

# Prediction of Precipitation Rate Based on Stationary Extreme Value Theory

**Justin Han**

The International School of Kuala Lumpur, Kuala Lumpur, Malaysia

**Email address:**

[justin.jaewoong.han@gmail.com](mailto:justin.jaewoong.han@gmail.com)

**To cite this article:**

Justin Han. Prediction of Precipitation Rate Based on Stationary Extreme Value Theory. *American Journal of Applied Mathematics*. Vol. 9, No. 5, 2021, pp. 186-191. doi: 10.11648/j.ajam.20210905.13

**Received:** September 23, 2021; **Accepted:** October 25, 2021; **Published:** October 30, 2021

---

**Abstract:** For the determination of the effectiveness of weather forecasts such as temperature or flooding forecast, the single variable linear regression or simple exponential smoothing will not be an effective way. To accurately predict the effectiveness of such a trend, iterative and statistical methods that can determine the status of temperature or flooding in the United States were chosen in this paper. The task of modeling the pattern in a focused period and performing data analysis was performed. For the analysis, the extreme value theory was used to assess extreme events within probability distributions by quantifying tail behavior. The python tools are utilized for the analysis of big data. By analyzing the maximum values of samples, it was possible to determine probabilities for extreme events. A comparison was made with events previously observed and analyzed for authenticity. As evident in our observations, lower values of data have much shorter return periods. In other words, they are more likely to reoccur; however, as the values increase for higher precipitation values, the length of the return periods increase exponentially. Therefore, there is a tendency for precipitation values to remain in lower ranges. In this paper, the USGS geographical information and Gumbel distribution were used to find the return period corresponding to the exceedance probability. The Gumbel distribution is applied to Allegheny River, New York and Whetstone River, San Diego.

**Keywords:** The Extreme Value Theory, Flooding Forecast, Precipitation

---

## 1. Introduction

### 1.1. Background

Injuries and deaths caused by natural disasters take up a significantly high proportion of news each day. Natural disasters involve the interaction between hazardous natural factors that are extraneous to man and any population. These hazards are violent and calamitous phenomena, which include floods, tornadoes, tsunamis, earthquakes, fires, volcano eruptions, and hurricanes. Research shows that the number of natural disasters around the world in the past decades has significantly risen [7-10]. The impact of these natural phenomena is catastrophic: buildings collapse, and many deaths follow. There are additional, long-lasting effects that exist after the disaster as well, such as famine, economic losses, several health risks, and emotional aftershocks.

Floods are the world's most common natural disaster, causing the majority of natural disaster fatalities. Floods are identified with a large, abnormal amount of water overflowing

over a normally dry land. They follow a normal law of nature and happen accordingly with seasons.

The destructiveness of floods and the severe winds that accompany the floods are mainly due to the mechanical force of the water and the debris it carries, along with the contamination and wetness of the flood water. Floods happen regularly in the U.S. following heavy precipitation in the spring or melting of snow and ice in the mountains and cause a high number of deaths and property damage. Most of the deaths caused by floods are related to the effects of flash floods, which happen when there is a fast runoff of water from heavy precipitation in a short amount of time [1, 2].

To minimize the damage caused by these natural disasters, it is important to prepare for these events. The predictability of catastrophic natural disasters is crucial but this task of modeling natural disasters in a focused area and performing data analysis is difficult especially when the data pattern is complex. Extracting patterns from the analytical process to characterize the natural disaster is a difficult job for the researchers.

Using the Gumbel distribution method, this research

focuses on the predictability of floods, for they account for a large amount of casualties out of all the natural disasters. By using statistical and computational simulations including a regression analysis using least square method, we investigated the exceedance probability, which is the probability that the event will exceed some critical value. Also the return period corresponding to this exceedance probability was found.

### 1.2. Extreme Value Theory

The risk of extreme, rare events, such as the 1755 Lisbon earthquake, is modeled using extreme value theory. Extreme value theory (EVT), also known as extreme value analysis (EVA), is a statistical branch that investigates deviations from the median of probability distributions. It aims to determine the likelihood of events that are more extreme than any previously observed from an ordered sample of a random variable.

Many disciplines, including structural engineering, finance, earth sciences, traffic prediction, and geological engineering, use extreme value analysis. For example, in the field of hydrology, EVA could be used to estimate the likelihood of a large flooding event, such as a 100-year flood. A coastal engineer would seek to estimate the 50-year wave and design the structure accordingly when designing a breakwater.

