PANDAS DATAFRAME & BASIC STATISTICS

ERT 474/574

Open-Source Hydro Data Analytics

Sep 17th 2025

University at Buffalo The State University of New York

Logistics

- HW2 due today

In the last lecture, we introduced **Numpy** (a powerful tool to generate data arrays and calculations)

How can we more effectively **manipulate** data?

```
import pandas as pd
```

Data Structure

Powerful functions

Data Structure

Data Structure	Dimension
Series	1
Data Frames	2

What is the difference between series and Data Frames?

Data Series

```
tom       105
bob       306
nancy    3560
dan      1200
eric       50
```

Data Framework

	Fav_number	Fav_color
tom	105	red
bob	306	blue
nancy	3560	orange
dan	1200	pink
eric	50	green

DataFrame syntax

column

Row

	Fav_number	Fav_color
tom	105	red
bob	306	blue
nancy	3560	orange
dan	1200	pink
eric	50	green

DataFrame syntax

Column name

index

	Fav_number	Fav_color
tom	105	red
bob	306	blue
nancy	3560	orange
dan	1200	pink
eric	50	green

DataFrame syntax

df.loc[index, column name]

Column name

index

	Fav_number	Fav_color
tom	105	red
bob	306	blue
nancy	3560	orange
dan	1200	pink
eric	50	green

DataFrame syntax

df.loc['tom','Fav_number']

Column name

index

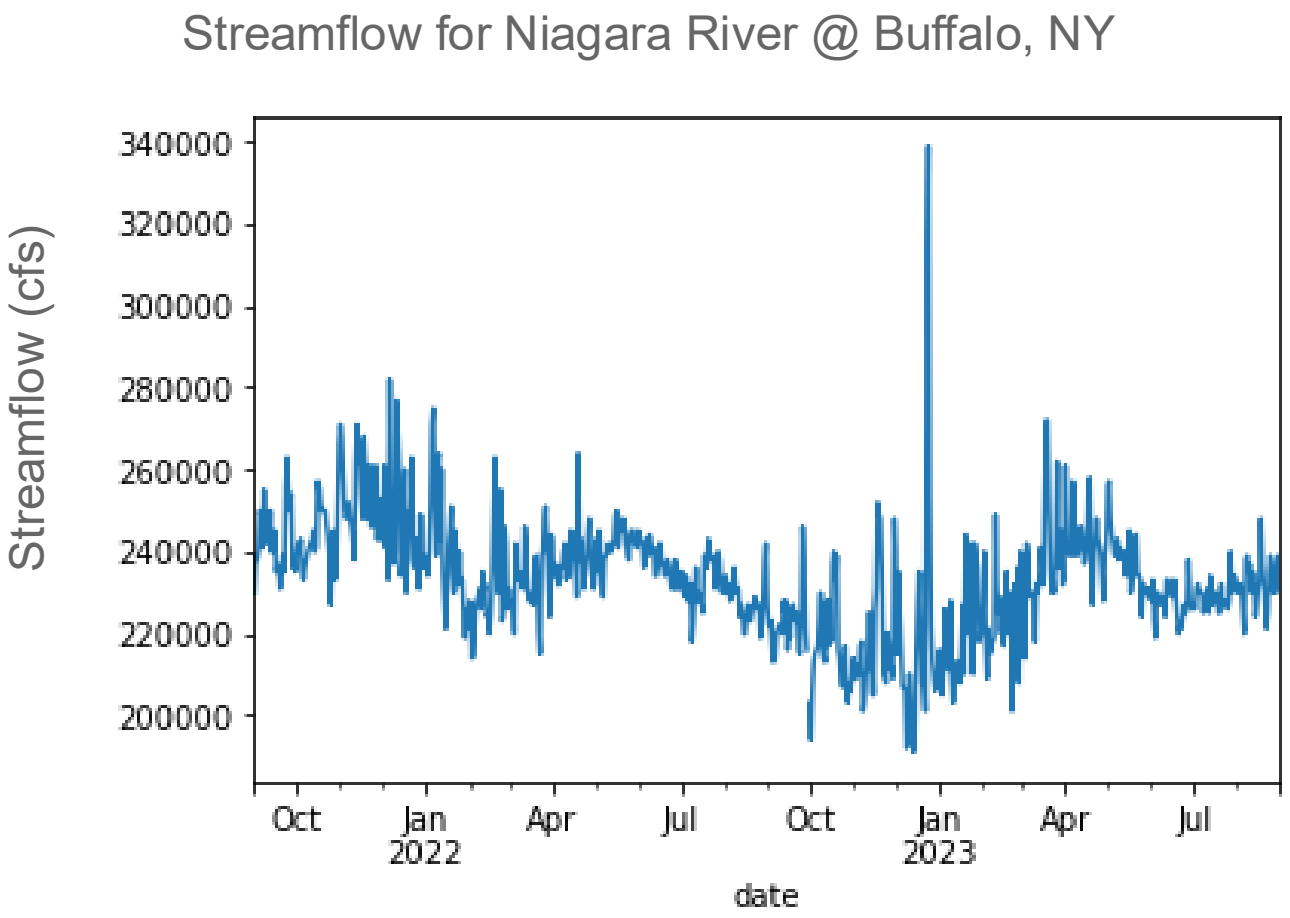
	Fav_number	Fav_color
tom	105	red
bob	306	blue
nancy	3560	orange
dan	1200	pink
eric	50	green

DataFrame syntax

df.iloc[0,0]

	0th column	1st column
	Fav_number	**Fav_color**
0th row **tom**	105	red
1st row **bob**	306	blue
2nd row **nancy**	3560	orange
3rd row **dan**	1200	pink
4th row **eric**	50	green

