BASIC STATISTICS

GLY606 Water Data Analysis & Modeling

Sep 23th 2024

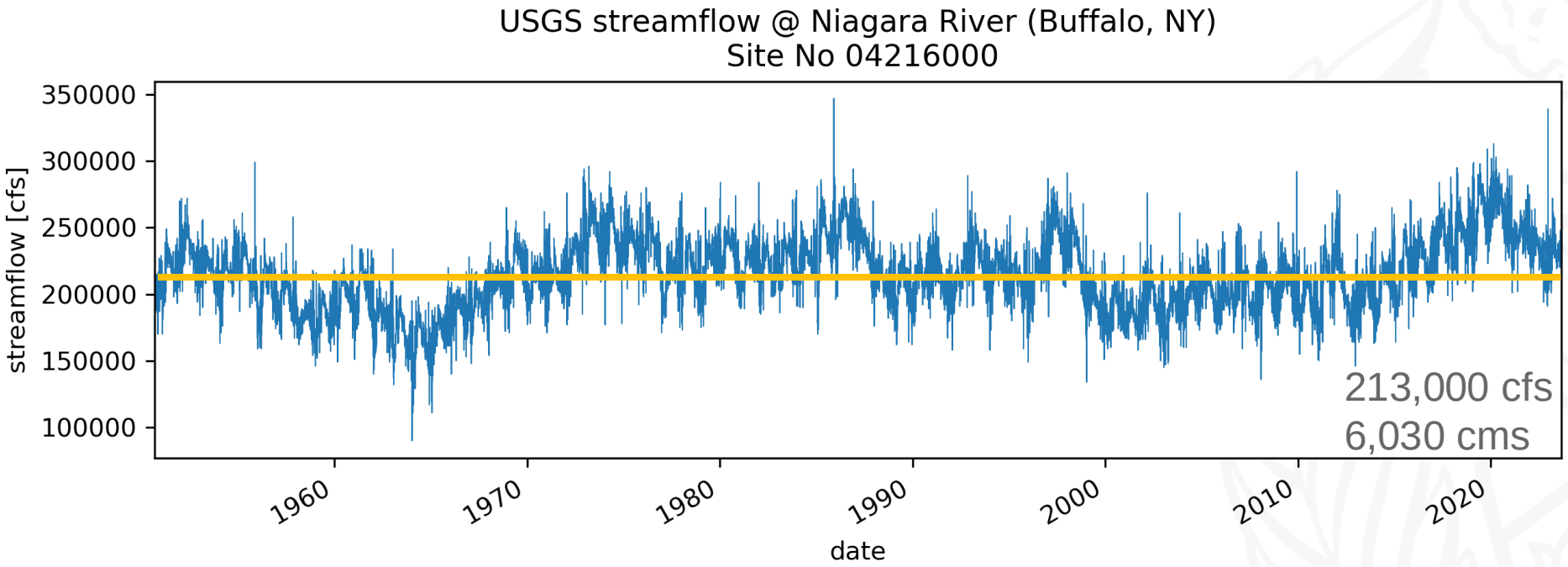
University at Buffalo The State University of New York

Statistical methods are widely used in hydrologic modeling

How do we do descriptive analysis when we get a data?

- Mean, variance, standard deviation (Box plot)
- PDF(Histogram), CDF (Quantile mapping), median (inter-quantile range)
- Extreme detection
  - z-score
  - 7Q10

If you were a state hydrologist, when you were asked to give a high-level introduction to Niagara Rivers at Buffalo, what information would you provide based on the streamflow observations?

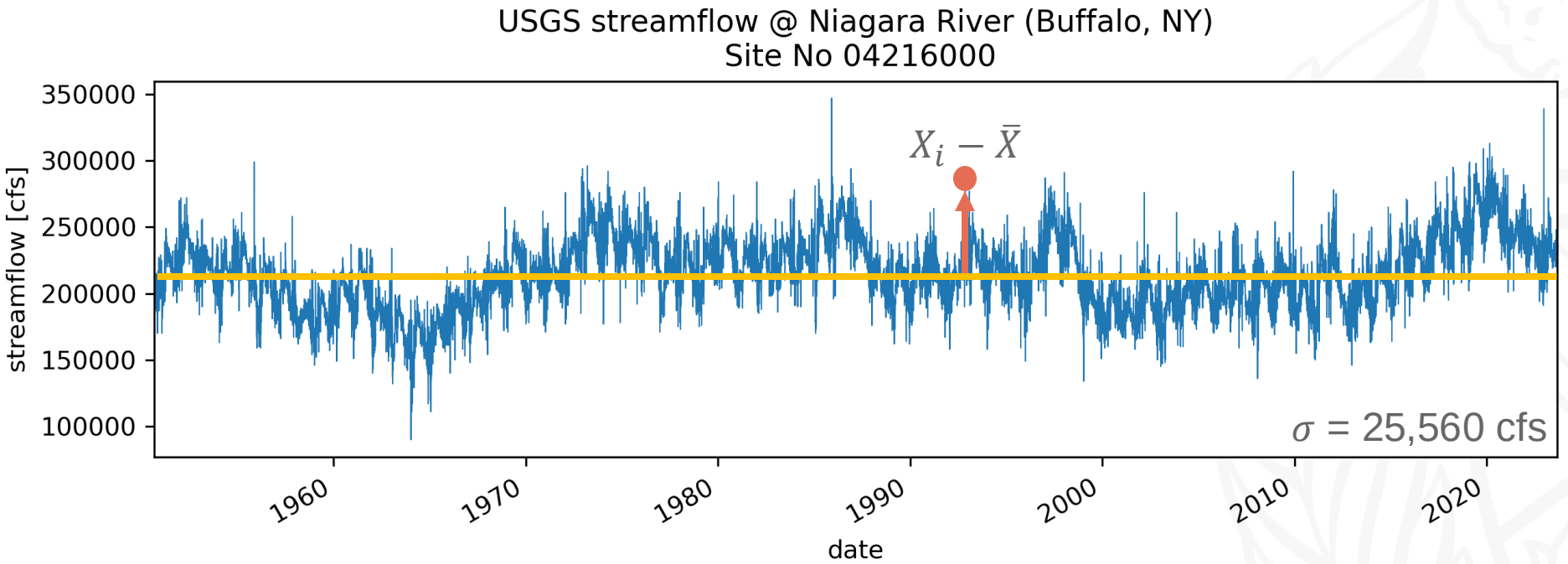


1. Sample Mean

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

Mean streamflow usually is used to evaluate the overall water availability for a region.

If you were a state hydrologist, when you were asked to give a high-level introduction to Niagara Rivers at Buffalo, what information would you provide based on the streamflow observations?



2. Sample Variance

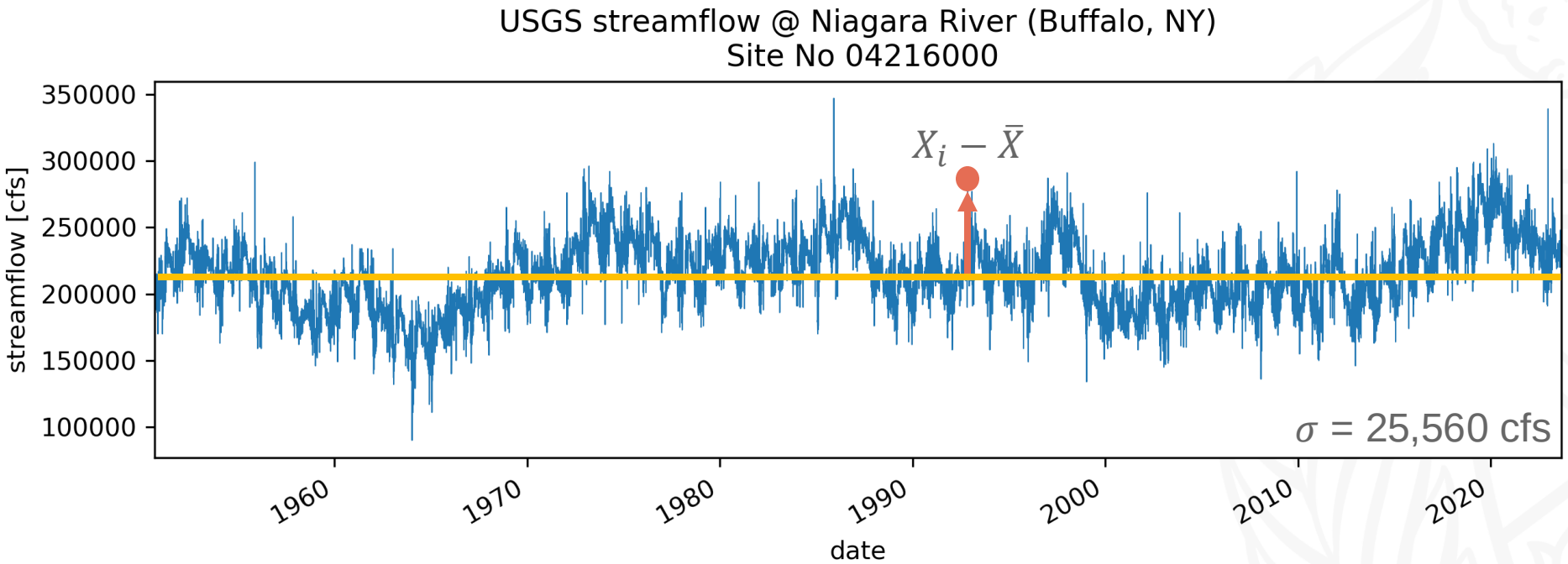
$$\sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

3. Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

"Variance" refers to a measure of how spread out a set of data is from its mean (average), essentially indicating how much variation exists within a data set