## 2. Procedures

The precipitation frequency curve was plotted using the Gumbel Distribution. [6]

1. The Gumbel distribution is used to characterize the maximum and lowest values of a set of samples drawn from a variety of distributions. [3]
2. The exceedance probability is the probability that an event will exceed a critical value (usually far from the mean).
3. A return period is a forecast of the likelihood of an event occurring. [4] Based on historical data, a statistical estimate of the average recurrence interval.
4. Theoretical probability is the number of times we expect an event to occur if we repeat the same experiment over and over (i.e., flipping a coin and getting heads or tails is each 0.50).
5. As the number of trials increases, the estimated probability approaches the theoretical probability. It's a theoretical probability approximation.

### 2.1. Conceptual Outcomes

This research demonstrates the understanding of return periods and frequency analysis of the natural disaster. We demonstrate the statistical parameters used in the frequency analysis.

### 2.2. Practical Outcomes

We utilized Python/Pandas and Plotly to make a statistical and computational analysis. We used statistical distribution methods and regressions to create a frequency curve.

### 2.3. Computing/Data Outputs

Numerical values: Theoretical and estimated return periods.

Graphical: graph based on the real data, regression, and forecast curves.

### 2.4. Data Collection - Real-time/Daily Data

To download data, the National Water Information System (NWIS) web interface was used.

## 3. Data and Results

Using the Gumbel Distribution to plot the Flood Frequency Curve:

1. The probability that an event will surpass a critical value is known as the exceedance probability (usually far from the mean).
2. A return period is a prediction of how likely an event will occur. The average recurrence interval is a statistical measurement based on historical data.
3. The number of times we expect an event to occur if we repeat the same experiment over and over is known as theoretical probability (i.e. flipping a coin and getting heads or tails is 50%).
4. As the number of trials increases, the estimated probability approaches the theoretical probability. It's a theoretical probability approximation.

### 3.1. Allegheny River at Salamanca NY

In this paper, to find the return period corresponding to the exceedance probability, Gumbel distribution is applied to the Allegheny River in Salamanda, NY and Whetstone River near Big Stone City, SD. The USGS number and geographical information are shown as follows [3, 5]:

USGS 03011020 ALLEGHENY RIVER AT SALAMANCA NY (Duration: 1904-2016)  
Cattaraugus County, New York  
Hydrologic Unit Code 05010001  
Latitude 42°09'23", Longitude 78°42'55" NAD83  
Drainage area 1,608 square miles  
Contributing drainage area 1,608 square miles  
Gage datum 1,358 feet above NGVD29  
USGS 05291000 WHETSTONE RIVER NEAR BIG STONE CITY, SD  
Grant County, South Dakota  
Hydrologic Unit Code 07020001  
Latitude 45°17'30", Longitude 96°29'14" NAD27  
Drainage area 398 square miles  
Gage datum 996.96 feet above NGVD29.

### 3.2. Procedures

Read raw data through Python and Pandas. Create a column 'Rank (i)' in descending order.

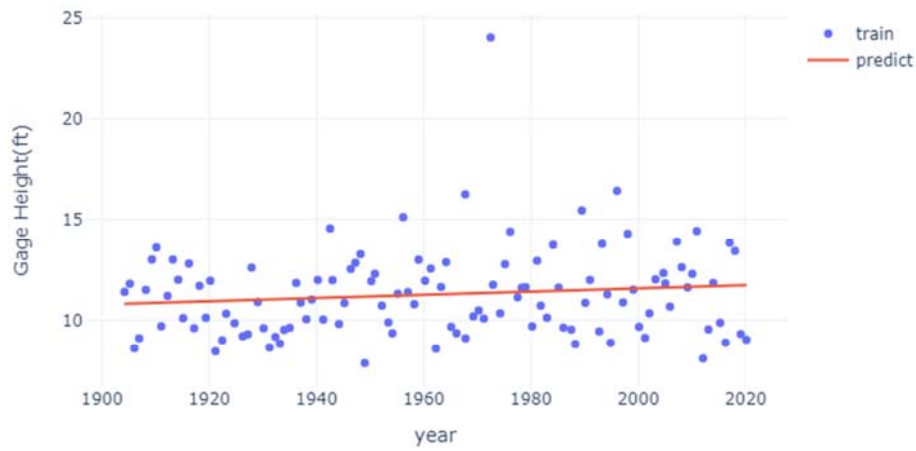
Create a fourth column called  $q_i$ . Gringorten's plotting position formula will be used to calculate the estimated exceedance probabilities relevant to past observations.

