Estimation of Precipitation Rate

Using Computational and Statistical Simulations

Contents

Abstract

It is **not easy** to use a **single variable linear regression** or simple exponential smoothing to determine the effectiveness of weather forecasts such as temperature or flooding forecast especially **when the data pattern is complicated**.

In order 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 were performed.

For the analysis, **the extreme value theory** was used to assess extreme events within probability distributions by quantifying tail behavior. 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.**

Introduction

Recently, Injuries and deaths caused by natural disasters take up a significantly high proportion of news each day. Research shows that the number of natural disasters around the world in the past decades has significantly risen. 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. 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. This damages the society and the buildings in the cities, and even kills people.

It is important to prepare for the catastrophic flood that may damage the society, and this is why I chose this topic and accurately and precisely predict the returning periods of floods based on its peak streamflow, or the extent of the effect of the flood on the society.

Objectives

1. Study hydrologic extremes for design and assessing the impacts of rare climatic events.

2. Introduce a framework for estimating stationary and non-stationary return levels, return periods, and risks of climatic extremes using Bayesian inference.

3. Find return levels and return periods framework implemented in the non-stationary extreme value analysis , explicitly designed to facilitate analysis of extremes in the geosciences.

4. Manage the risks of extreme events and disasters figuring out how the global warming and precipitation would change temporal and spatial pattern of climatic extremes.

Wetland and Dryland

This study examined space-time historical flood and drought variability spanning for a century for the:

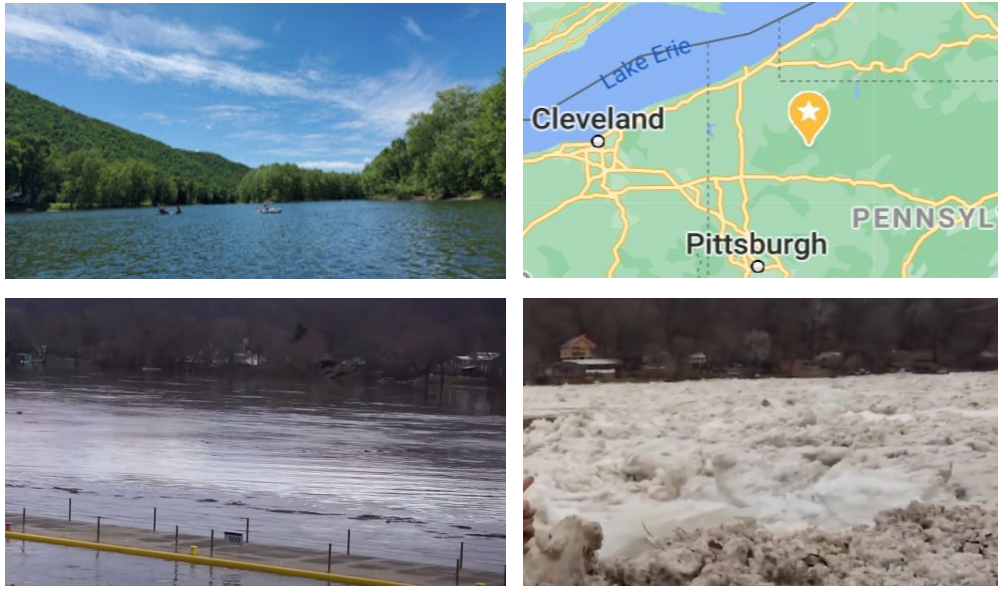
1. **Wetland**(Allegheny river, NY) and

2. **Dryland**(Great Plains)

Wetland - Allegheny River(NY)

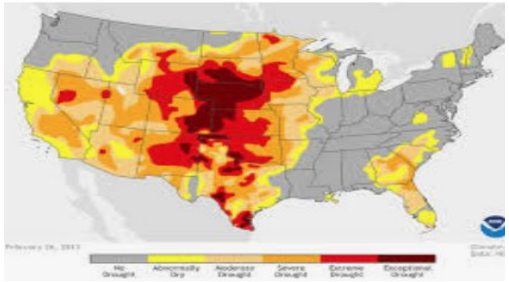
This study examined space-time historical flood and drought variability spanning for a century for the:

1. **wetland**(Allegheny river)
2. **dryland**(Great Plains)



Flooding(summer) and ice jam(winter) on the Allegheny River

Dryland - Great Plains



Dry conditions across Great Plains

Background Information - Extreme Value Theory

- Extreme value theory is used to model the risk of extreme, rare events, such as the 1755 Lisbon earthquake.
- Extreme value theory or extreme value analysis (EVA) is a branch of statistics dealing with the extreme deviations from the median of probability distributions.
- It seeks to assess, from a given ordered sample of a given random variable, the probability of events that are more extreme than any previously observed.