Data manipulation for time series data



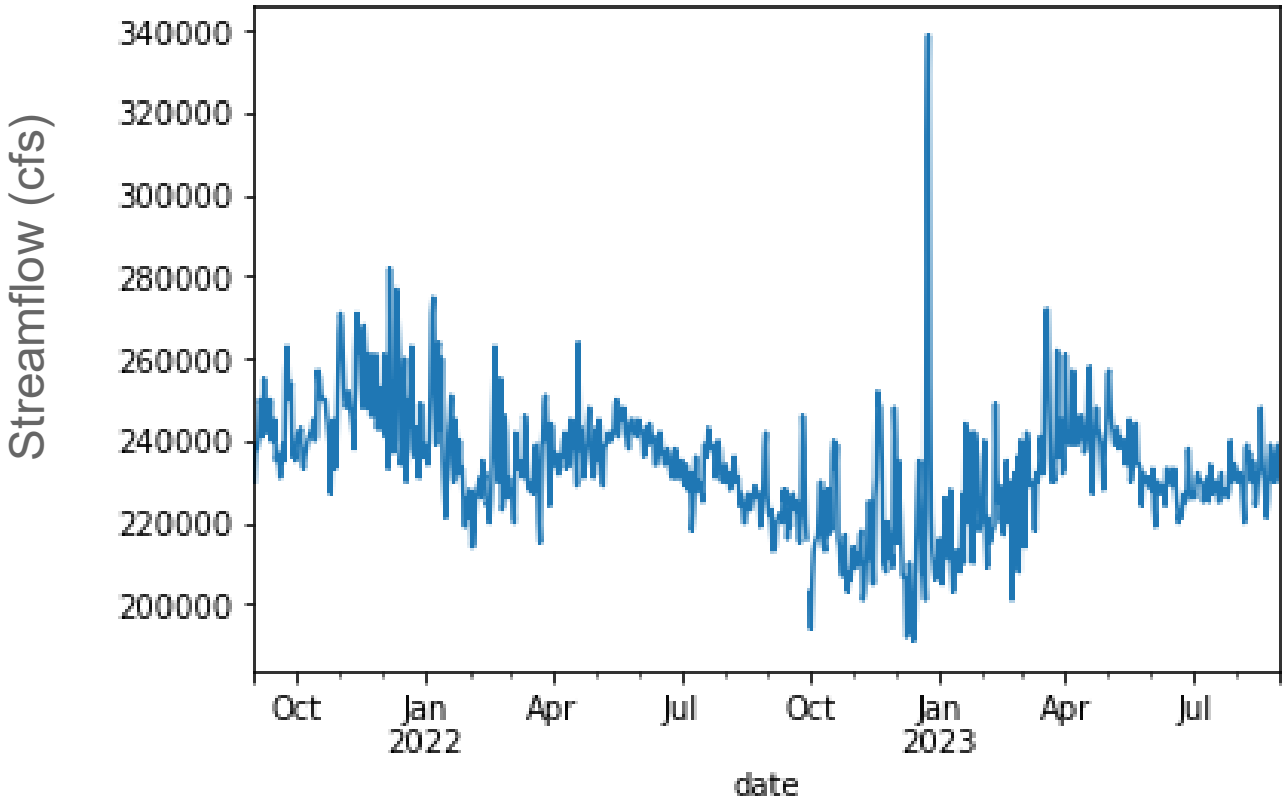
Out[70]:

date	streamflow	quality_flag
2022-10-01	203000.0	A
2022-10-02	194000.0	A
2022-10-03	206000.0	A
2022-10-04	213000.0	A
2022-10-05	215000.0	A
...	...	...
2022-12-28	209000.0	A
2022-12-29	206000.0	A
2022-12-30	207000.0	A
2022-12-31	211000.0	A
2023-01-01	216000.0	A