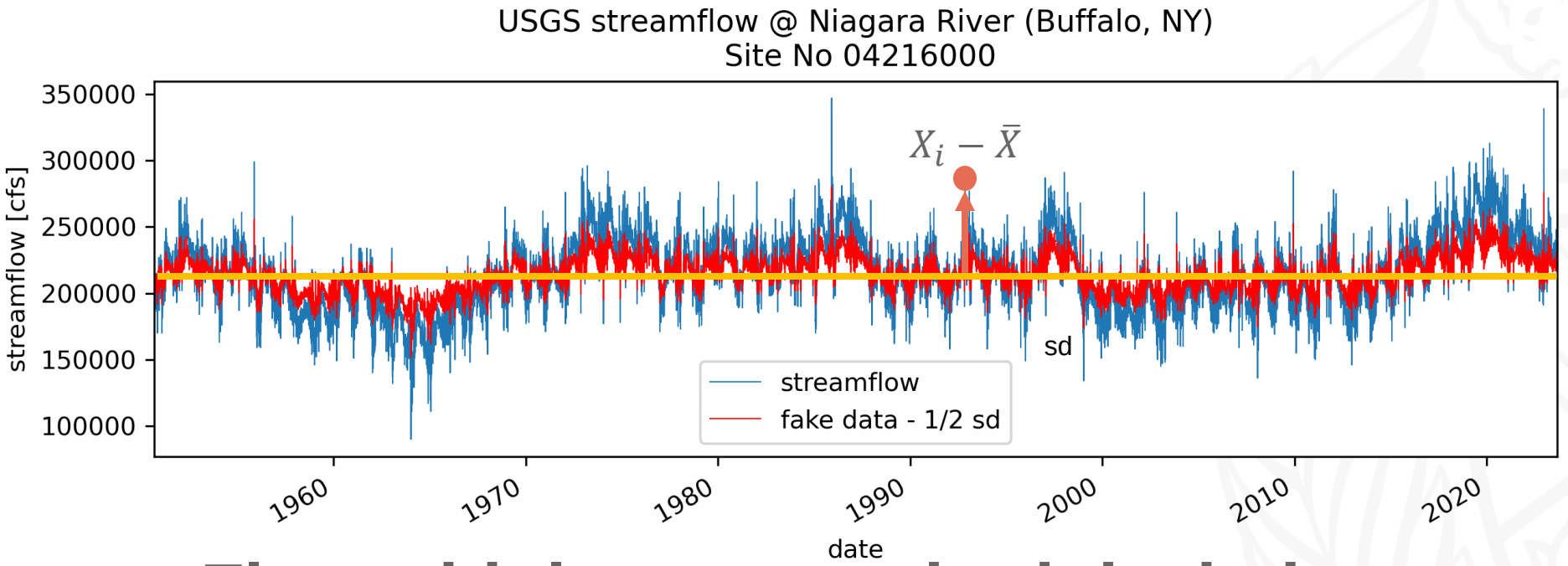
If you were a state hydrologist, when you were asked to give a high-level introduction to Niagara Rivers at Buffalo, what information would you provide based on the streamflow observations?



How do we create a time series with same $\bar{X}$ but 0.5 σ?

$$X_{i,fake} = \bar{X} + \frac{1}{2}(X_i - \bar{X})$$

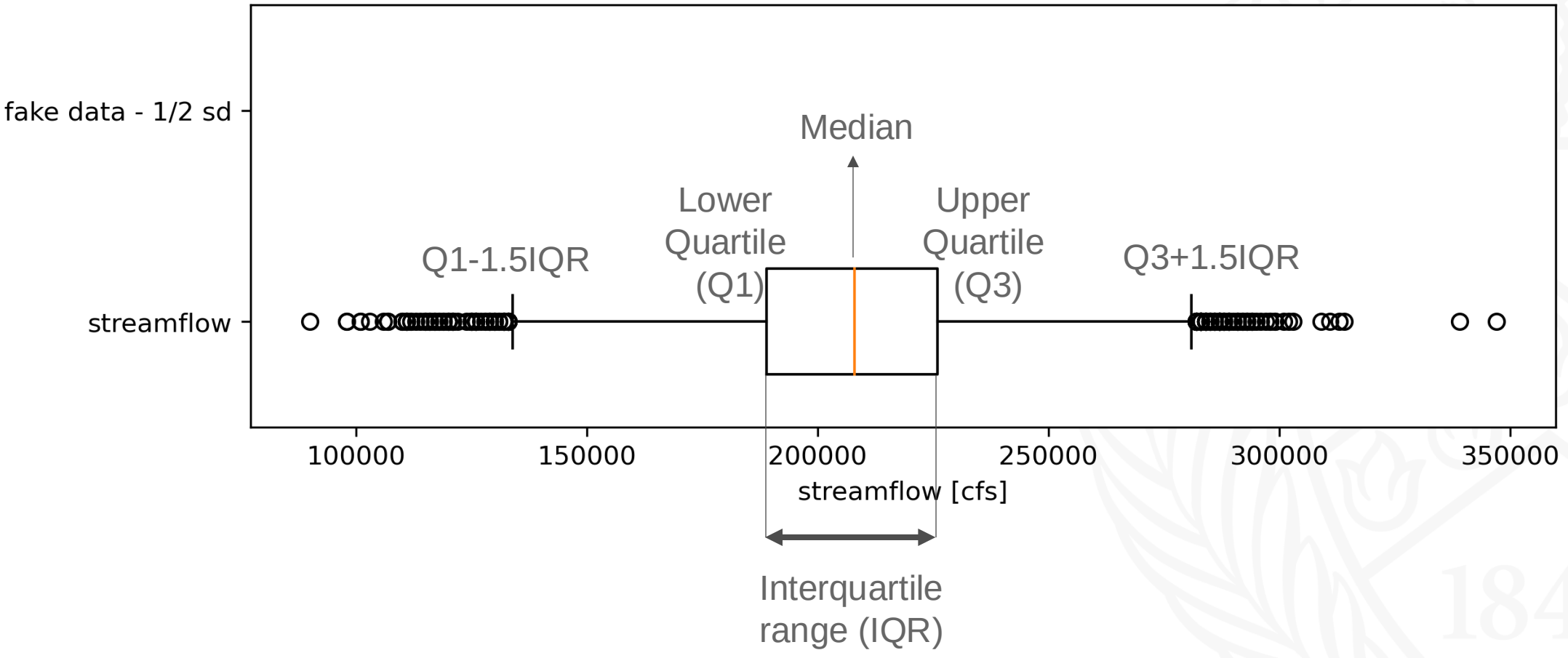
If you were a state hydrologist, when you were asked to give a high-level introduction to Niagara Rivers at Buffalo, what information would you provide based on the streamflow observations?



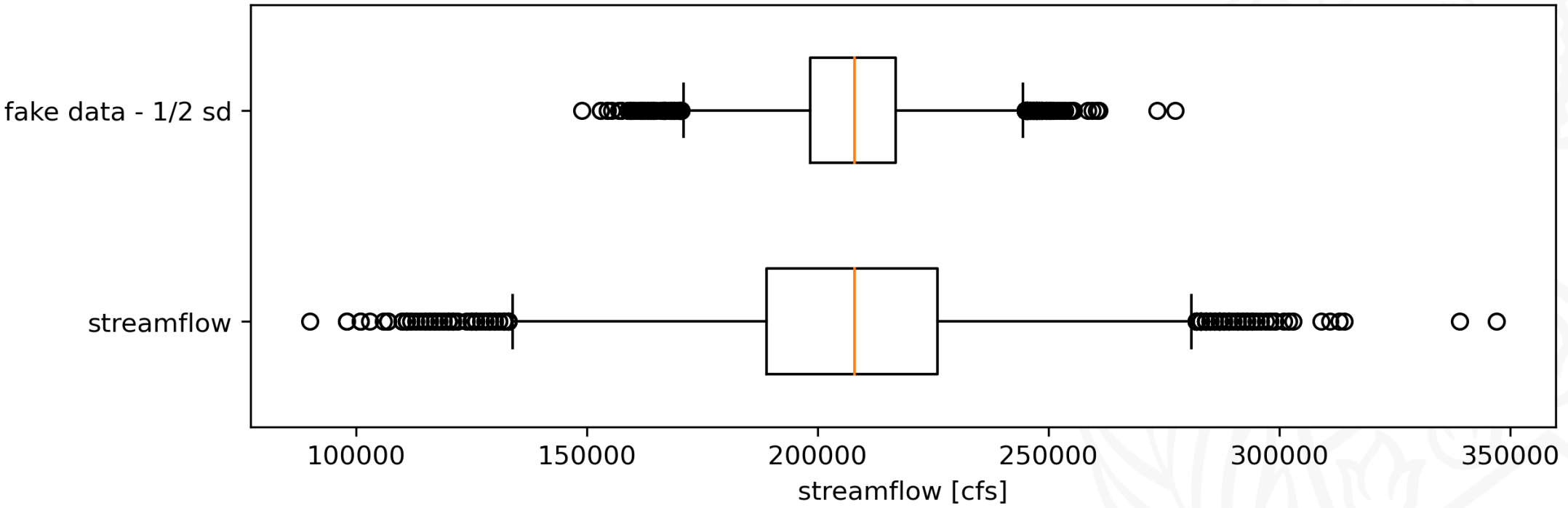
A more straightforward way to visualize the spread of the dataset?

Flow with lower standard deviation are more centered around the mean value!

Box-plot


fake data - 1/2 sd
streamflow
Median
Lower Quartile (Q1)
Upper Quartile (Q3)
Q1-1.5IQR
Q3+1.5IQR
100000
150000
200000
250000
300000
350000
streamflow [cfs]
Interquartile range (IQR)

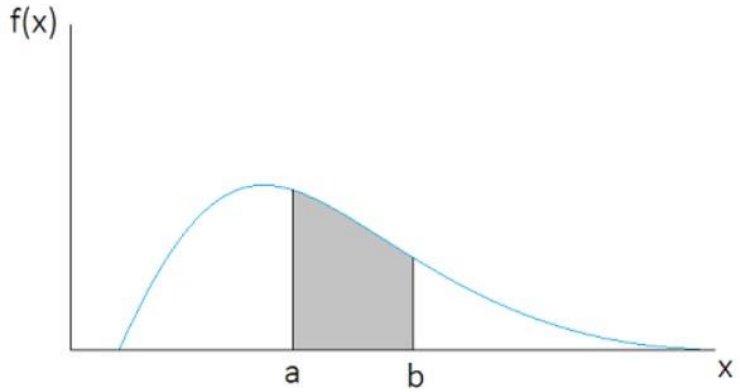
Box-plot


fake data - 1/2 sd
streamflow
100000
150000
200000
250000
300000
350000
streamflow [cfs]

How do we evaluate the distribution of streamflow data?

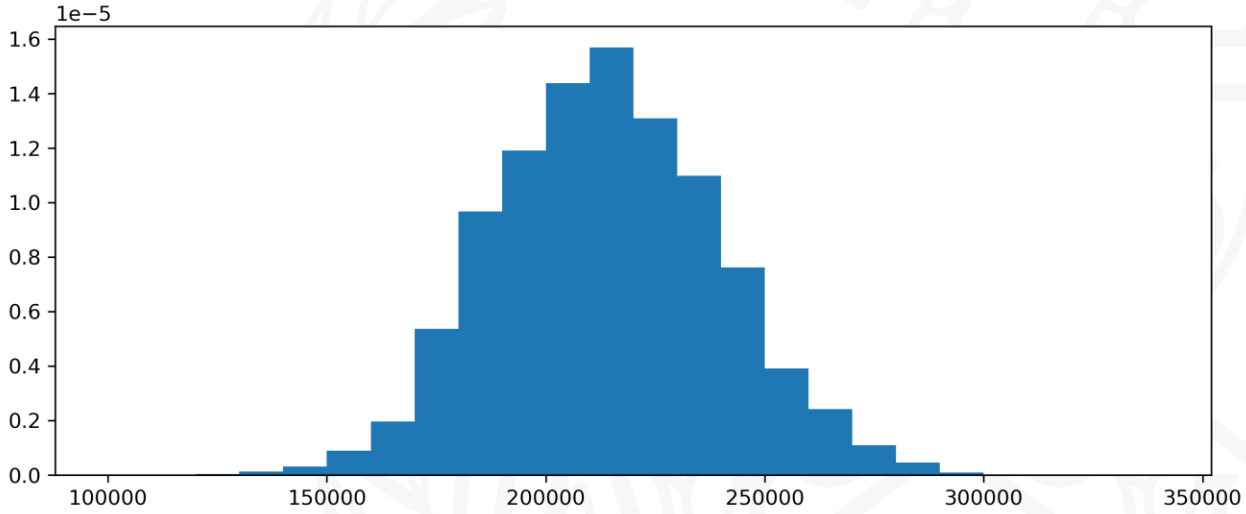
- Probability Density Function

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$



The total area below a PDF is 1.