- Extreme Value Theory (EVT) can provide a rigorous framework for analysis of climate extremes and their return levels
- Under a wide range of conditions, the distribution of the maxima or minima converges to one of the three limiting distributions: Gumbel, Fréchet, or Weibull

Extreme Value Theory

3 distribution parameters $\theta=(\mu,\sigma,\xi)$:

(1) the location parameter (μ) specifies the center of the distribution

(2) the scale parameter (σ) determines the size of deviations around the location parameter

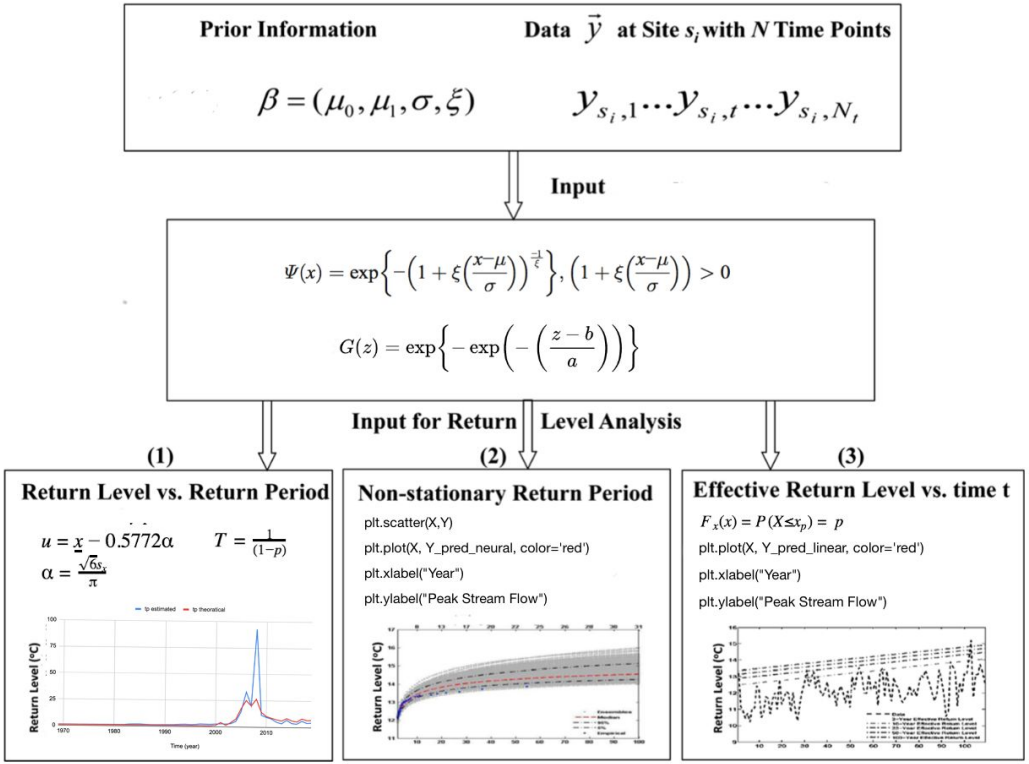
(3) the shape parameter (ξ) governs the tail behavior of the distribution.

The limiting case of $\xi \rightarrow 0$ gives the Gumbel distribution, $\xi < 0$ the Weibull distribution and $\xi > 0$ the Fréchet distribution.

Additional Background Information

- Gumbel distribution is used to model the distribution of the maximum/minimum of a number of samples of various distributions.
- Exceedance probability is the probability that the event will exceed some critical value (usually far from the mean).
- Return period is an estimate of the likelihood of an event to occur. A statistical measurement based on historic data denoting the average recurrence interval.
- Theoretical probability is the fraction of times we expect the 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).
- Estimated probability approaches the theoretical probability as the number of trials gets larger. It is an approximation of theoretical probability.

Procedure


Prior Information
Data $\vec{y}$ at Site s_i with N Time Points
$\beta = (\mu_0, \mu_1, \sigma, \xi)$
$y_{s_i,1} \cdots y_{s_i,t} \cdots y_{s_i,N_t}$
Input
$\Psi(x) = \exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right\}, \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right) > 0$
$G(z) = \exp\left\{-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right\}$
Input for Return Level Analysis
(1)
Return Level vs. Return Period
$u = \underline{x} - 0.5772\alpha$
$T = \frac{1}{(1-p)}$
$\alpha = \frac{\sqrt{6}s_x}{\pi}$
Return Level (°C)
Time (year)
(2)
Non-stationary Return Period
plt.scatter(X,Y)
plt.plot(X, Y_pred_neural, color='red')
plt.xlabel("Year")
plt.ylabel("Peak Stream Flow")
Return Level (°C)
(3)
Effective Return Level vs. time t
$F_x(x) = P(X \le x_p) = p$
plt.plot(X, Y_pred_linear, color='red')
plt.xlabel("Year")
plt.ylabel("Peak Stream Flow")
Return Level (°C)