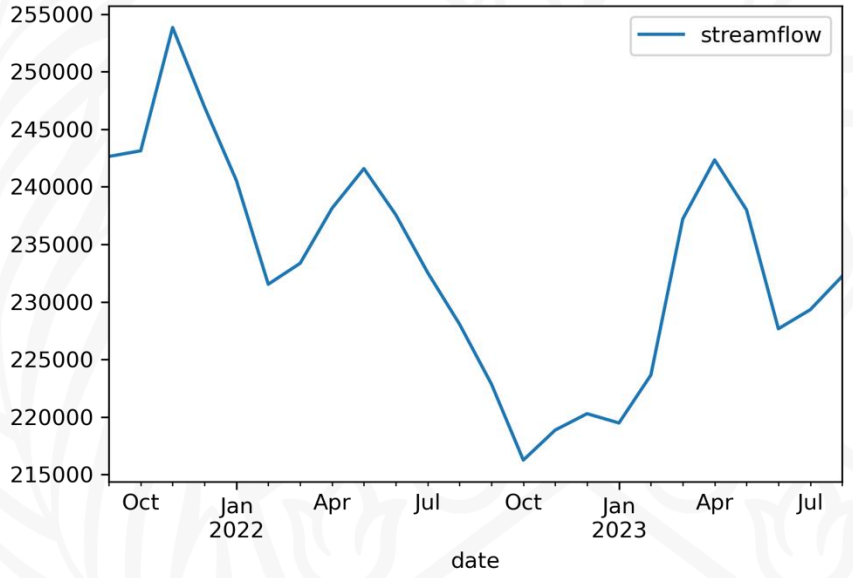
93 rows × 2 columns

Data manipulation for time series data

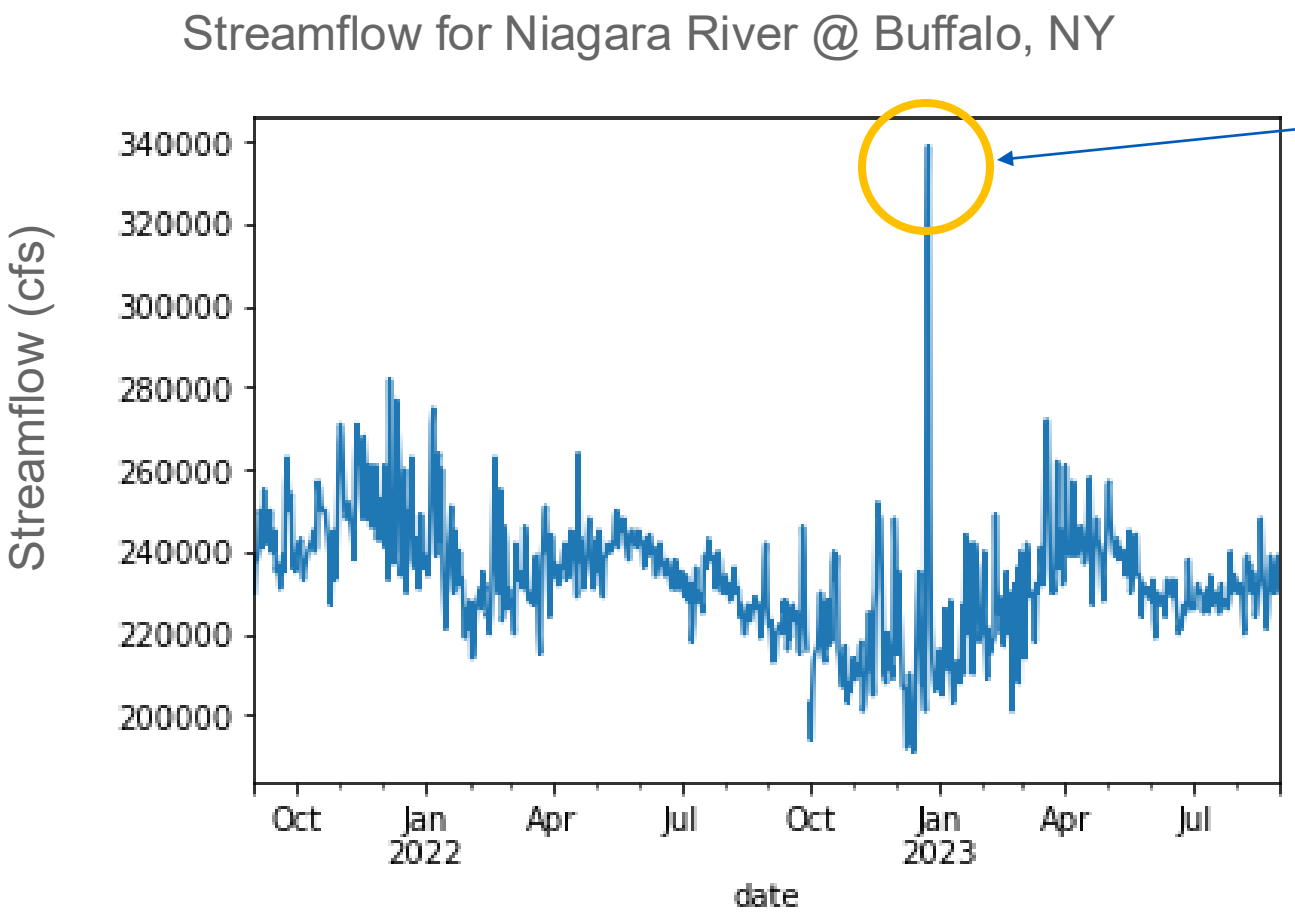
Streamflow for Niagara River @ Buffalo, NY



How can we change the frequency of data from daily to monthly?



Data manipulation for time series data



How can we identify the extreme high flow events?

When is that event?

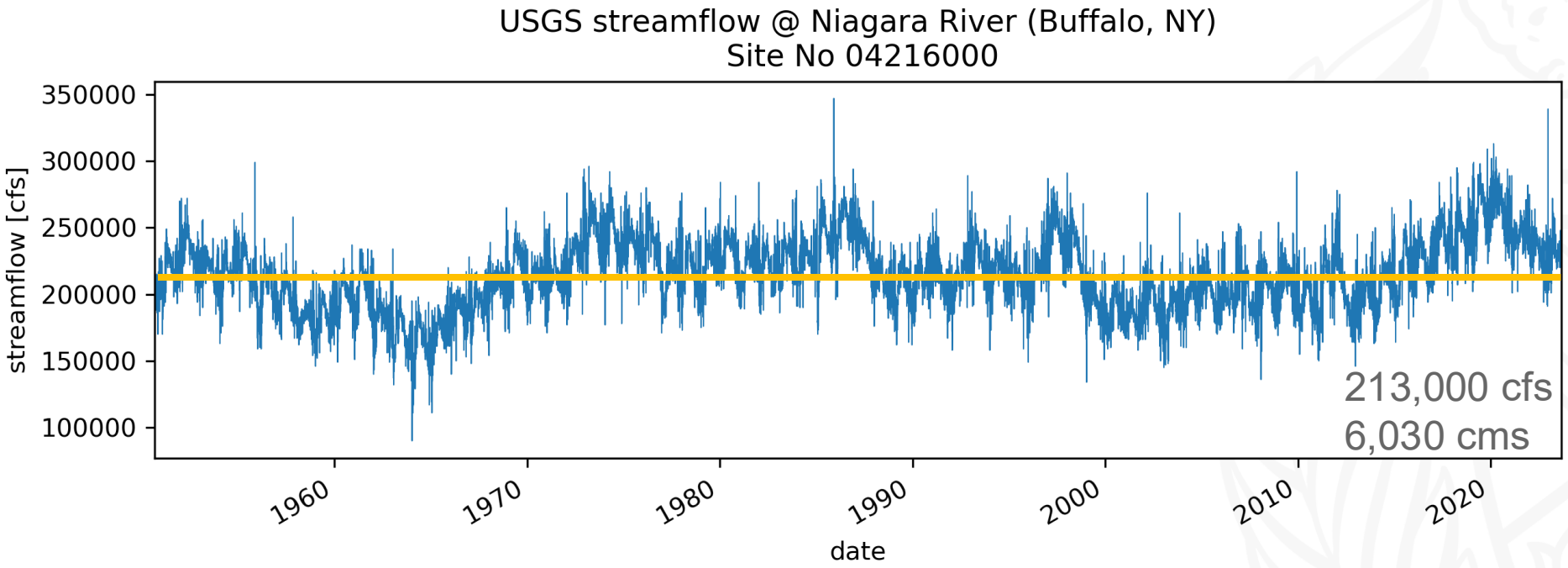
We will practice more in today's lab session!