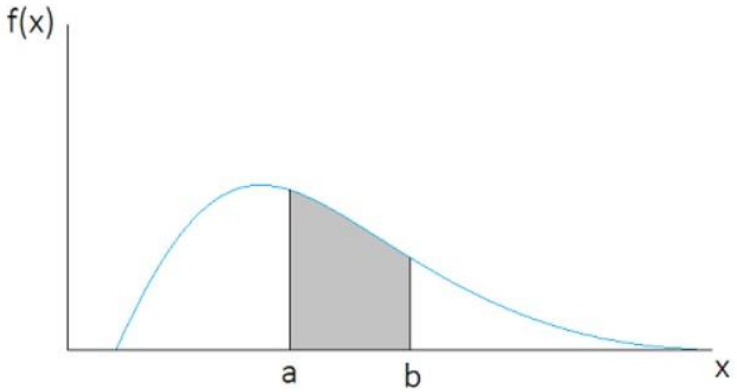
We can use histogram to visualize the PDF for time series data.



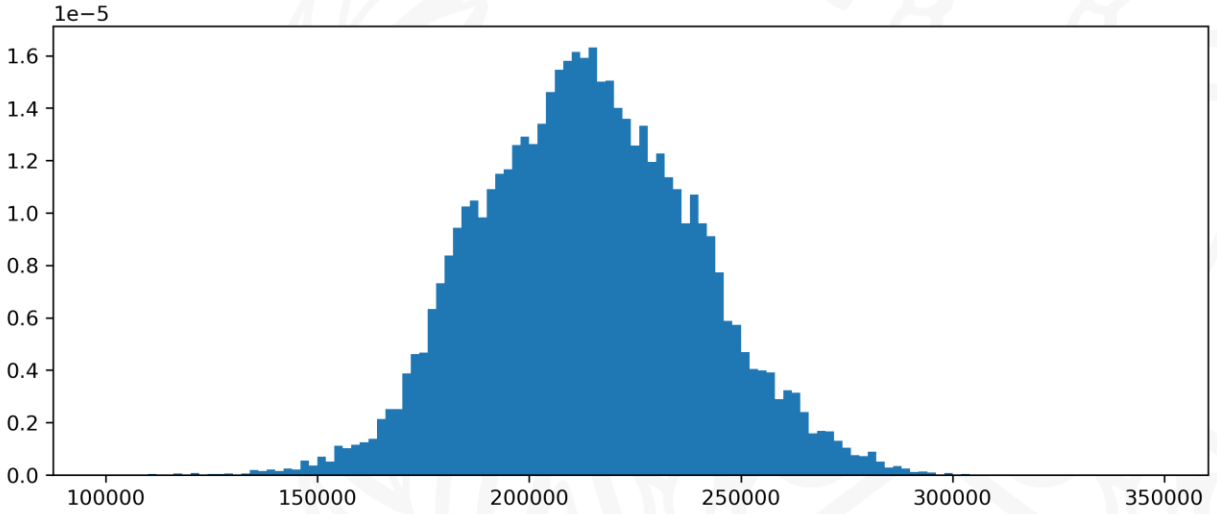
How do we evaluate the distribution of streamflow data?

- Probability Density Function

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$



We can use histogram to visualize the PDF for time series data



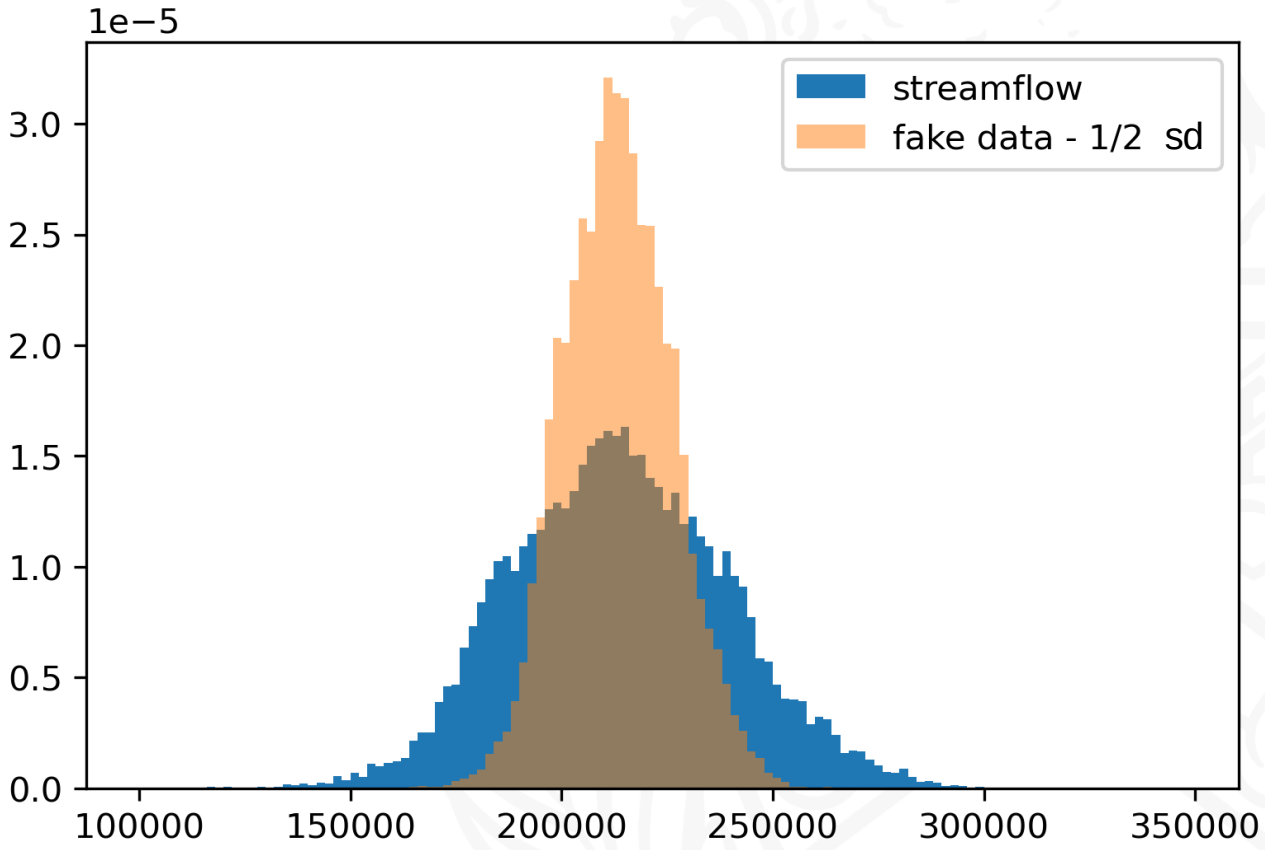
The histogram might look a bit different after we change the bin size

How do we evaluate the distribution of streamflow data?

- Probability Density Function

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

What does the PDF for our fake streamflow data look like?



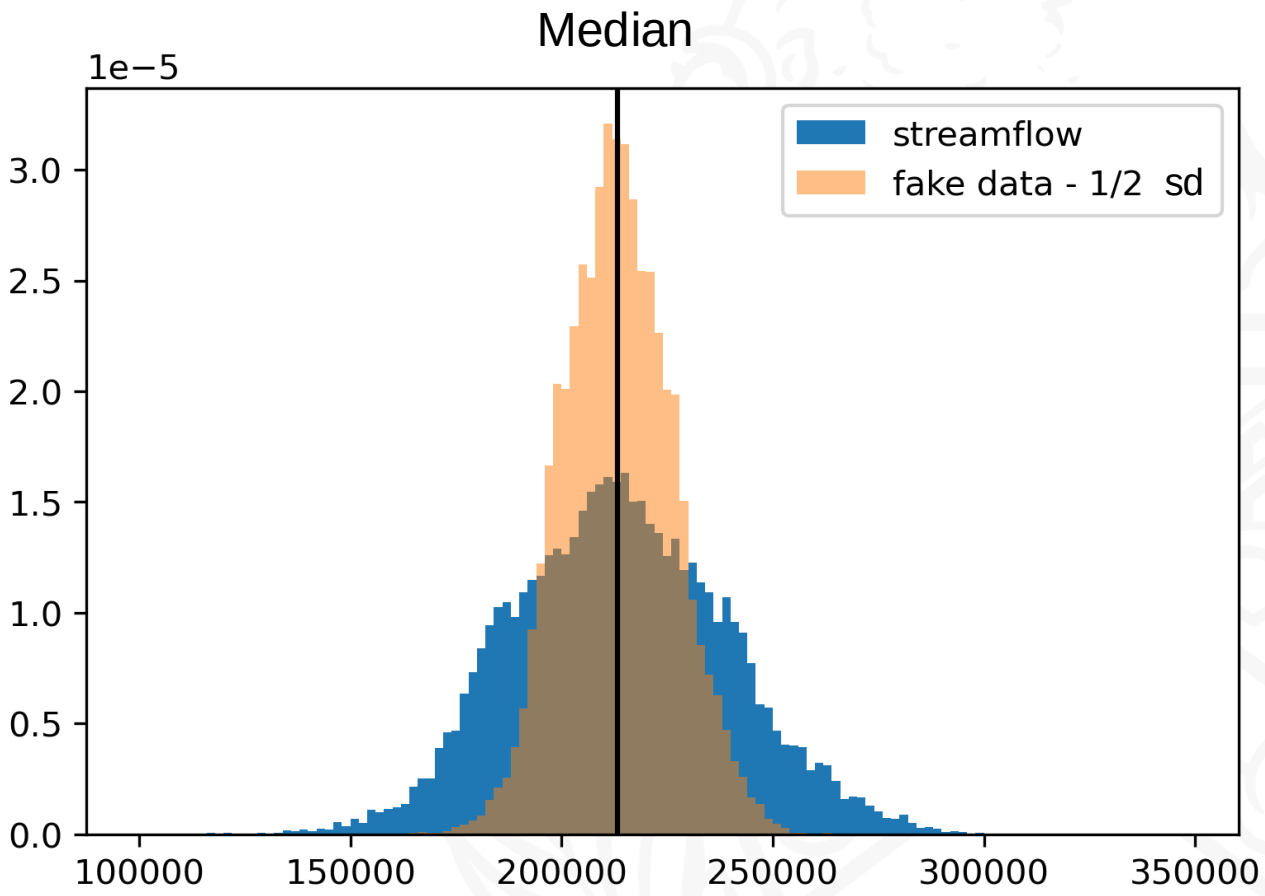
How do we evaluate the distribution of streamflow data?

$$P(X \leq median) = 0.5$$

- Probability Density Function

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

What does the PDF for our fake streamflow data look like?