Procedure

1. Label column A(first column) as ‘time’ and column B(second column) as the ‘annual peak streamflow (cfs)’ and enter in appropriate values in each cell.
2. Select the streamflow values in column B then sort the values from smallest to largest by clicking on the ‘Sort and Filter’ tool. Allow expanding the selection.
3. Label column C as ‘Rank (i)’ and rank the data in decreasing order (from N to 1).
4. Create a fourth column called qi. Gringorten plotting position formula will be used to calculate the estimated exceedance probabilities relevant to past observations.
5. Make another column and label it pi. Then make it equal to 1-qi. pi refers to the non-exceedance probability.
6. Create one more column and label it ‘Tp estimated’ and evaluate the values in pi using the equation for the return period.

Procedure

- We will follow the ‘Gumbel’ or ‘Extreme Value Type 1’ distribution. The CDF (Cumulative distribution of function) of the Gumbel distribution is the following: x is the observed discharge data; u and are the calculated parameters of the distribution. This distribution will allow us to calculate the theoretical estimate of p.
- Create two columns labeled ‘(x-u)/α’ and ‘p-theoretical. Using the following equations, calculate $\bar{x}$, sx, u and α.
- Use the peak streamflow values (x) and calculate the column (x-u)/α as shown:
- Use the CDF equation from step 7 to calculate the value of p-theoretical.
- Use the equation used to calculate 'Tp estimated’ and use it to calculate 'Tp theoretical’ using the p theoretical values.

Part A: Precipitation in Wetland

ALLEGHENY RIVER AT SALAMANCA NY

Hydrologic Unit Code 05010001, Latitude 42°09'23", Longitude 78°42'55" NAD83, Drainage area 1,608 square miles

Data

To find the return period corresponding to the exceedance probability, Gumbel distribution is applied to the Allegheny River in Salamanda, NY. The USGS number and geographical information are shown as follows

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

Procedure & Data collection

	A	B	C	D	E	F	G	H	I
1	**Time**	**Peak Streamflow(cfs)**	**Rank (i)**	**q_i**	**p_i**	**T_p estimated**	**(x-u)/a**	**p theoretical**	**T_p theoretical**
2	Jan. 06, 1949	10,300	103	0.9066478	0.0933522	1.10296412	-1.912349	0.00114888	1.001150198
3	Jan. 28, 2012	11,900	102	0.8978076	0.1021924	1.11382434	-1.589016	0.0074546	1.007510585
4	Apr. 08, 1962	12,800	101	0.8889675	0.1110325	1.12490056	-1.407141	0.01683558	1.017123875
5	Mar. 07, 1921	13,500	100	0.8801273	0.1198727	1.13619928	-1.265682	0.02885387	1.029711151
6	Apr. 04, 1988	13,700	99	0.8712871	0.1287129	1.14772727	-1.225266	0.03320448	1.034344879
7	Nov. 02, 1994	14,100	98	0.862447	0.137553	1.15949159	-1.144432	0.04325433	1.045209847
8	Apr. 12, 2016	14,100	97	0.8536068	0.1463932	1.17149959	-1.144432	0.04325433	1.045209847
9	Apr. 08, 2001	14,700	96	0.8447666	0.1552334	1.18375889	-1.023182	0.06191245	1.065998579
10	Jan. 24, 1906	14,800	95	0.8359264	0.1640736	1.1962775	-1.002974	0.06545594	1.070040507
11	Apr. 11, 1931	15,000	94	0.8270863	0.1729137	1.2090637	-0.962557	0.07292061	1.07865627
12	Mar. 15, 1933	15,500	93	0.8182461	0.1817539	1.22212619	-0.861516	0.09378549	1.103491485
13	Oct. 20, 1967	15,500	92	0.8094059	0.1905941	1.23547401	-0.861516	0.09378549	1.103491485
14	Sep. 23, 1992	15,500	91	0.8005658	0.1994342	1.24911661	-0.861516	0.09378549	1.103491485
15	Jul. 03, 1987	15,800	90	0.7917256	0.2082744	1.26306387	-0.800891	0.107795	1.120818644

Procedure & Data collection

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.

$$q_i = \frac{i - a}{N + 1 - 2a}$$

q_i = exceedance probability associated with a specific observation
N = number of annual maxima observations
i = Rank of specific observation (i=1 is the largest and i=N is the smallest)
a = constant for estimation (0.44)