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?

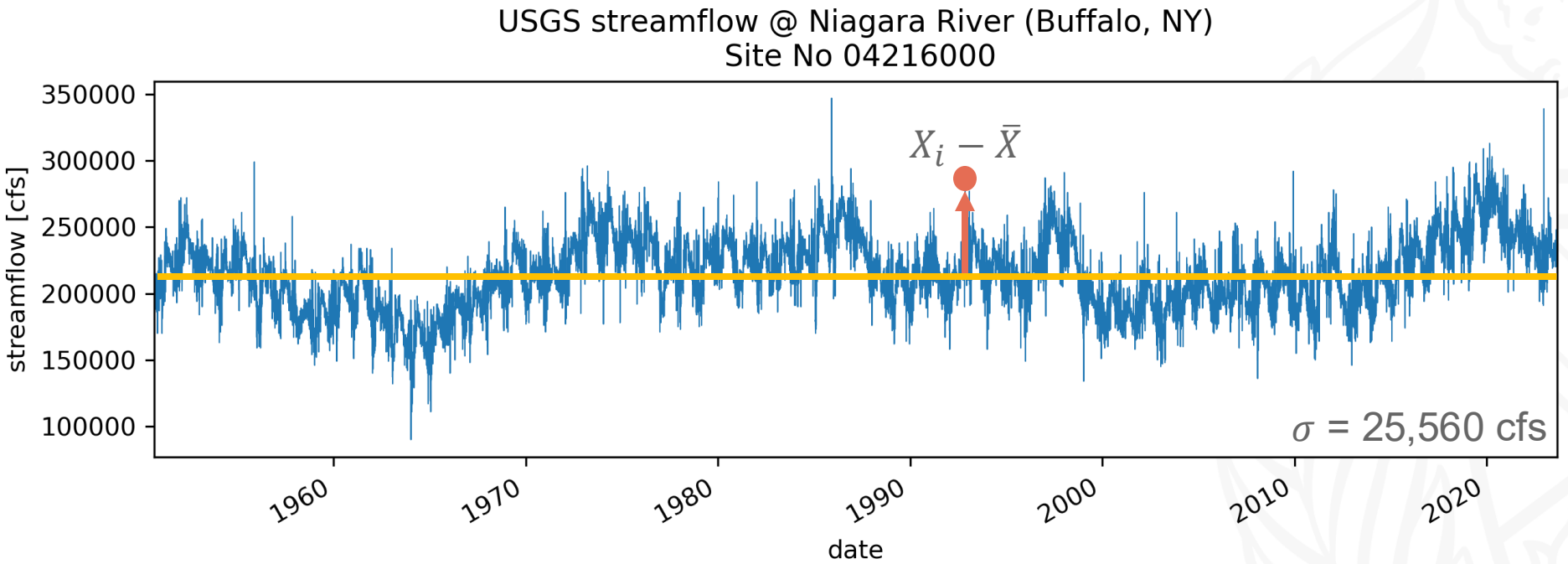


1. Sample Mean

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

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

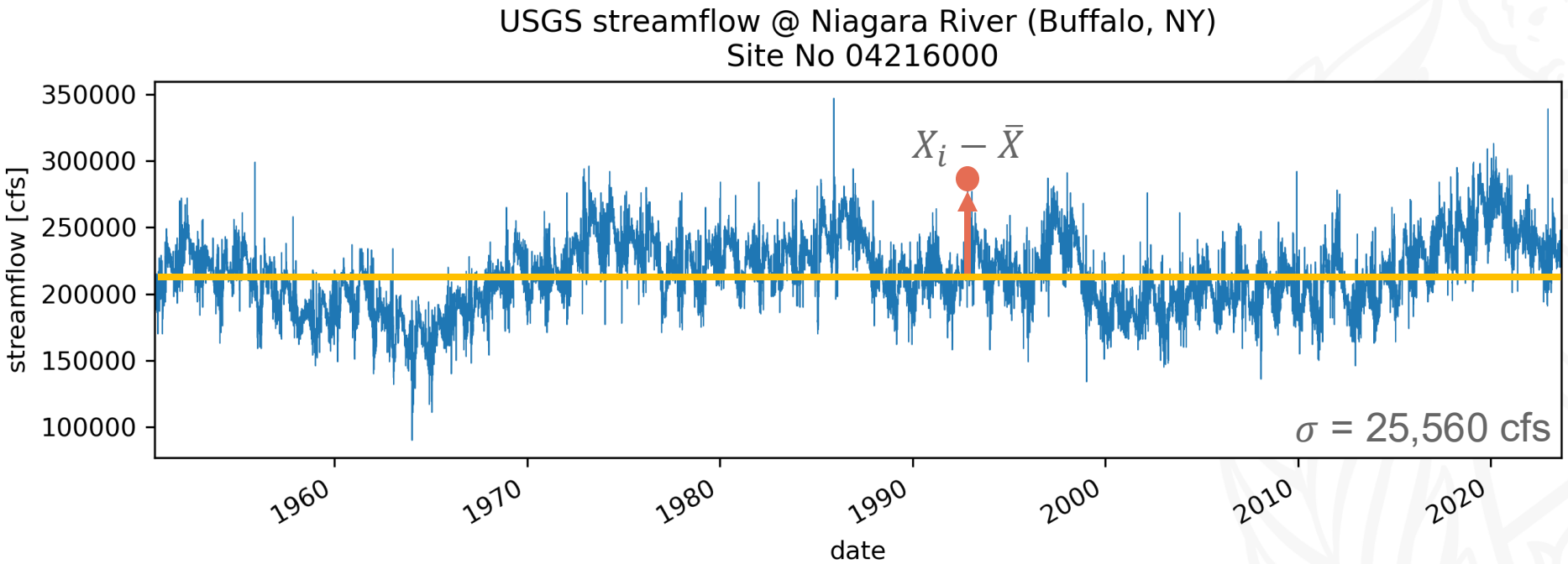
$$\sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

