning Models: Literature review*. *Water*, 10(11), 1536. <https://doi.org/10.3390/w10111536>

Retallack, G. and Conde, G. (n.d.). *Flooding Induced by Rising Atmospheric Carbon Dioxide Fiscal Year 2020 Annual Report*. [online] doi:10.1130/GSATG427.1.

WHO (2021). *Floods*. [online] [www.who.int](http://www.who.int). Available at: [https://www.who.int/health-topics/floods#tab=tab\\_1](https://www.who.int/health-topics/floods#tab=tab_1)

wri-indonesia.org. (n.d.). *Main Causes of Floods in Indonesia and How to Prevent Them* | WRI Indonesia. [online] Available at: <https://wri-indonesia.org/en/blog/3-main-causes-floods-indonesia-and-how-prevent-them>.

Yu, W., Nakakita, E., Kim, S., & Yamaguchi, K. (2015). *Improvement of rainfall and flood forecasts by blending ensemble NWP rainfall with radar prediction considering orographic rainfall*. *Journal of Hydrology*, 531, 494–507. <https://doi.org/10.1016/j.jhydrol.2015.04.055>