	A	B	C	D
1	**Time**	**Peak Streamflow(cfs)**	**Rank (i)**	**qi**
2	Jan. 06, 1949	10,300	103	0.906647808
3	Jan. 28, 2012	11,900	102	0.897807638
4	Apr. 08, 1962	12,800	101	0.888967468
5	Mar. 07, 1921	13,500	100	0.880127298
6	Apr. 04, 1988	13,700	99	0.871287129
7	Nov. 02, 1994	14,100	98	0.862446959
8	Apr. 12, 2016	14,100	97	0.853606789
9	Apr. 08, 2001	14,700	96	0.84476662
10	Jan. 24, 1906	14,800	95	0.83592645
11	Apr. 11, 1931	15,000	94	0.82708628
12	Mar. 15, 1933	15,500	93	0.81824611
13	Oct. 20, 1967	15,500	92	0.809405941
14	Sep. 23, 1992	15,500	91	0.800565771
15	Jul. 03, 1987	15,800	90	0.791725601

Procedure & Data collection

If X is a random variable with a cumulative distribution function $F_x(x)$, the probability that X is less than equal (not exceeding) to a given event x_p is:

$$F_x(x) = P(X \leq x_p) = p$$

The probability that this event will be exceeded is now $1 - p$, and the percent exceedance would be $100(1 - p)\%$.

For an event x_p, the return period corresponding to this exceedance probability is denoted by T.

$$T = \frac{1}{(1-p)}$$

	A	B	C	D	E	F
1	**Time**	**Peak Streamflow(cfs)**	**Rank (i)**	**qi**	**pi**	**Tp estimated**
2	Jan. 06, 1949	10,300	103	0.9066478	0.0933522	1.10296412
3	Jan. 28, 2012	11,900	102	0.8978076	0.1021924	1.11382434
4	Apr. 08, 1962	12,800	101	0.8889675	0.1110325	1.12490056
5	Mar. 07, 1921	13,500	100	0.8801273	0.1198727	1.13619928
6	Apr. 04, 1988	13,700	99	0.8712871	0.1287129	1.14772727
7	Nov. 02, 1994	14,100	98	0.862447	0.137553	1.15949159
8	Apr. 12, 2016	14,100	97	0.8536068	0.1463932	1.17149959
9	Apr. 08, 2001	14,700	96	0.8447666	0.1552334	1.18375889
10	Jan. 24, 1906	14,800	95	0.8359264	0.1640736	1.1962775
11	Apr. 11, 1931	15,000	94	0.8270863	0.1729137	1.2090637
12	Mar. 15, 1933	15,500	93	0.8182461	0.1817539	1.22212619
13	Oct. 20, 1967	15,500	92	0.8094059	0.1905941	1.23547401
14	Sep. 23, 1992	15,500	91	0.8005658	0.1994342	1.24911661
15	Jul. 03, 1987	15,800	90	0.7917256	0.2082744	1.26306387

Procedure & Data collection

We will follow the 'Gumbel' or 'Extreme Value Type 1' distribution. The CDF (Cumulative distribution of function) of the Gumbel distribution is the following:

$$F_x(x) = \exp\left[-exp\left(-\frac{x-u}{\alpha}\right)\right] = p$$

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

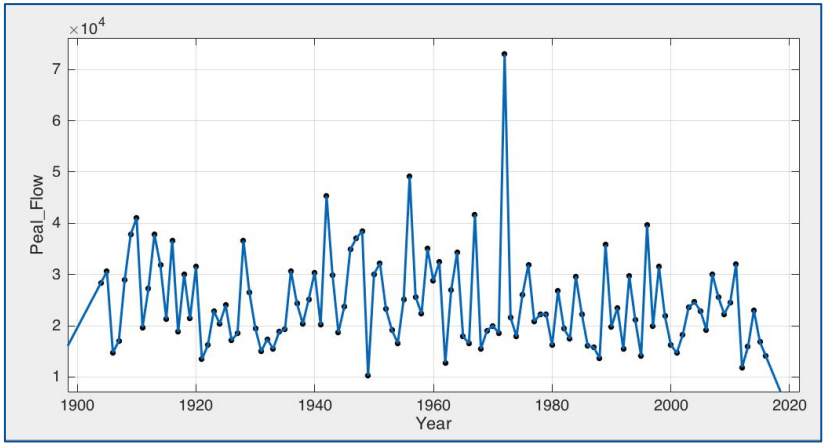
$$s_x{}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