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

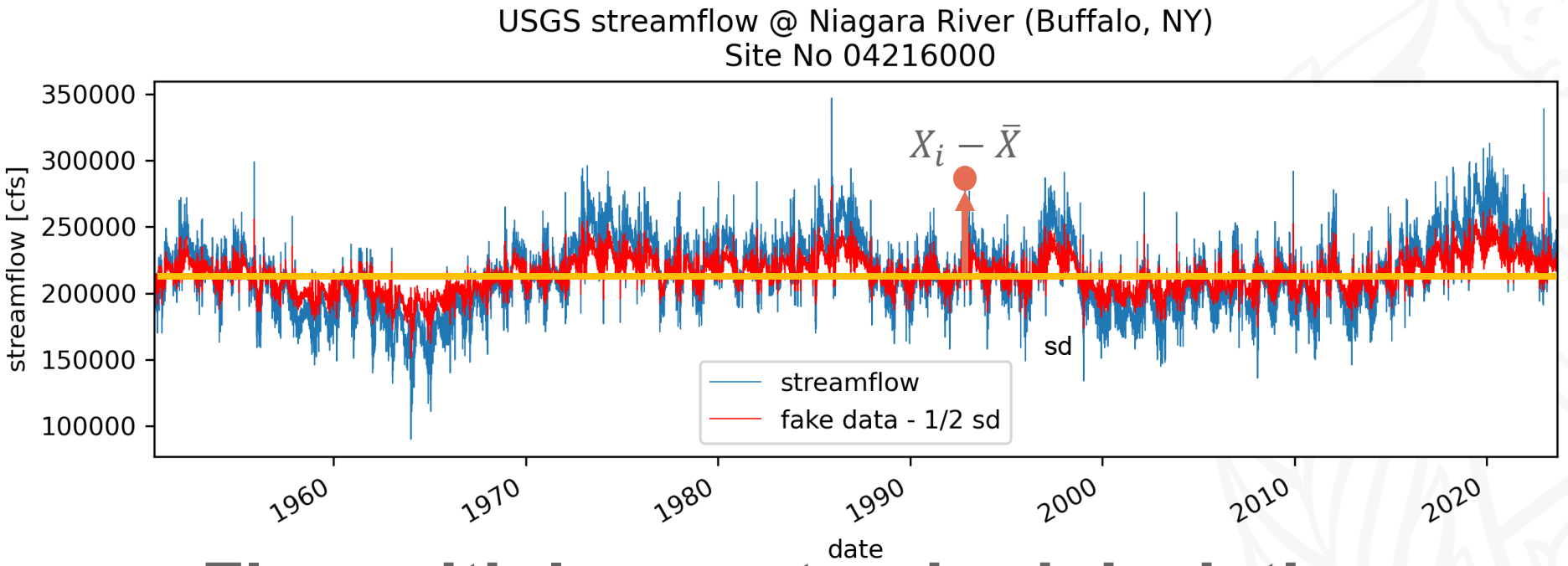
If you were a state hydrologist, when you were asked to give a high-level introduction to Niagara Rivers at Buffalo, what information would you provide based on the streamflow observations?



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

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

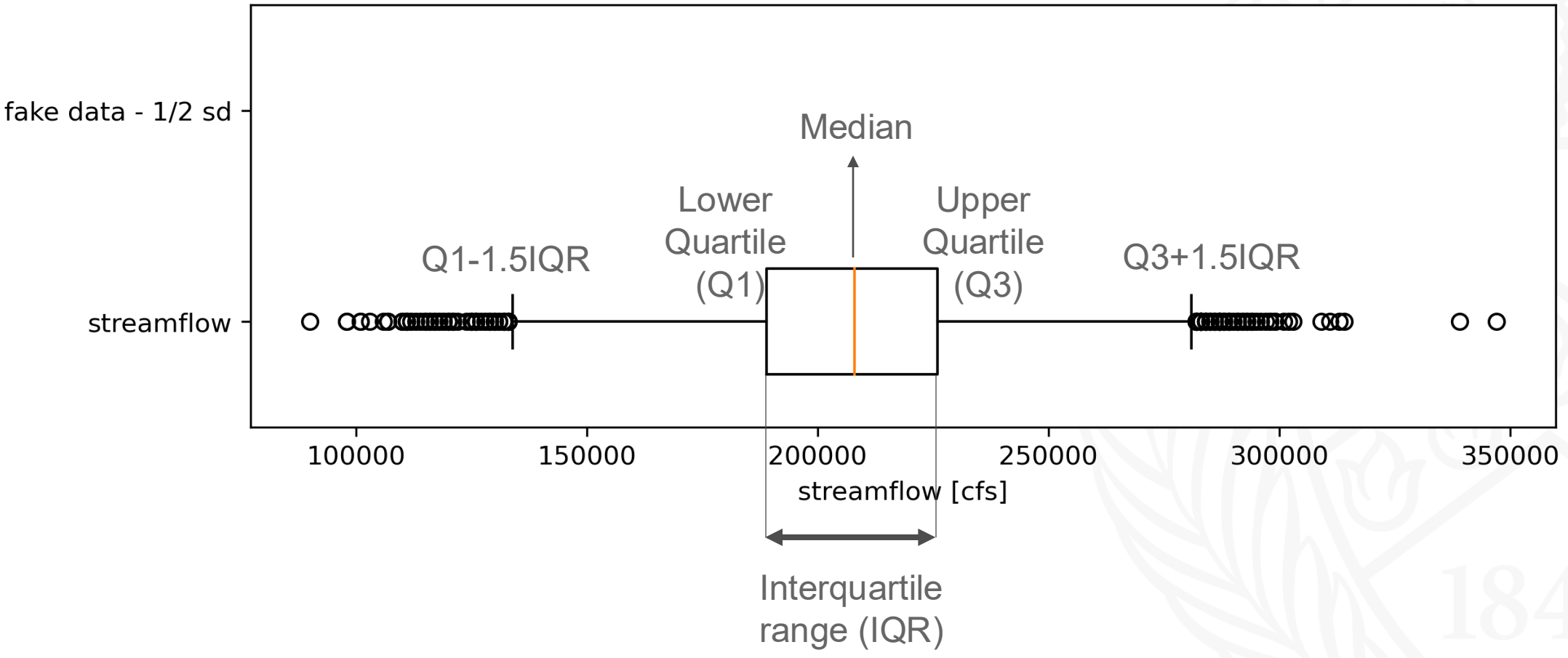
If you were a state hydrologist, when you were asked to give a high-level introduction to Niagara Rivers at Buffalo, what information would you provide based on the streamflow observations?