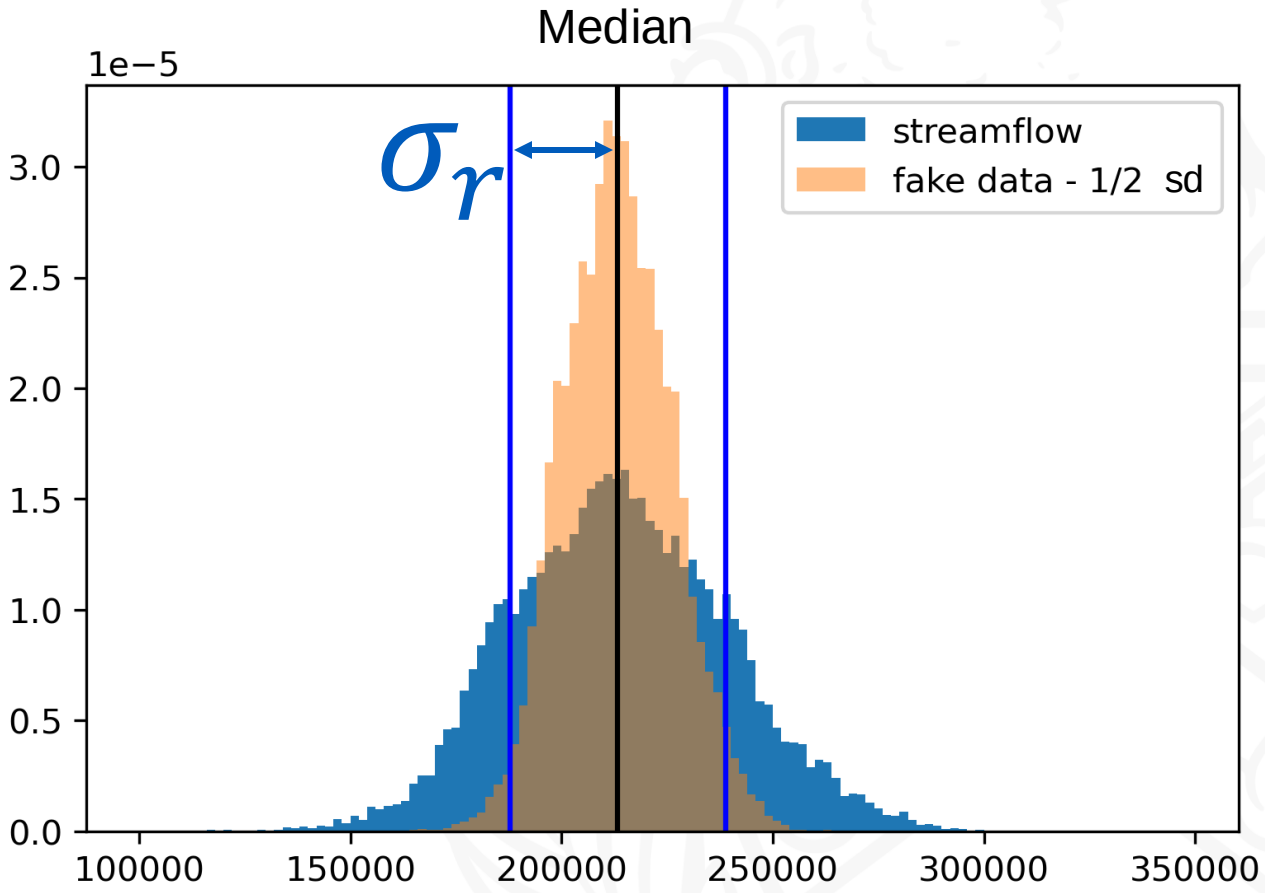


How do we evaluate the distribution of streamflow data?

- Probability Density Function

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

What does the PDF for our fake streamflow data look like?

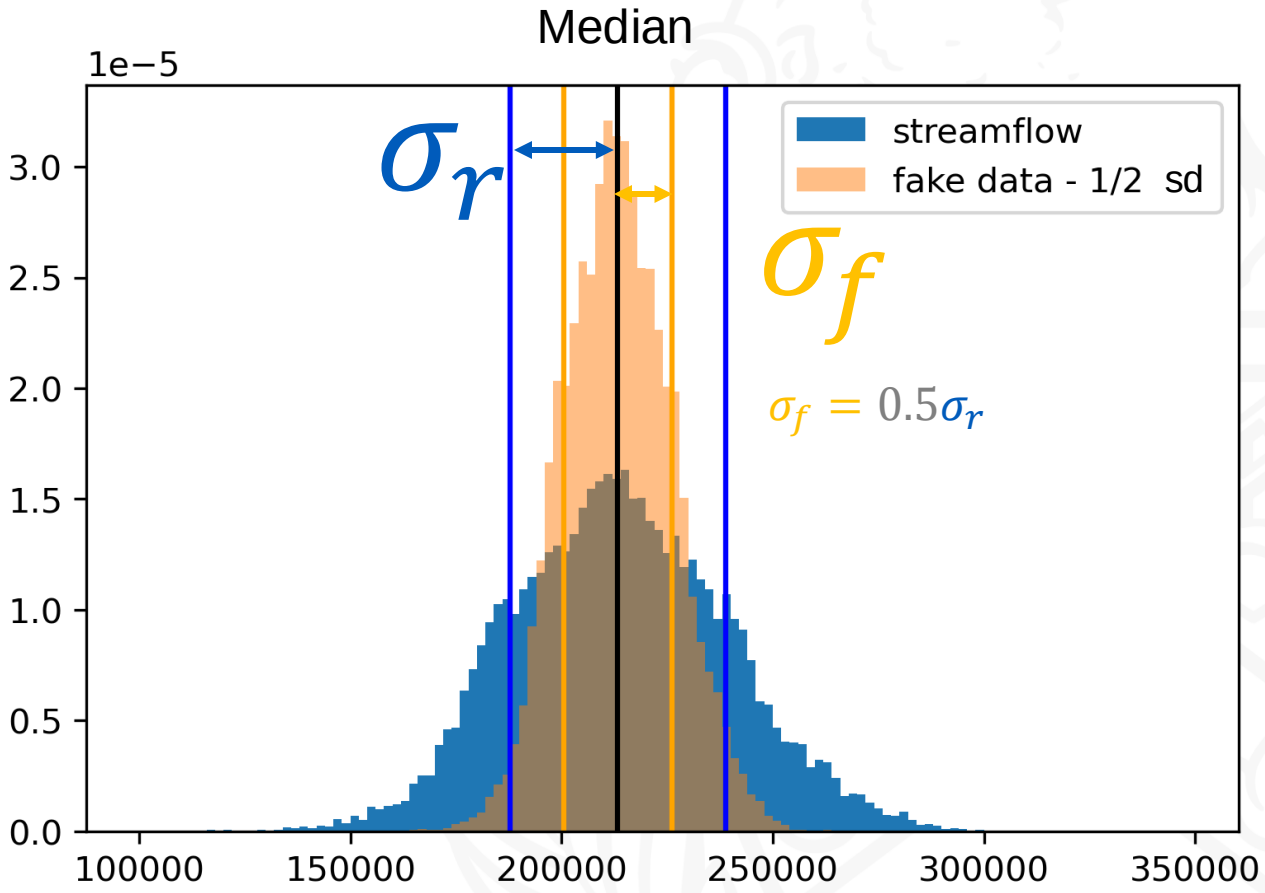


How do we evaluate the distribution of streamflow data?

- Probability Density Function

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

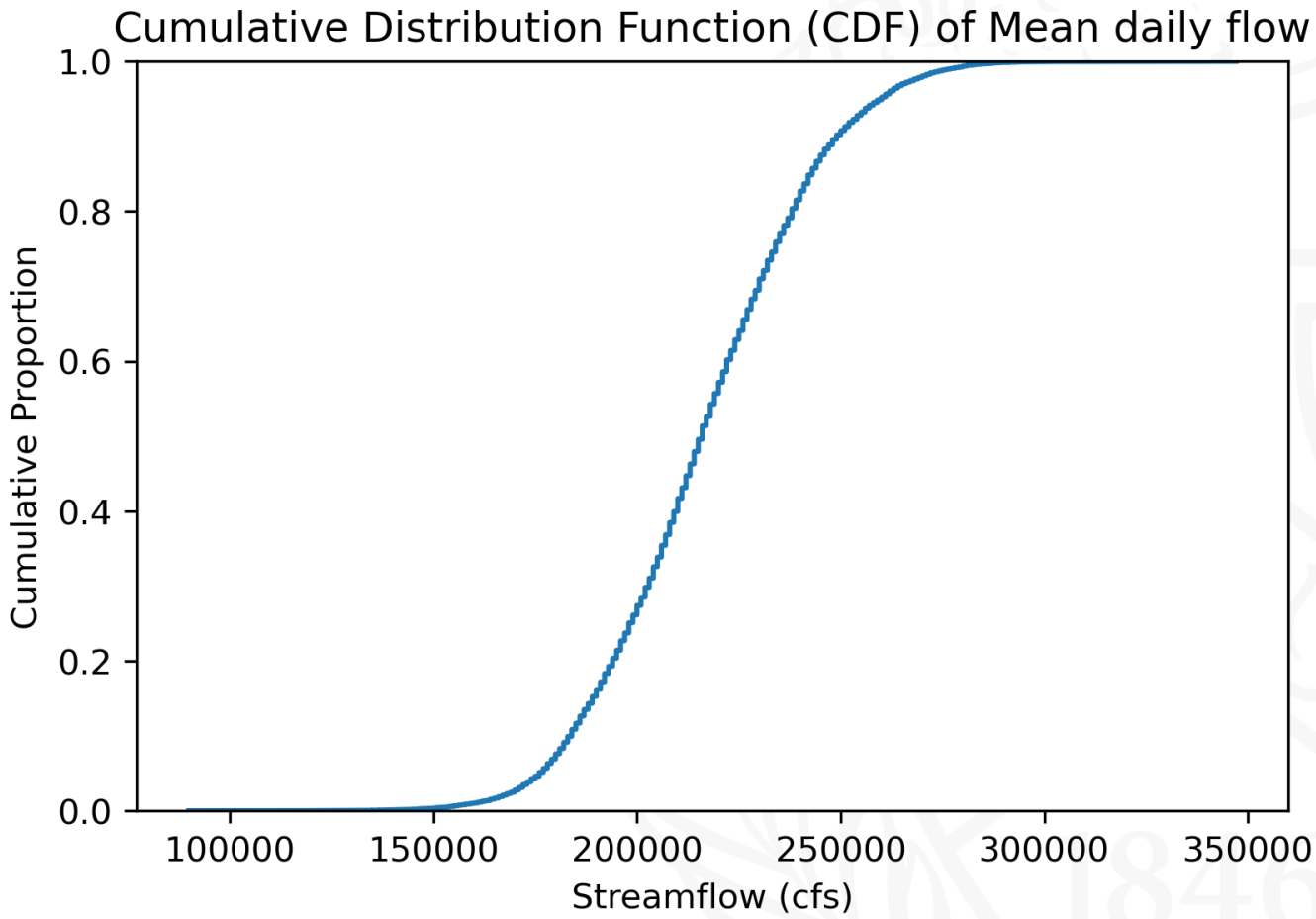
What does the PDF for our fake streamflow data look like?