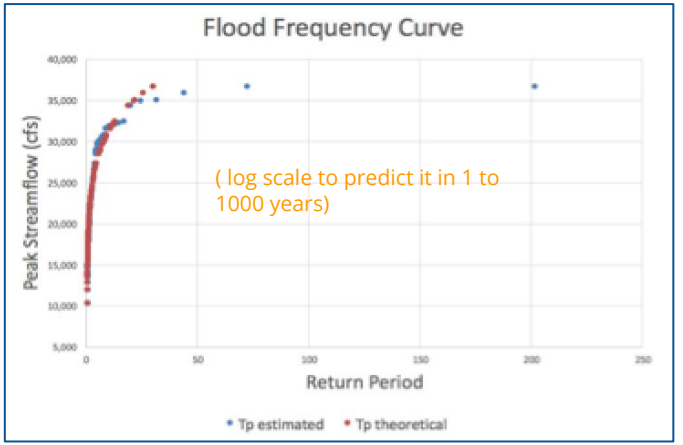
$$u = \bar{x} - 0.5772\alpha$$

$$\alpha = \frac{\sqrt{6} s_x}{\pi}$$

G	H	I
(x-u)/a	p theoretical	T_p theoretical
-1.912349	0.00114888	1.001150198
-1.589016	0.0074546	1.007510585
-1.407141	0.01683558	1.017123875
-1.265682	0.02885387	1.029711151
-1.225266	0.03320448	1.034344879
-1.144432	0.04325433	1.045209847
-1.144432	0.04325433	1.045209847
-1.023182	0.06191245	1.065998579
-1.002974	0.06545594	1.070040507
-0.962557	0.07292061	1.07865627
-0.861516	0.09378549	1.103491485
-0.861516	0.09378549	1.103491485
-0.861516	0.09378549	1.103491485
-0.800891	0.107795	1.120818644

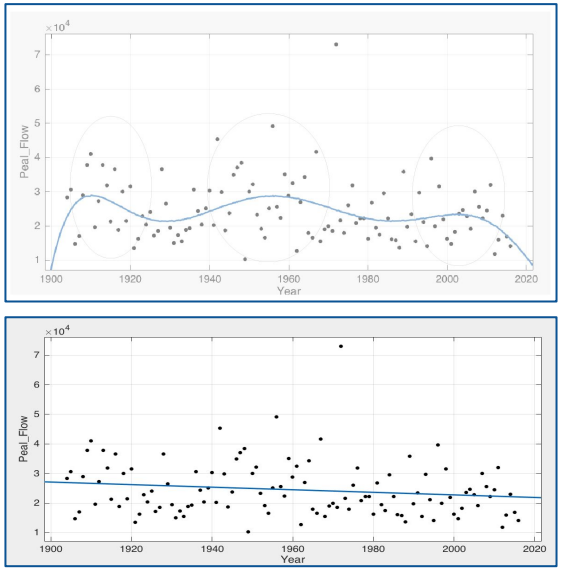
Allegheny River

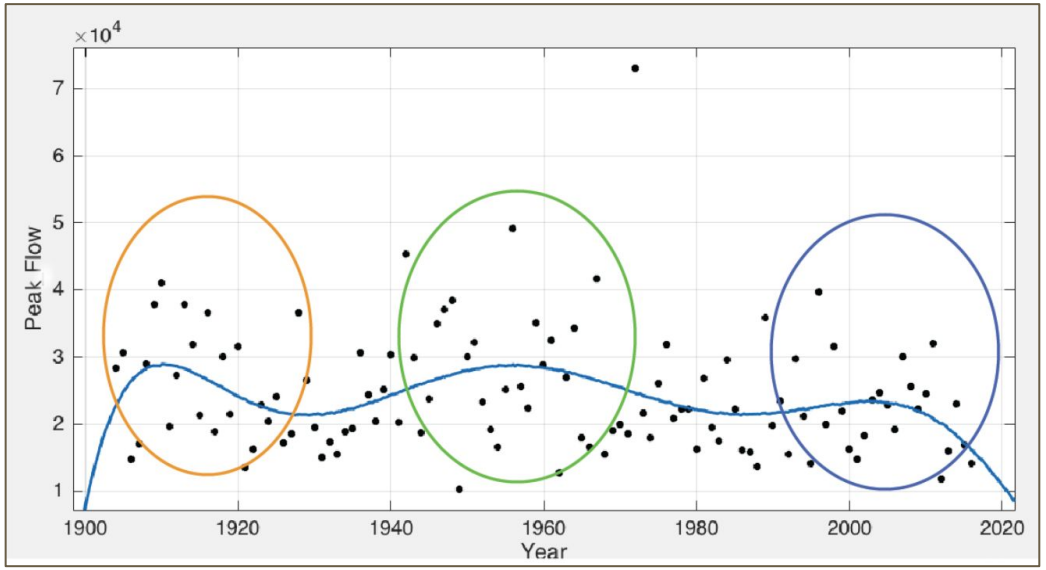




Plot of estimated' vs Annual Streamflow.
On the same graph, also plot theoretical' vs Annual Streamflow was drawn.