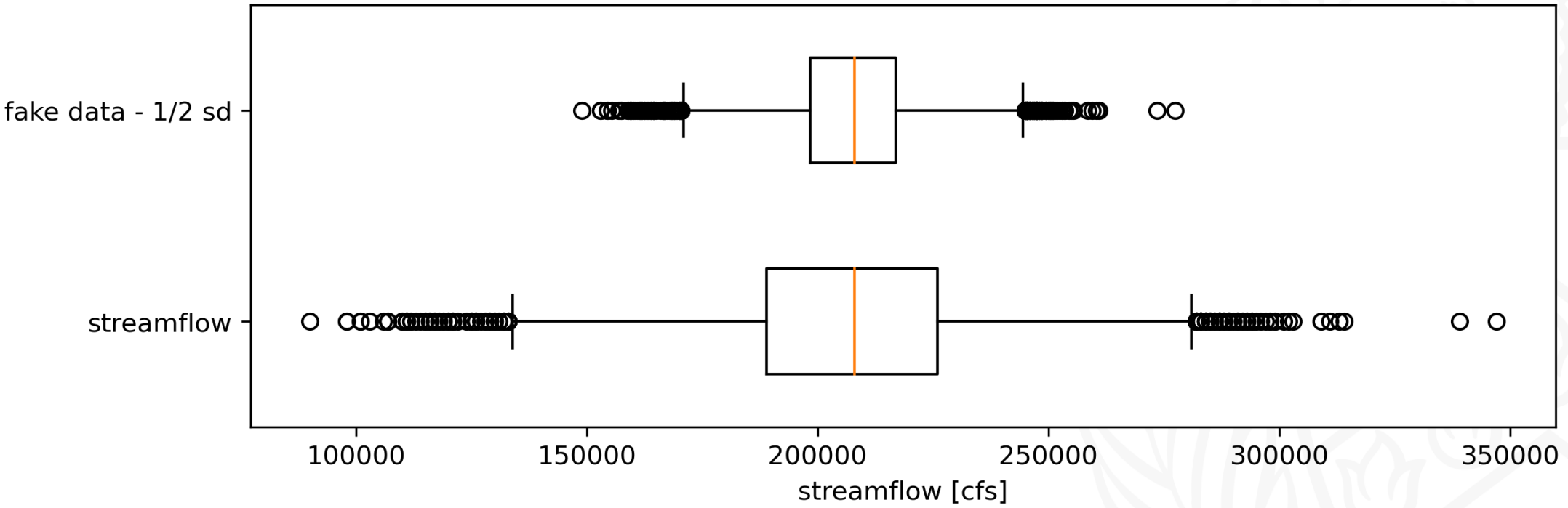
A more straightforward way to visualize the spread of the dataset?

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)

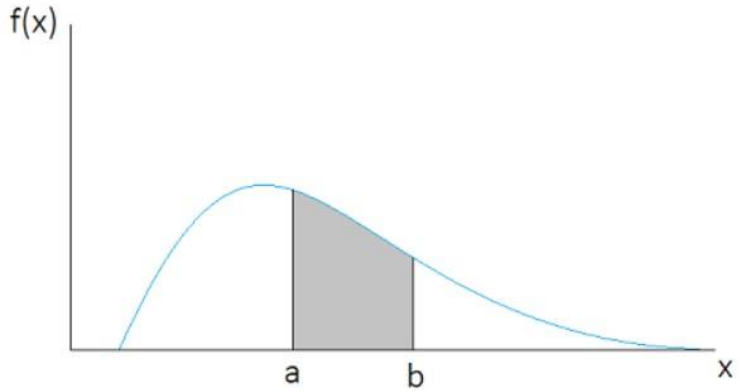
Box-plot


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

How do we evaluate the distribution of streamflow data?