How do we evaluate the distribution of streamflow data?

- Cumulative Distribution Functions

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

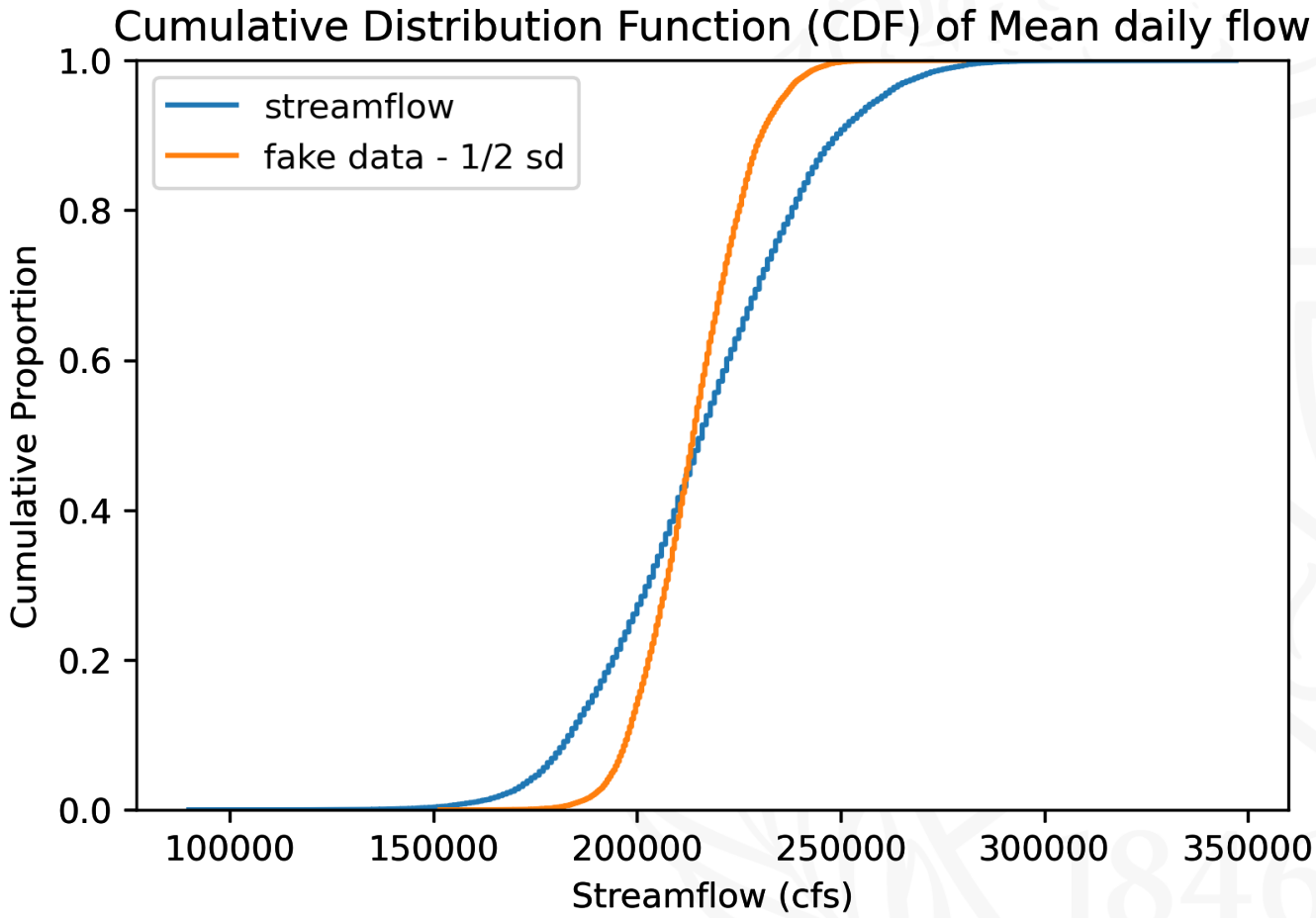


How do we evaluate the distribution of streamflow data?

- Cumulative Distribution Functions

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

What does the CDF for our fake streamflow data look like?

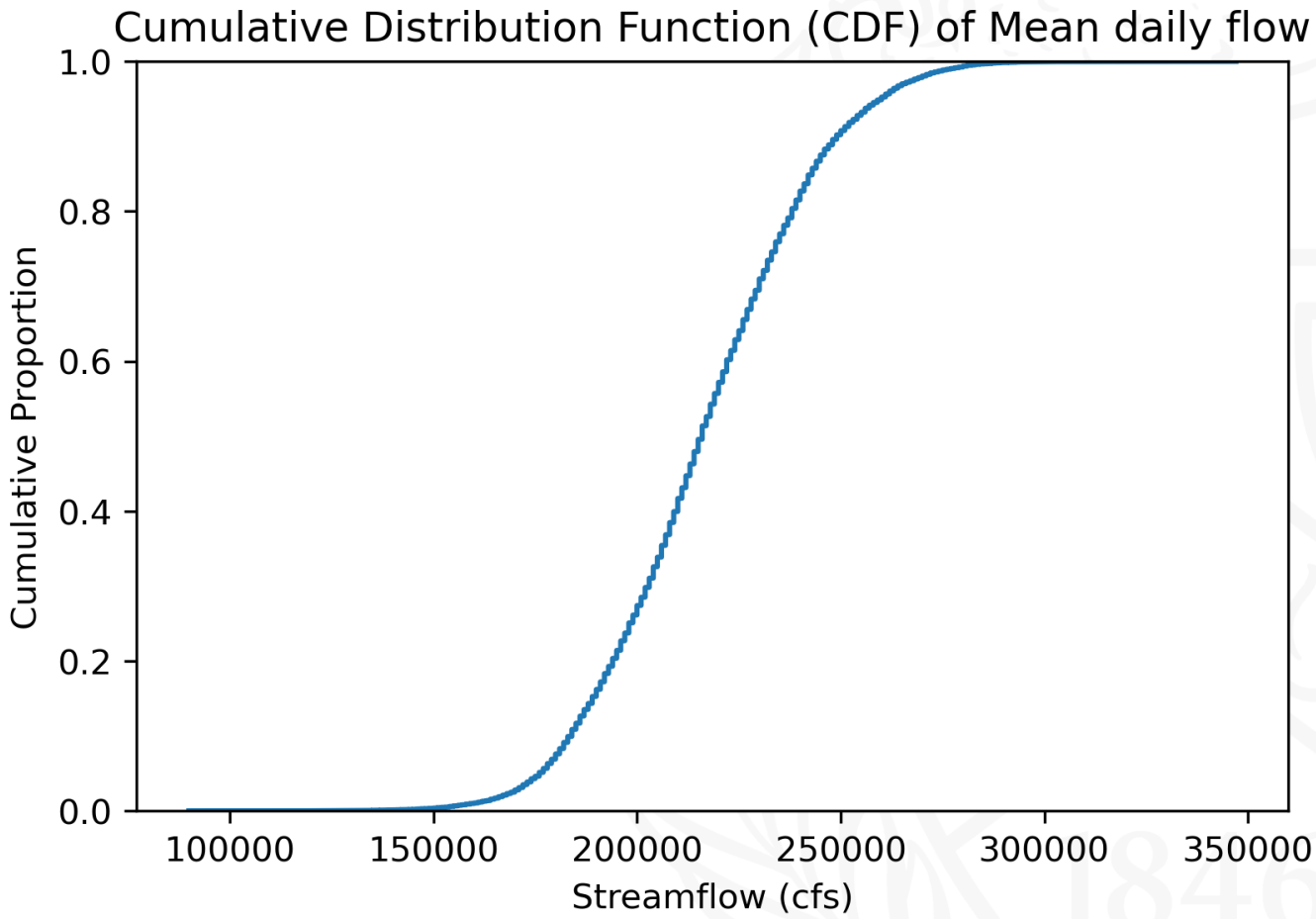


How do we evaluate the distribution of streamflow data?

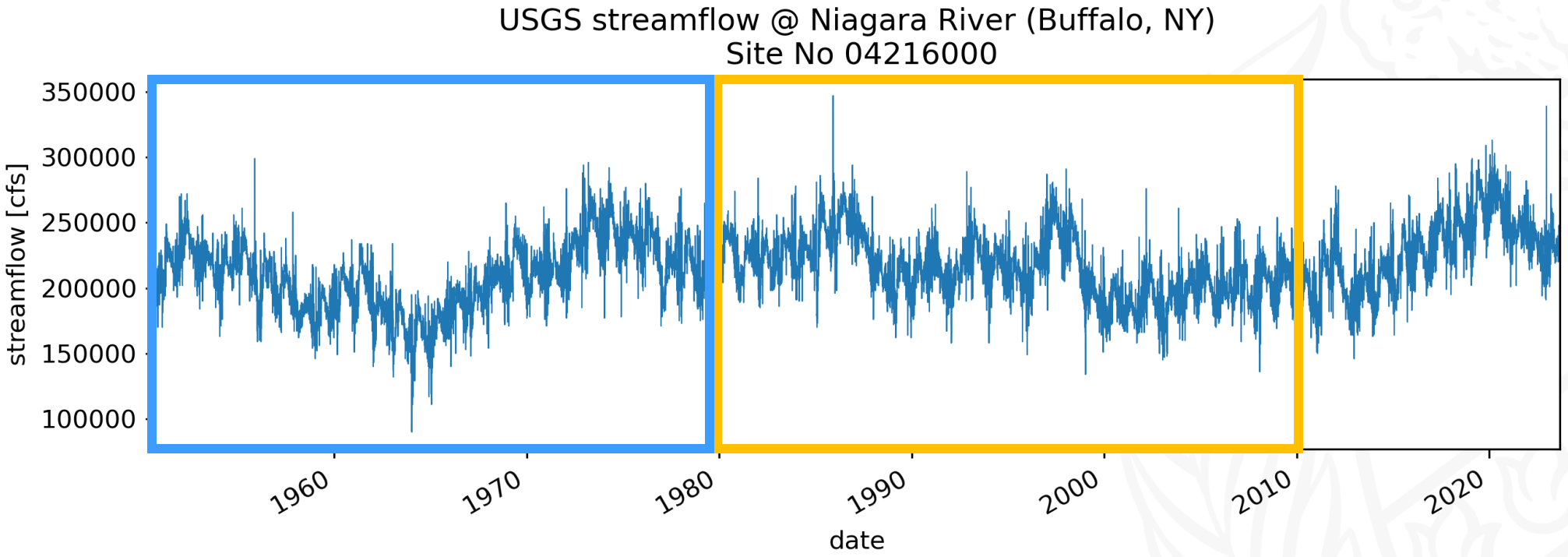
- Cumulative Distribution Functions

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

- In the analysis, CDF can be a powerful tool for visualizing the shifting of hydrologic regimes.