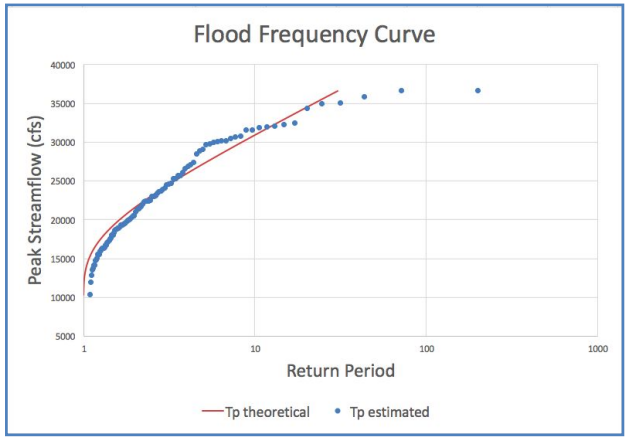
Allegheny River

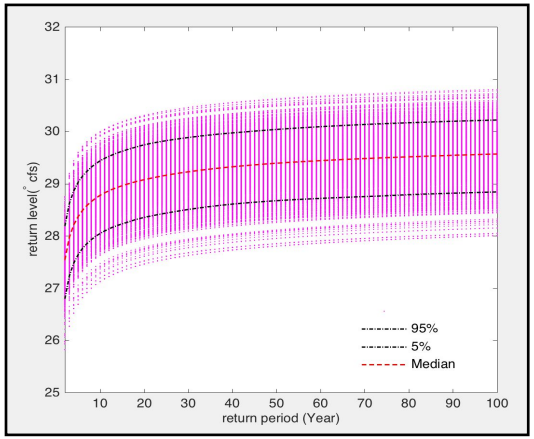


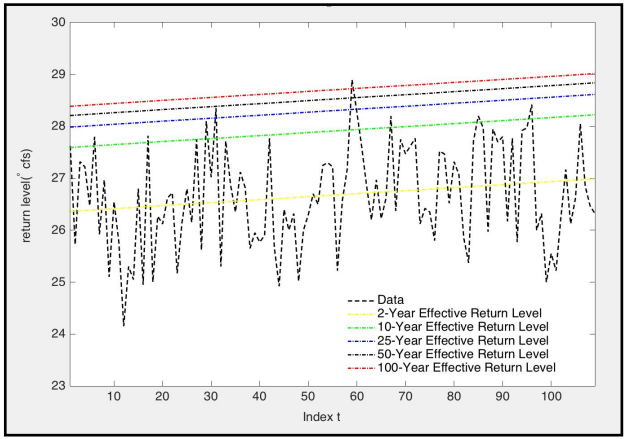


Linear model Poly8:: f(x) = p1*x^8 + p2*x^7 + p3*x^6 + p4*x^5 + p5*x^4 + p6*x^3 + p7*x^2 + p8*x + p9
Coefficients (with 95% confidence bounds):
p1 = 1.121e-10 (-7.068e-10, 9.309e-10)
p2 = -1.716e-06 (-1.452e-05, 1.109e-05)
p9 = 1.969e+16 (-1.545e+17, 1.938e+17)

Graphing the flood frequency curve (Wetland-NY)







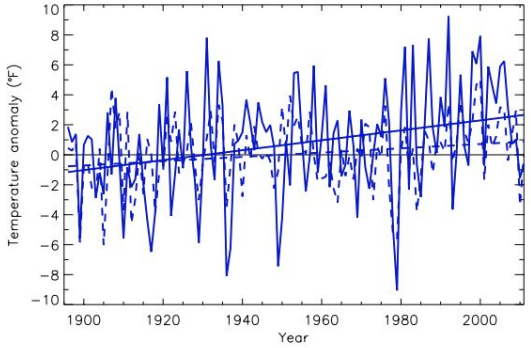
The orange line represents the theoretical distribution, while the blue dots stand for the fit of the annual peak streamflow data with respect to a Gumbel distribution. You can predict streamflow values corresponding to any return period from 1 to 100. The curve follows the distribution very well for low flows, but starts to drift away from the theoretical at higher flows. (Left: log scale to predict it in 1 to 1000 years)