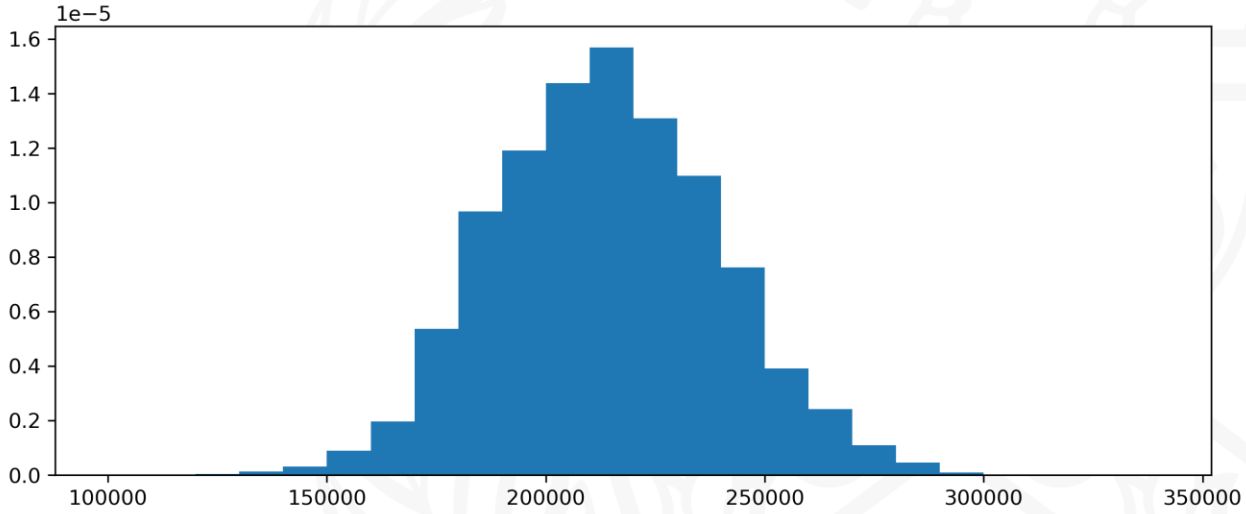
- Probability Density Function

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



The total area below a PDF is 1.

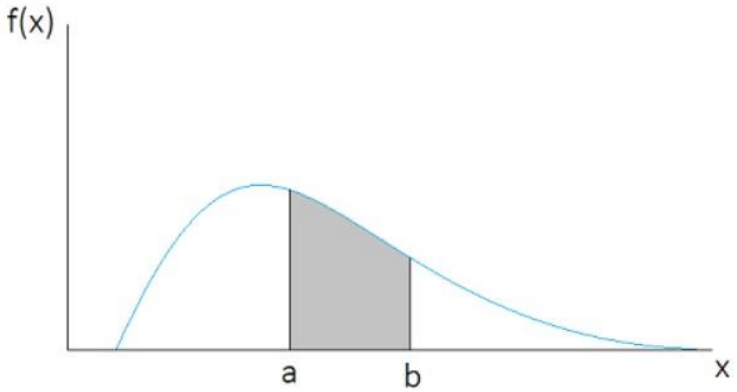
We can use histogram to visualize the PDF for time series data.



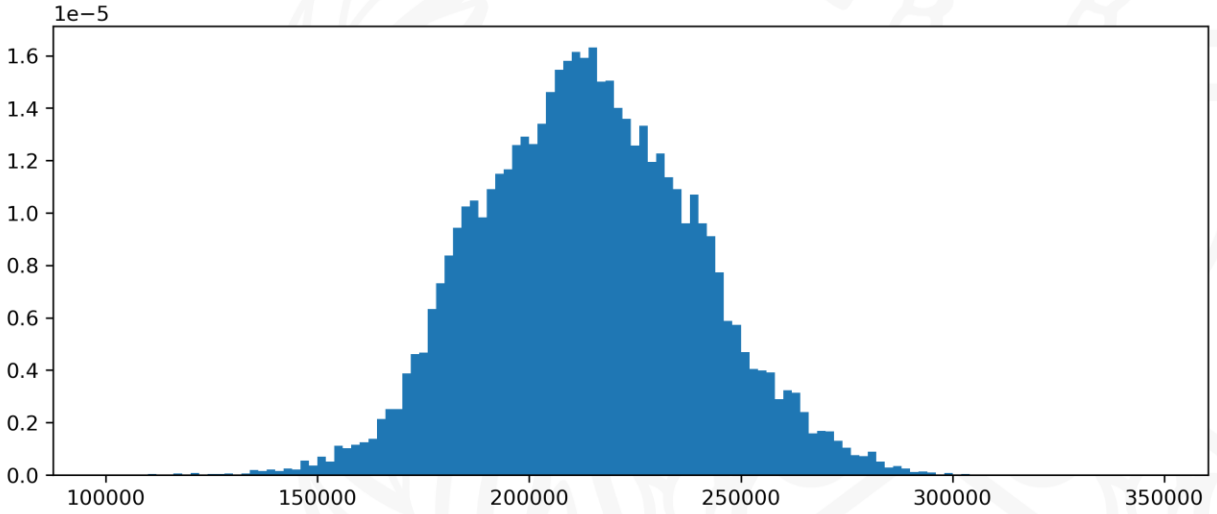
How do we evaluate the distribution of streamflow data?