Make another column and label it  $p_i$ . Then make it equal to

$$1 - q_i \cdot p_i$$

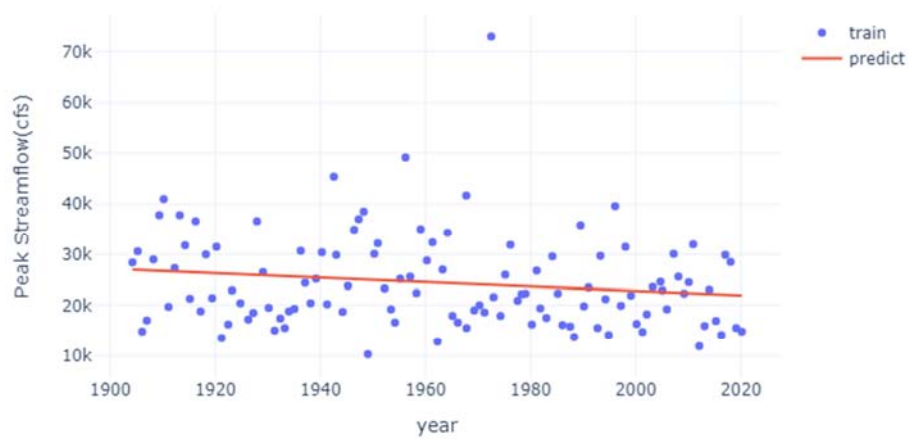
Refers to the non-exceedance probability.

Gage Height(ft) vs Year



(a) ALLEGHENY RIVER AT SALAMANCA NY

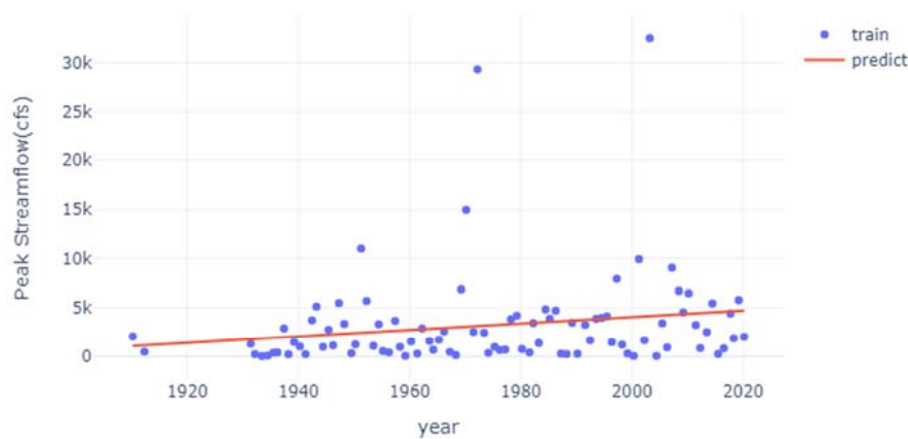
Streamflow(cfs) vs Year



(b) WHETSTONE RIVER NEAR BIG STONE CITY

**Figure 1.** Streamflow (cfs) of the Allegheny river in NY (1904-2016) – with pandas and plotly.

Streamflow(cfs) vs Year



(a) ALLEGHENY RIVER AT SALAMANCA NY

	Time	Peak Streamflow(cfs)	Gage Height(ft)	Rank(i)	qi	pi
0	1904-03-26	28400	11.40	37.0	0.314847	0.685153
1	1905-03-20	30600	11.80	25.0	0.211505	0.788495
2	1906-01-24	14800	8.60	107.0	0.917671	0.082329
3	1906-12-15	17000	9.10	92.0	0.788495	0.211505
4	1908-03-16	29000	11.50	34.0	0.289011	0.710989
...	...	...	...	...	...	...
112	2016-04-12	14100	8.88	111.0	0.952118	0.047882
113	2017-01-13	29900	13.83	31.0	0.263176	0.736824
114	2018-01-13	28500	13.43	36.0	0.306235	0.693765
115	2019-02-08	15500	9.31	105.0	0.900448	0.099552
116	2020-02-27	14800	9.03	108.0	0.926283	0.073717

116 rows × 6 columns

(b) WHETSTONE RIVER NEAR BIG STONE CITY

*Figure 2. Exceedance probability associated with a specific observation.*

### 3.3. Statistical Definition of Return Period

Generate a new column called 'T p estimated,' and use the equation for the return time to evaluate the values in p i. The estimated distribution of 35 years of data is referred to as 'T p estimated.' To estimate the parameters, we assume that the data follows a certain distribution.