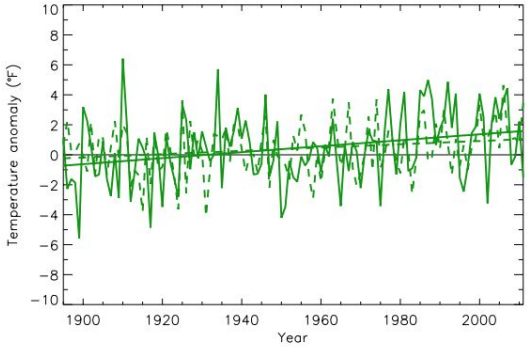
Part B: Precipitation in Dryland

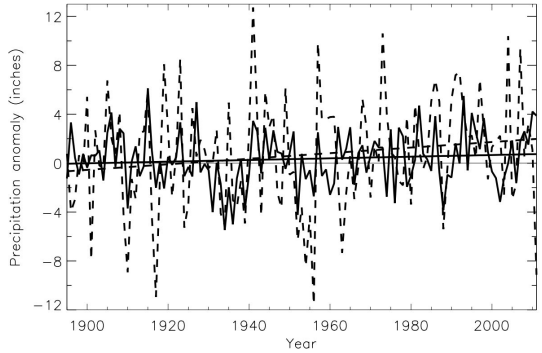
GREAT PLAINS AT Grant County, South Dakota

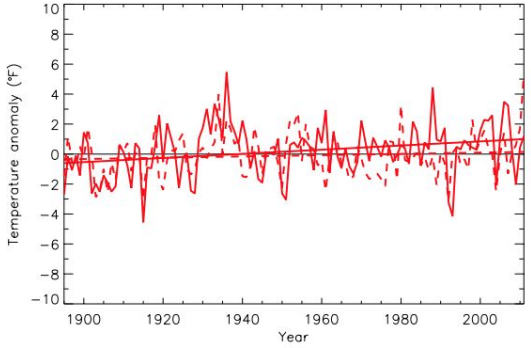
Hydrologic Unit Code 07020001, Latitude 45°17'30", Longitude 96°29'14" NAD27, Drainage area 398 square miles, Gage datum 996.96 feet above NGVD29

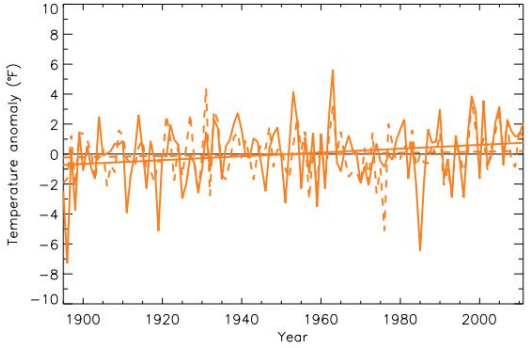
Great Plains(Dryland)- Precipitation





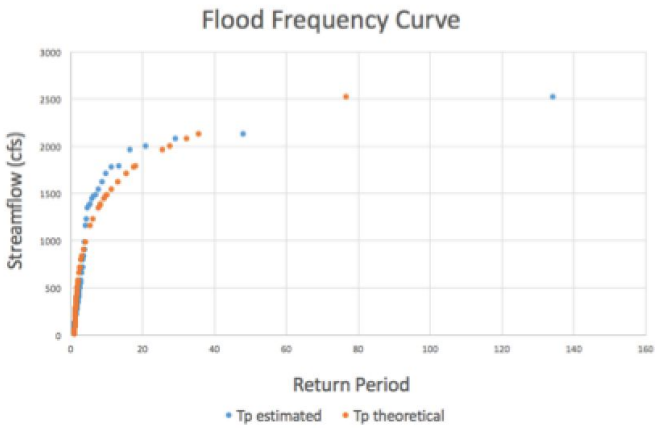


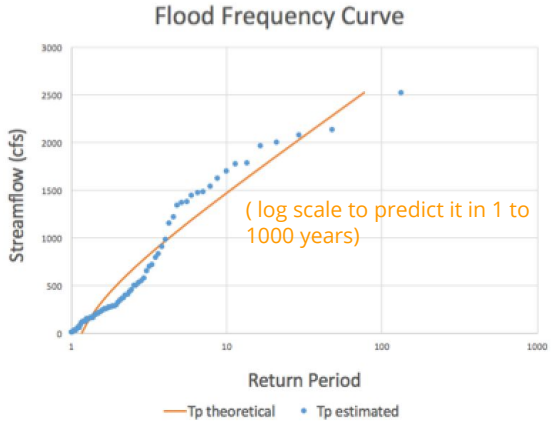




Precipitation anomaly (inches) for annual (black), winter(blue), spring (green), summer (red), and fall (orange), for the northern (solid lines) and southern (dashed lines) U.S. Great Plains. Dashed lines indicate the best fit by minimizing the chi-square error statistic