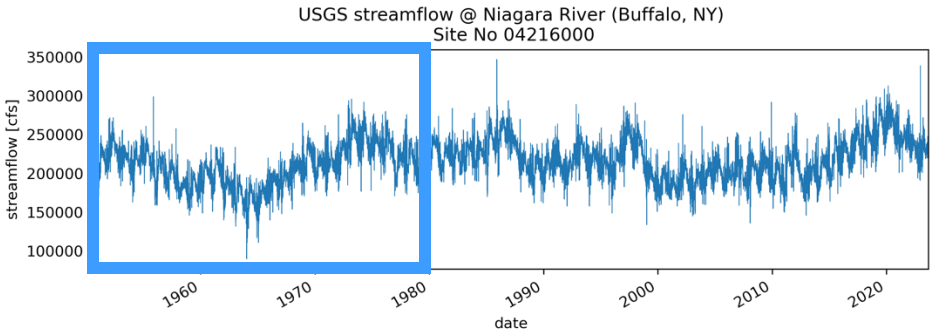
Use CDF to visualize the shift in hydrologic regimes



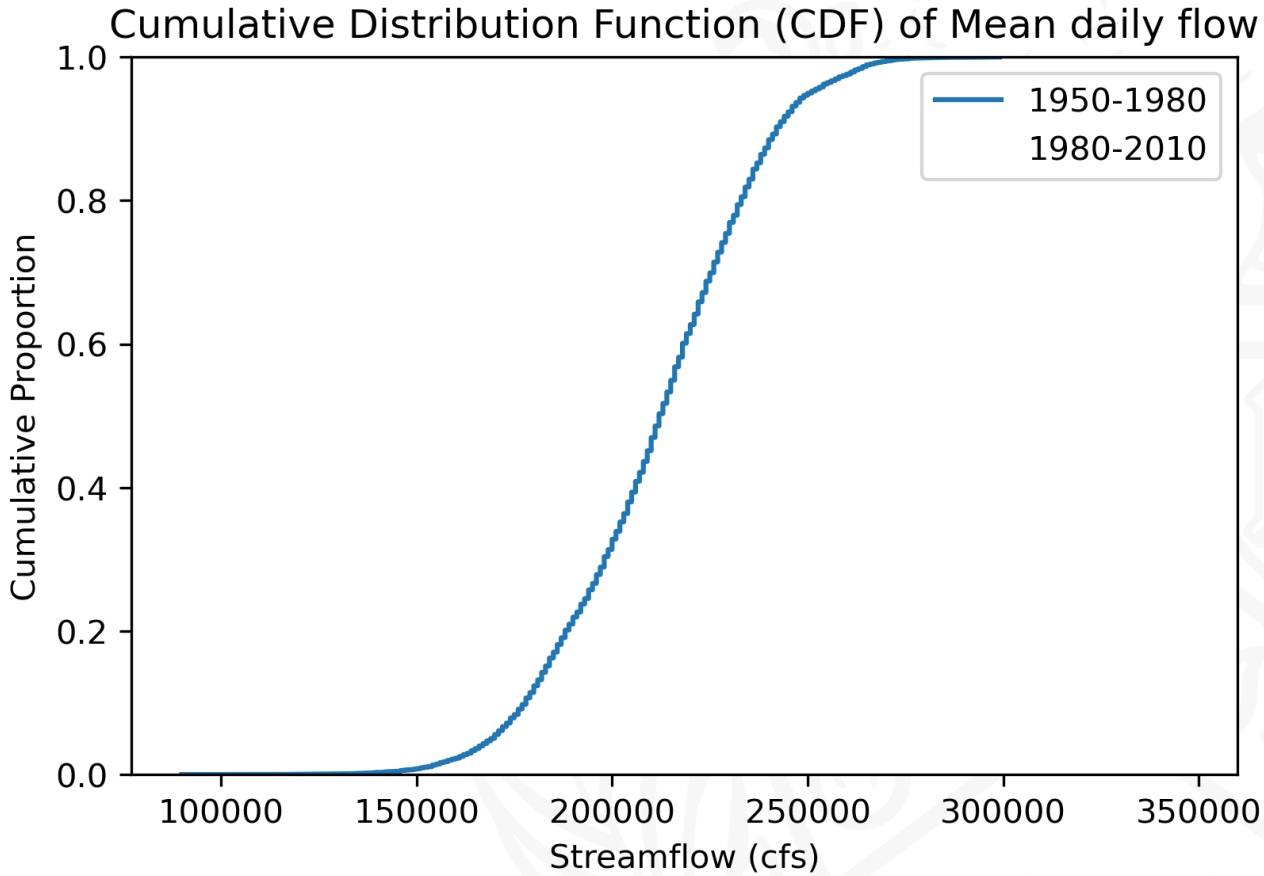
1950-1980

1980-2010

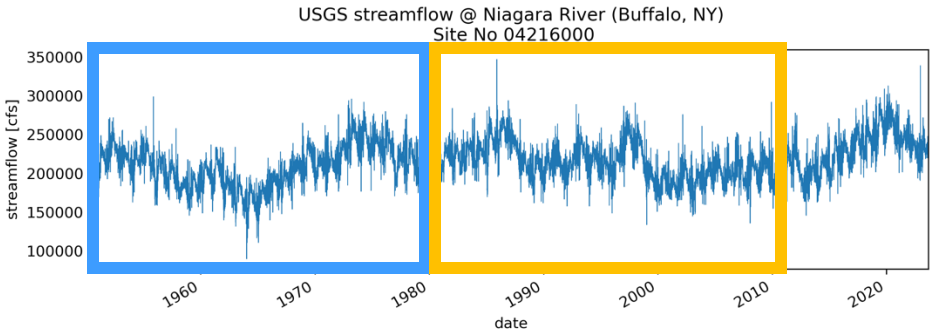
Use CDF to visualize the shift in hydrologic regimes



1950-1980 1980-2010

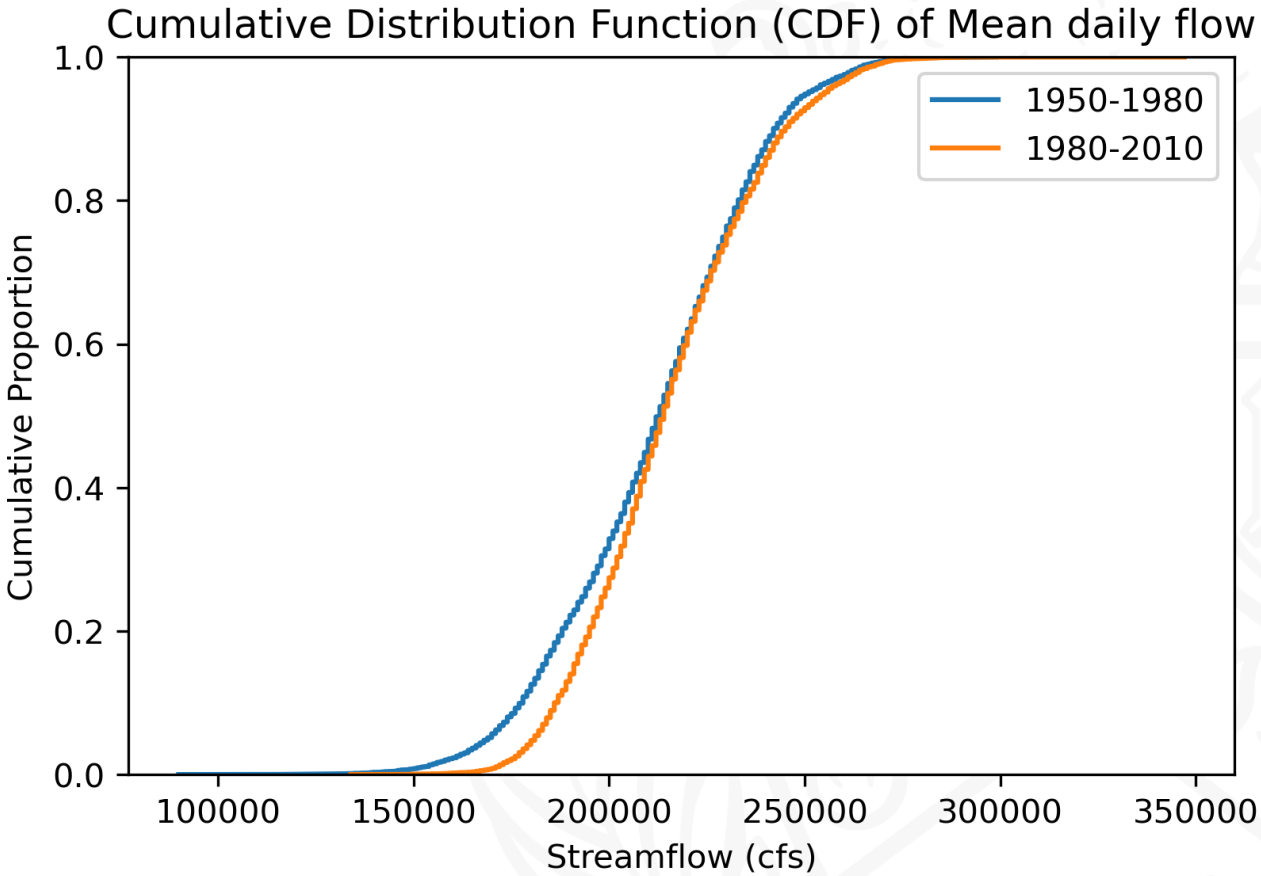


Use CDF to visualize the shift in hydrologic regimes



1950-1980 1980-2010

What information can we read from the plot in the right?



CDF for identifying floods?

New Data Reveals Hidden Flood Risk Across America

By Christopher Flavelle, Denise Lu, Veronica Penney, Nadja Popovich and John Schwartz June 29, 2020

Nearly twice as many properties may be susceptible to flood damage than previously thought, according to a new effort to map the danger.