The 'Gumbel' or 'Extreme Value Type I' distribution will be followed. The Gumbel distribution's CDF (Cumulative Distribution of Function) is as follows: The observed discharge data is x, and the computed distribution parameters

are u and. We can derive the theoretical estimate of p using this distribution.

Make two columns with the labels '(x-u)/' and 'p-theoretical.' Calculate x, sx, u, and using the formulae below.

Calculate the column (x-u)/ using the peak streamflow values (x) as shown:

Calculate the value of p-theoretical using the CDF equation from step 7.

Using the p theoretical values, use the equation used to obtain 'T p estimated' to calculate 'T p theoretical'.

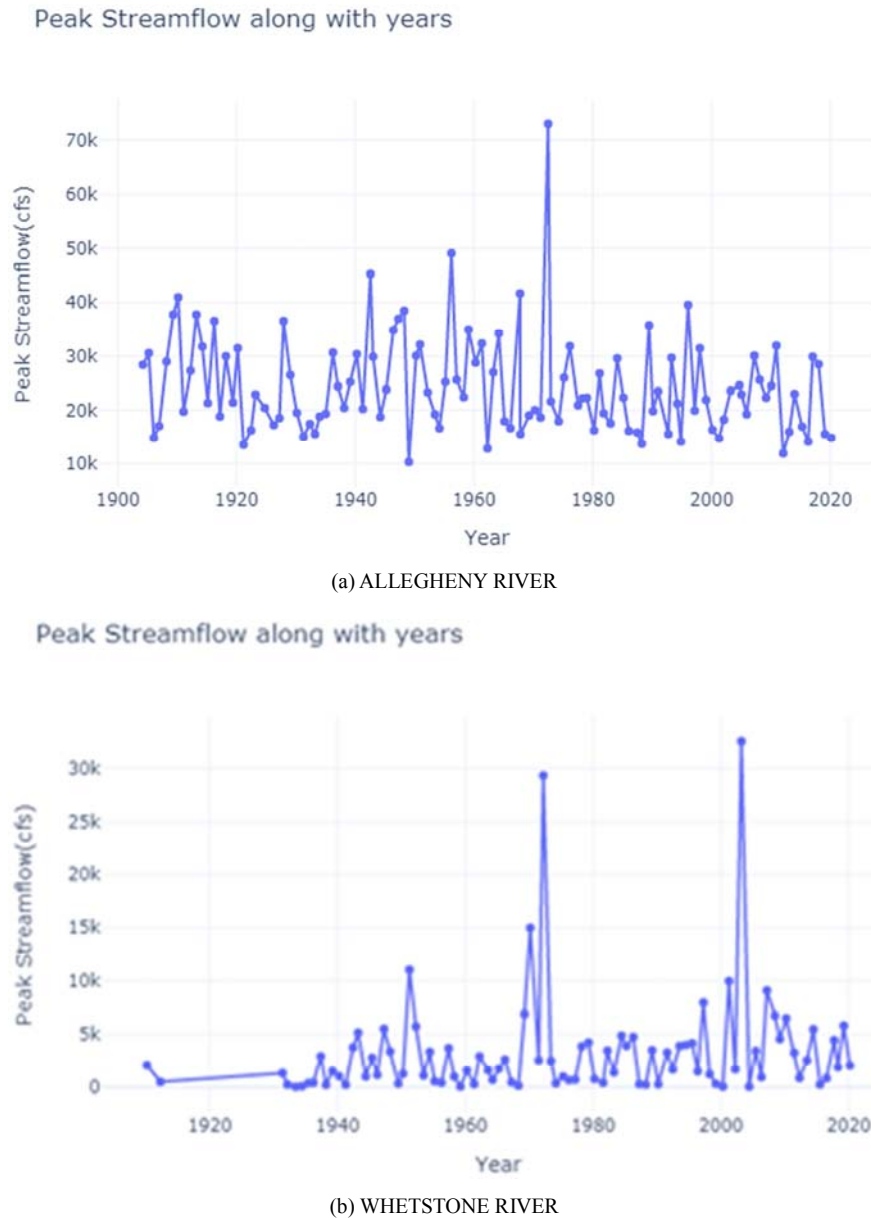
	Time	Peak Streamflow(cfs)	Gage Height(ft)	Rank(i)	qi	pi
0	1910-03-08	2070.0	11.60	42.0	0.451151	0.548849
2	1912-04-15	507.0	5.20	70.0	0.755102	0.244898
3	1931-05-28	1320.0	7.10	53.0	0.570560	0.429440
4	1932-03-01	244.0	5.70	84.0	0.907078	0.092922
5	1933-06-06	45.0	2.84	92.0	0.993921	0.006079
...	...	...	...	...	...	...
88	2016-07-12	847.0	5.46	64.0	0.689970	0.310030
89	2017-09-25	4430.0	11.25	19.0	0.201476	0.798524
90	2018-04-23	1890.0	7.63	44.0	0.472861	0.527139
91	2019-03-22	5800.0	11.97	11.0	0.114633	0.885367
92	2020-03-09	2040.0	7.57	43.0	0.462006	0.537994