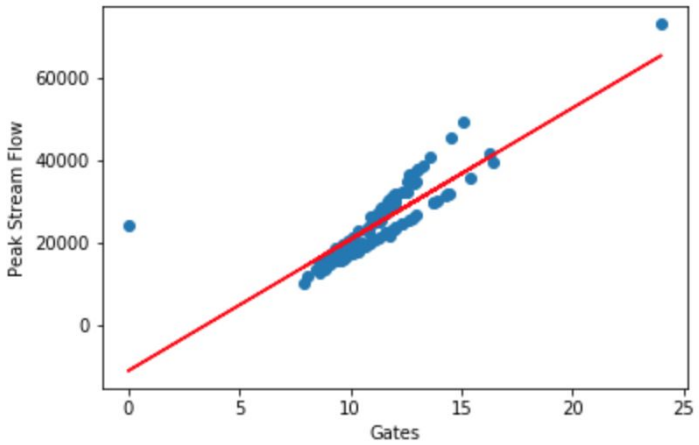
Graphing the flood frequency curve (Dryland-G.P.)

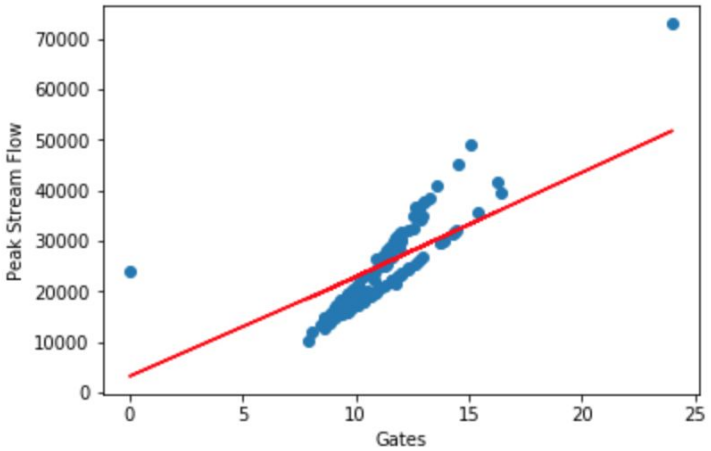




The orange line represents the theoretical distribution, while the blue dots stand for the fit of the annual peak streamflow data with respect to a Gumbel distribution. You can predict streamflow values corresponding to any return period from 1 to 100. The curve follows the distribution very well for low flows, but starts to drift away from the theoretical at higher flows. (Right: log scale to predict it in 1 to 1000 years)

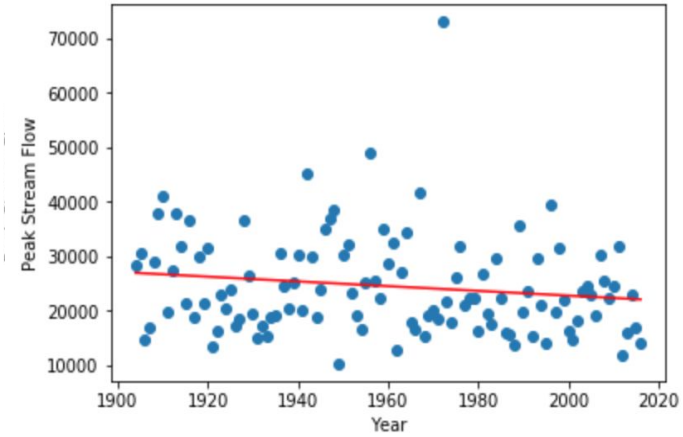
Stationary EVA Using Machine Learning(SVM)

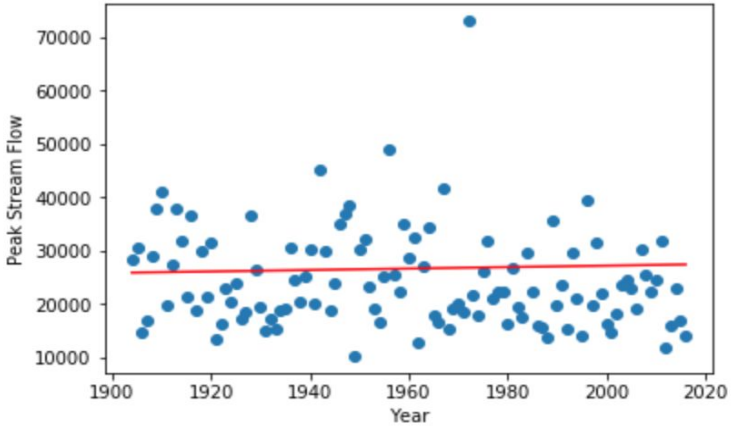




Linear regression vs. Support Vector Machine(SVM) - NY Allegheny River

Non-Stationary EVA Using AI(Neural Networks)





Linear regression(left) vs. Neural Network(right) - NY Allegheny River

Discussion

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 San Jose in CA and Allegheny River NY.