Across much of the United States, the flood risk is far greater than government estimates show, new calculations suggest, exposing millions of people to a hidden threat — and one that will only grow as climate change worsens.

That new calculation, which takes into account sea-level rise, rainfall and flooding along smaller creeks not mapped federally, estimates that 14.6 million properties are at risk from what experts call a 100-year flood, far more than the 8.7 million properties shown on federal government flood maps. A 100-year flood is one with a 1 percent chance of striking in any given year.

What is a "100-year flood"?

Instead of the term "100-year flood" a hydrologist would rather describe *this extreme hydrologic event as a flood having a 100-year recurrence interval.*

Recurrence intervals

Recurrence intervals and probabilities of occurrences

Recurrence interval, years	Annual exceedance probability, percent
2	50
5	20
10	10
25	4
50	2
100	1
200	0.5
500	0.2

The term "100-year flood" is used to define a flow events that statistically has this same 1-percent chance of occurring.

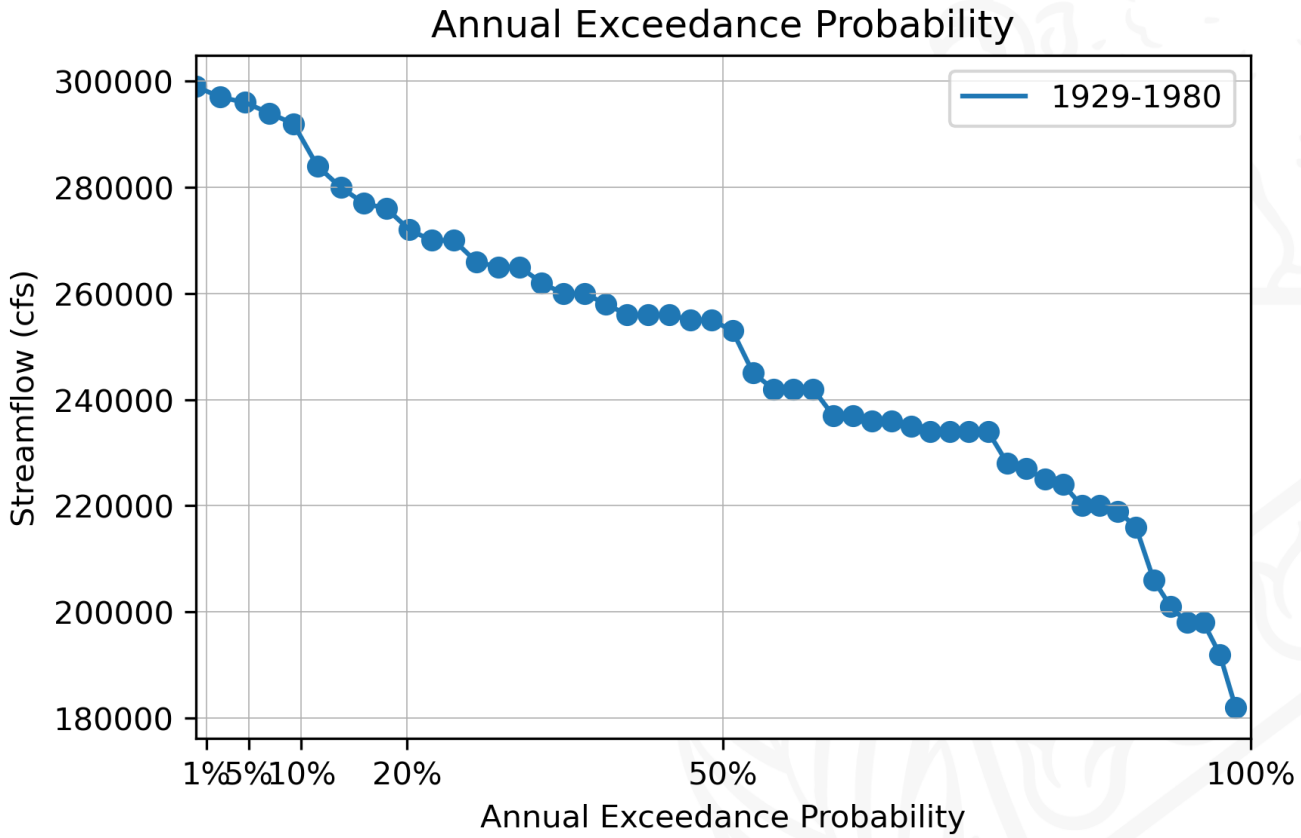
In other words, over the course of 1 million years, these events would be expected to occur 10,000 times.

But, just because it rained 10 inches in one day last year doesn't mean it can't rain 10 inches in one day again this year.

Annual exceedance probability (AEP)

- Step 1: Identify the peak flow for each year
- Step 2: Calculate the CDF for annual peak flow
- Step 3: AEP = 1 - CDF

If you were a hydrologist back in the 1980s, the construction of a bridge requires the information of 100-year flood

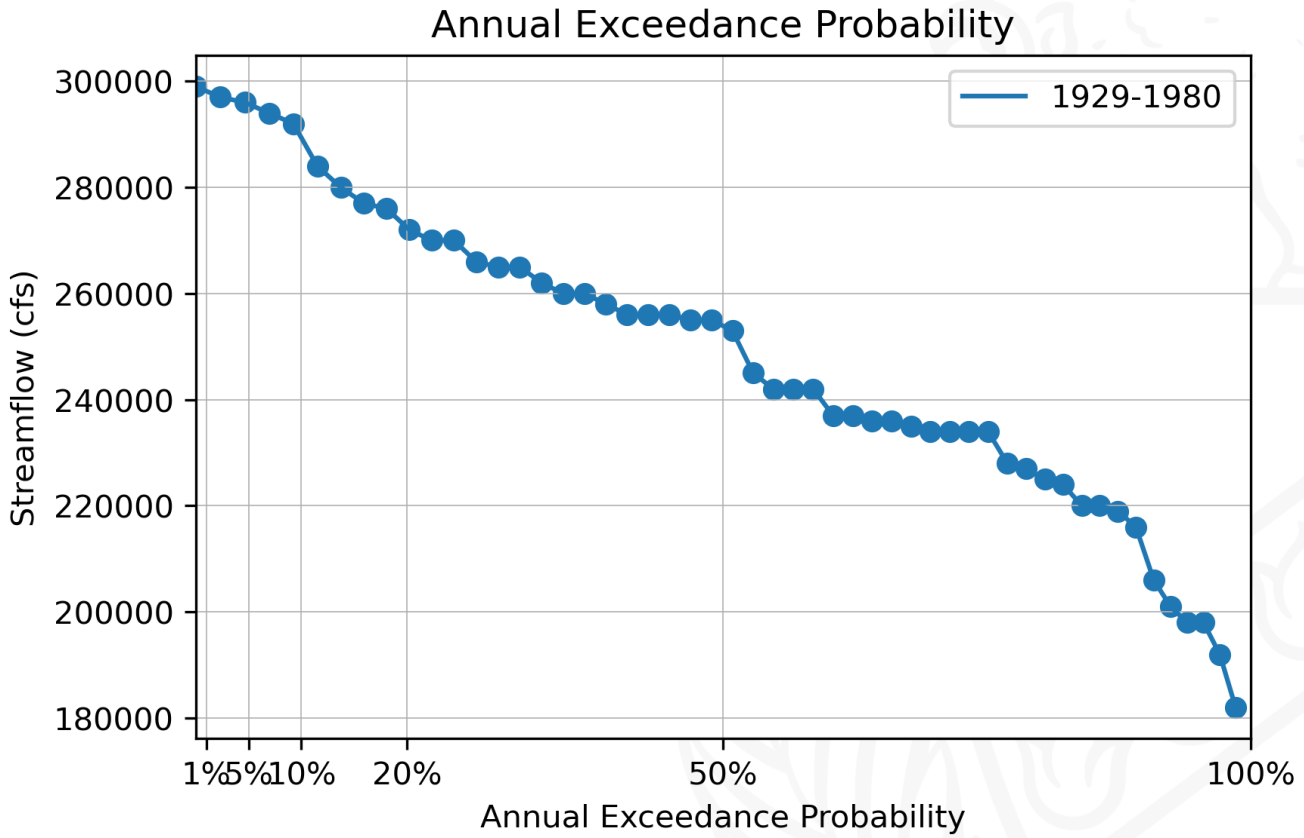


We can use the AEP plot on the right to retrieve the flood information, which corresponds to the value with **1% AEP!**

With more observations, it is important to revisit the flood designs

Now we are back to the 21st century, so we have observations for four more decades.

If we update the plots using more data, what have we observed?

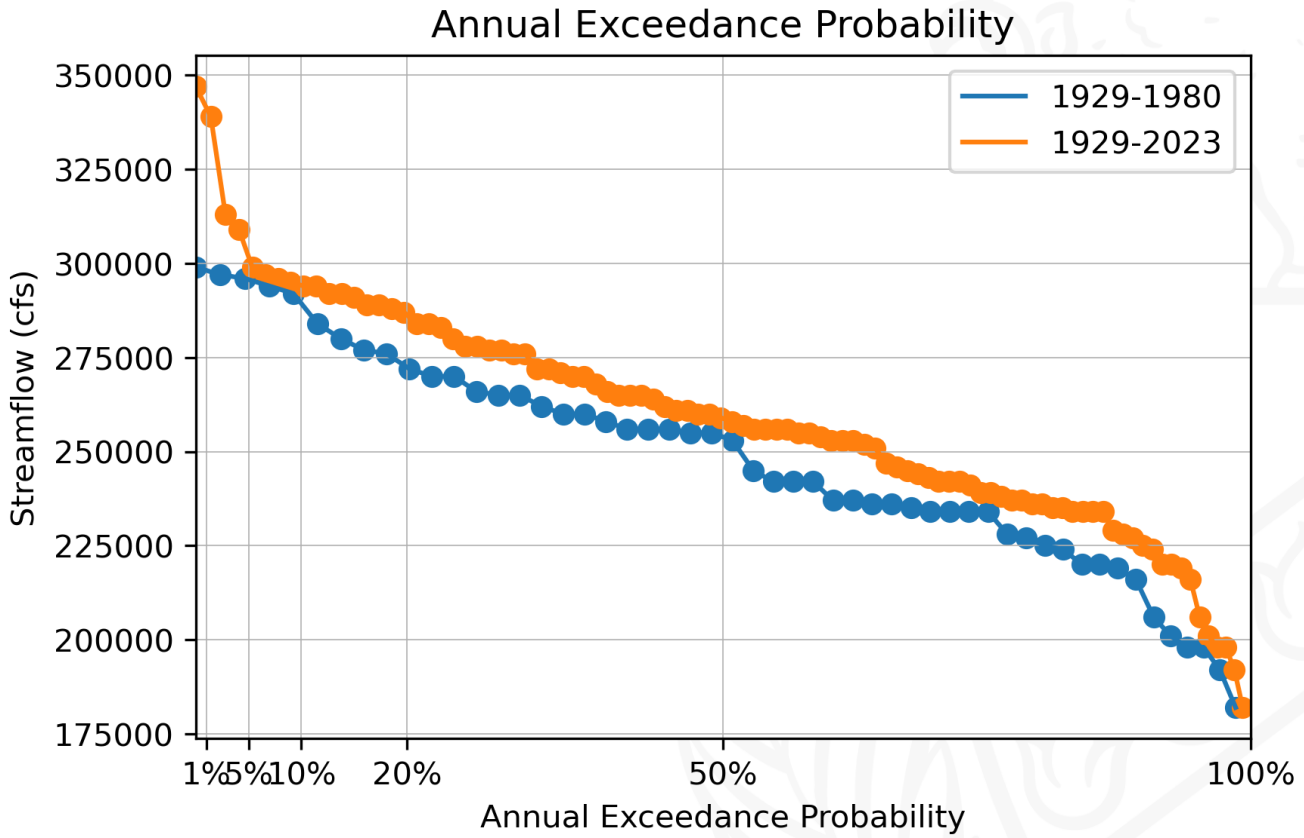


With more observations, it is important to revisit the flood designs

Now we are back to the 21st century, so we have observations for four more decades.