92 rows × 6 columns

*Figure 3. Calculation of T\_p estimated and T\_p theoretical using the p theoretical values (ALLEGHENY RIVER AT SALAMANCA NY).*

### 3.4. Graphing the Flood Frequency Curve

Figure 4 shows peak streamflow over years.



**Figure 4.** Peak Streamflow on years.

Plot ' $T_p$  estimated' vs Annual Streamflow. On the same graph, also plot ' $T_p$  theoretical' vs Annual Streamflow.

## 4. Conclusion

Flooding happens all throughout the United States, despite the climate differences. There are high numbers of flood events observed in the Northeast into the Midwest. Many floods occur in the spring in these regions from rainfall and the melting of snow. The area by the Red River, along the border of Minnesota and North Dakota, has noticeable numbers of river flooding. As mentioned previously, floods account for a large significant amount of damage done by all types of natural disasters. More than 2.5 billion people were affected by floods in the past few decades, with many people injured, homeless, or dead [3].

The prediction of increased flood frequency and damage

causes for an increased focus on ways to prevent floods and minimize its damage. One main solution is to make sure people are aware of flood warnings as well as the appropriate responses of what to do when there is a flood. For example, they should be aware of risky behaviors such as entering flood waters, which may increase their chances of getting injured or drowning. In addition to the increased awareness, continuing to study and learn more about flood losses and patterns of natural disasters can help form future solutions and strategies to continue minimizing flood casualties.

Although the climate is dry, the Deserts in the Southwest are still prone to floods. Because the Southwestern areas are larger, they tend to have a higher concentration of flood events — a larger geographic footprint for floods. Flash floods also happen quite often in the summer from thunderstorms due to the Southwest monsoon. Because of the large amount of precipitation, streams, rivers, creeks, and other bodies of water

are quickly filled up and lead to high water levels.

In this paper, the USGS geographical information and Gumbel distribution were used to find the return period corresponding to the exceedance probability. The Gumbel distribution is applied to Allegheny River, NY and Whetstone River, SD.

---

## References

- [1] Chang, Li-Chiu; Shen, Hung-Yu; Chang, Fi-John (2014). "Regional flood inundation nowcast using hybrid SOM and dynamic neural networks". *Journal of Hydrology*. 519 (Part A): 476–489. doi: 10.1016/j.jhydrol.2014.07.036.
- [2] F. D. Mwale (2014), Application of self-organising maps and multi-layer perceptron-artificial neural networks for streamflow and water level forecasting in data-poor catchments: the case of the Lower Shire floodplain, Malawi.
- [3] J. Schaake (2016), Hydrologic Ensemble Prediction Experiment an informal yet highly active group of researchers in the field of predictive hydrologic uncertainty.
- [4] "AMS Glossary". allenpress.com (2015) Archived from the original on 16 July 2012. Retrieved 9 July 2015.
- [5] Surface Water for USA: Peak Streamflow. (2021). U.S. Department of the Interior. [https://nwis.waterdata.usgs.gov/nwis/peak?site\\_no=03011020&agency\\_cd=USGS&format=html](https://nwis.waterdata.usgs.gov/nwis/peak?site_no=03011020&agency_cd=USGS&format=html)
- [6] P. G. Guest, Philip George Guest (2013) *Numerical Methods of Curve Fitting*. Page 349.
- [7] Marriten K. Van Aalst, "The impacts of climate change on the risk of natural disasters", *Disasters*, (2006) Wiley Online Library.
- [8] P. Hoeppe, "Trends in weather related disasters-Consequences for insurers and society", *Weather and climate extremes* (2016) Elsevier.
- [9] D. Alexander, "The study of natural disasters, 1977-97: Some reflections on a changing field of knowledge", *Disasters*, (1997), Wiley Online Library.
- [10] M. Masozera, M. Bailey, C. Kerchner, "Distribution of impacts of natural disasters across income groups: A case study of New Orleans", *Ecological Economics* (2007), Elsevier.