If we update the plots using more data, what have we observed?

The water volume of 100-year flood increased more than 10%!

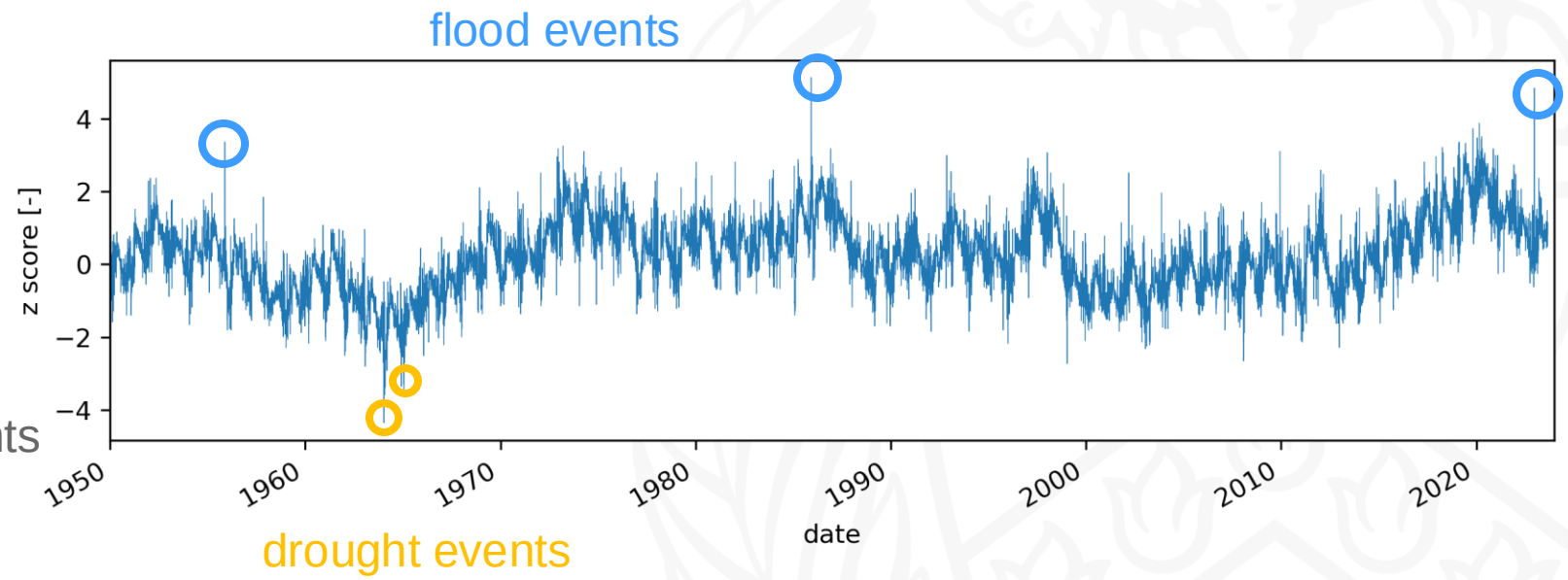


How can we more directly identify individual extreme events?

- **Z score**

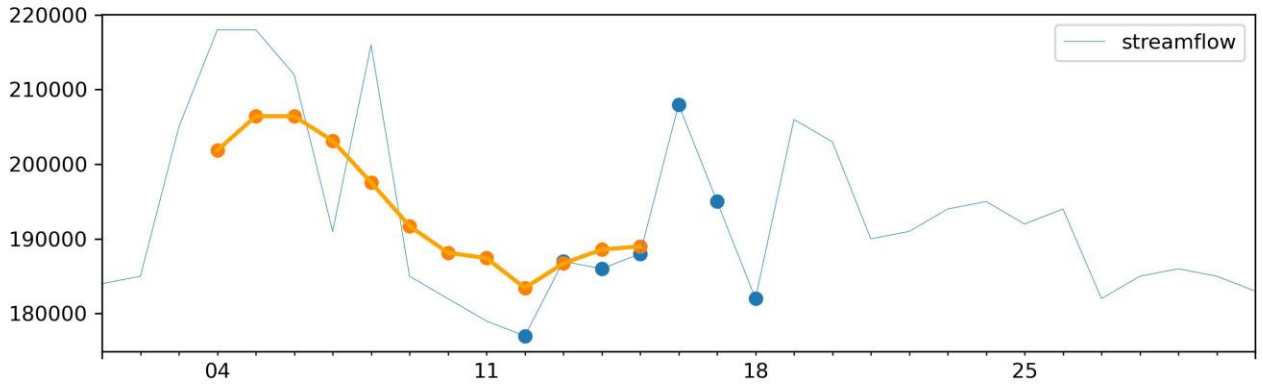
$$z = \frac{X_i - \bar{X}}{\sigma}$$

- Z scores will be assigned to every data points
- It is useful to identify extreme events across basins
- Not only useful for identify flood events, but also for low flow events



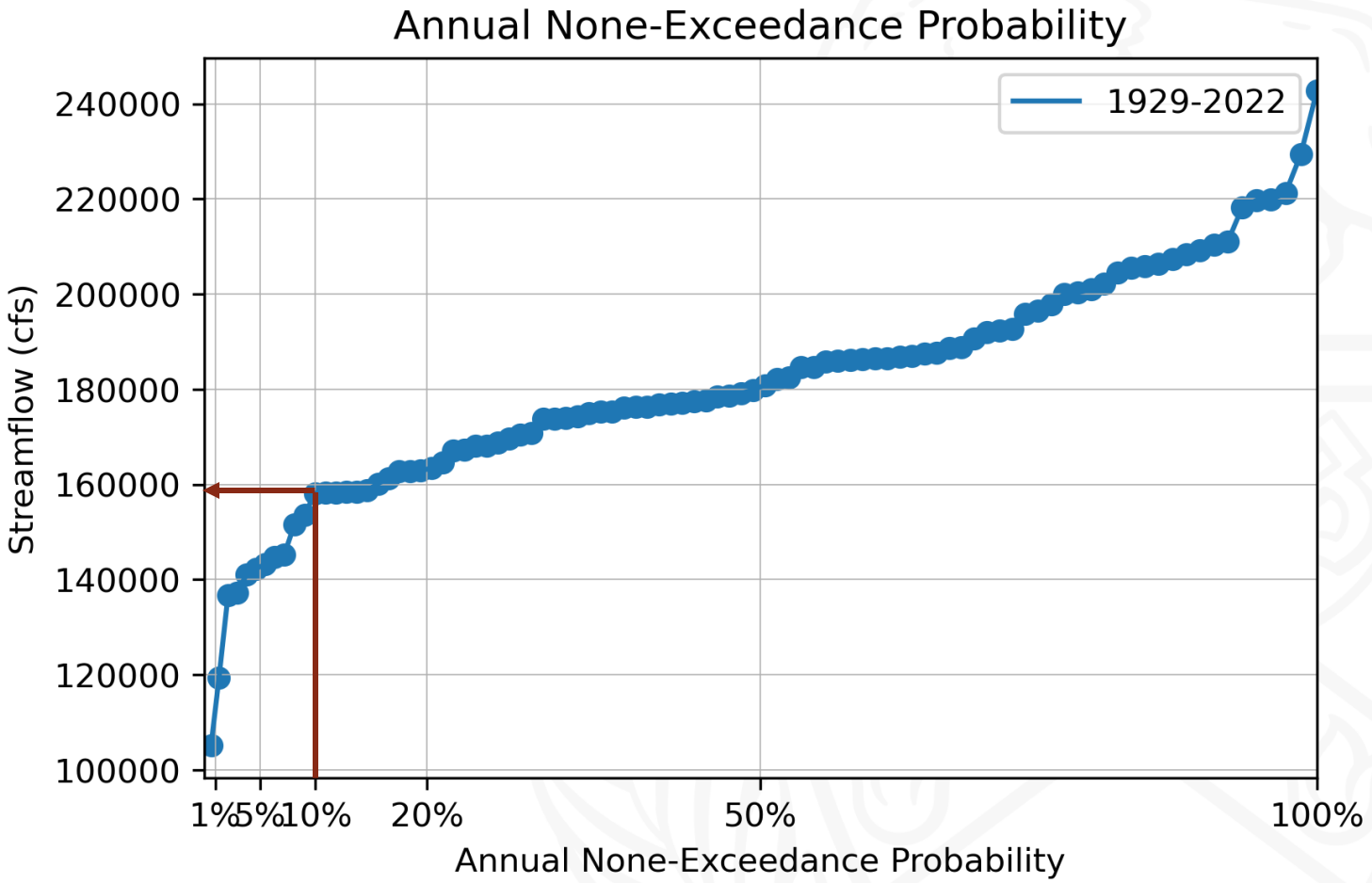
Statistical metrics for drought (low-flow events)?

- **7Q10**
  - The 7Q10 is the lowest 7-day average flow with a recurrence interval of 10 years.
- How do we calculate 7Q10?
  - Step 1: calculate the **7-day rolling average**



Statistical metrics for drought (low-flow events)?

- **7Q10**
  - The 7Q10 is the lowest 7-day average flow with a recurrence interval of 10 years.
- How do we calculate 7Q10?
  - Step 1: calculate the **7-day rolling average**
  - Step 2: Find annual minimum 7-day average flow
  - Step 3: Annual Non-exceedance probability



Recap

- Basic statistics
  - Mean, variance, standard deviation
  - Boxplot (Interquartile range, IQR)
  - PDF, CDF, AEP (recurrence probability)
  - Identify extreme events (z score, 7Q10)


1846