- 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



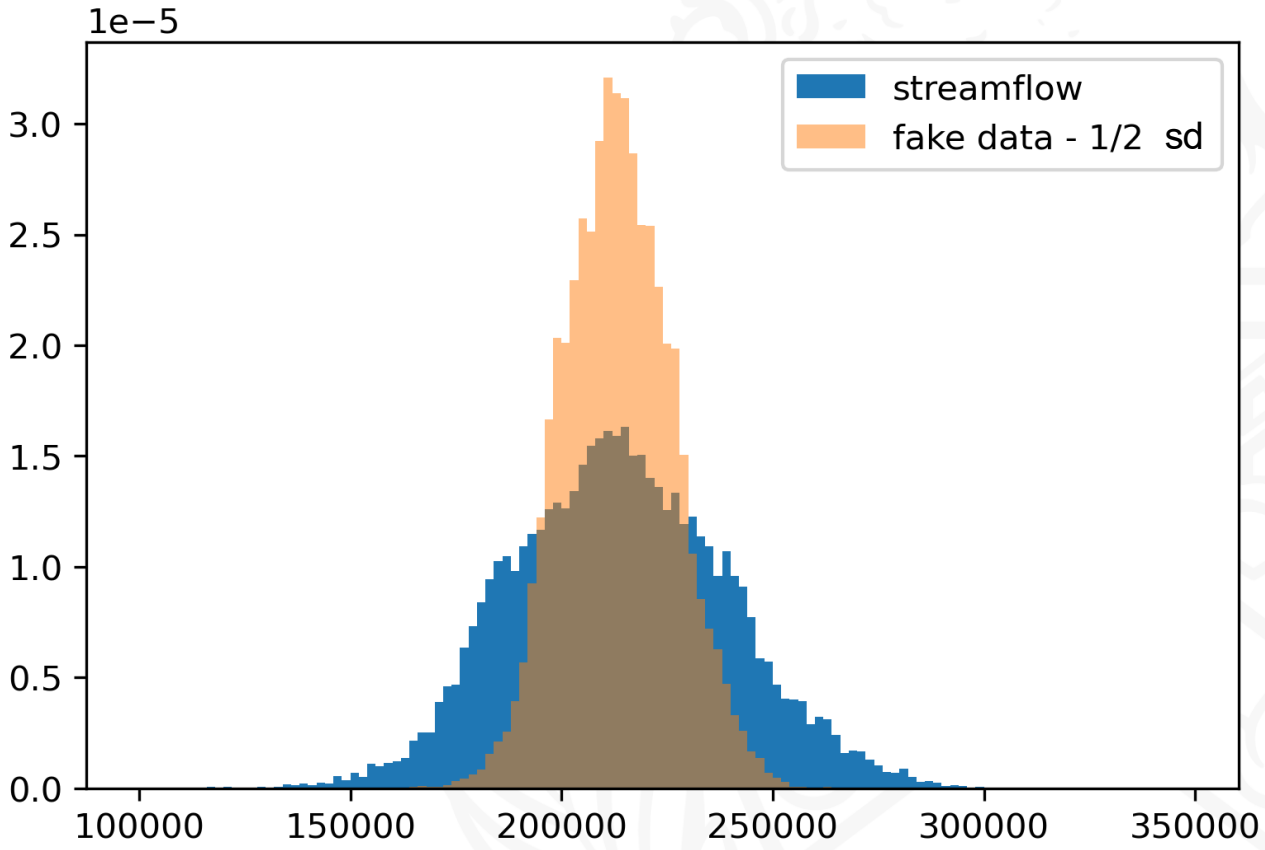
The histogram might look a bit different after we change the bin size

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?



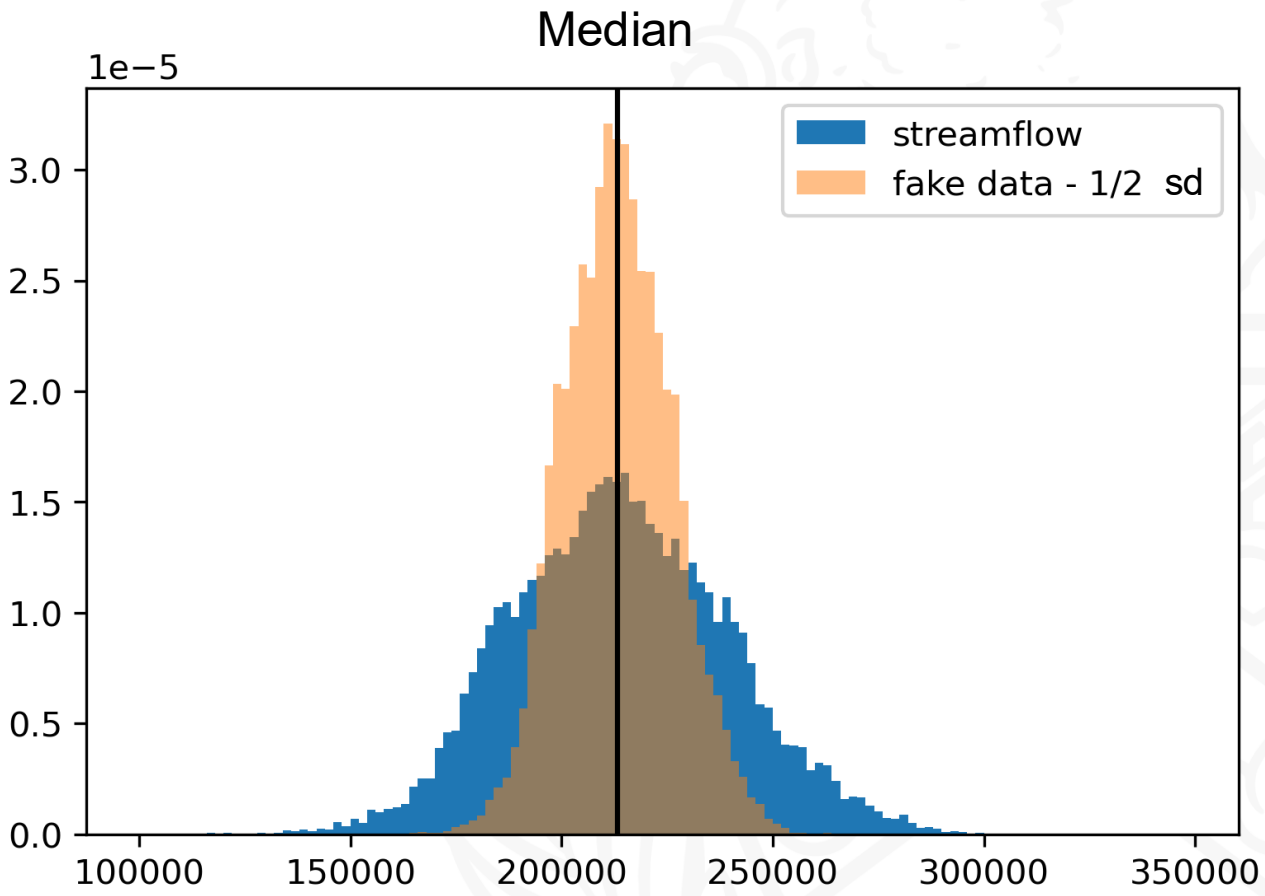
How do we evaluate the distribution of streamflow data?

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

- 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?

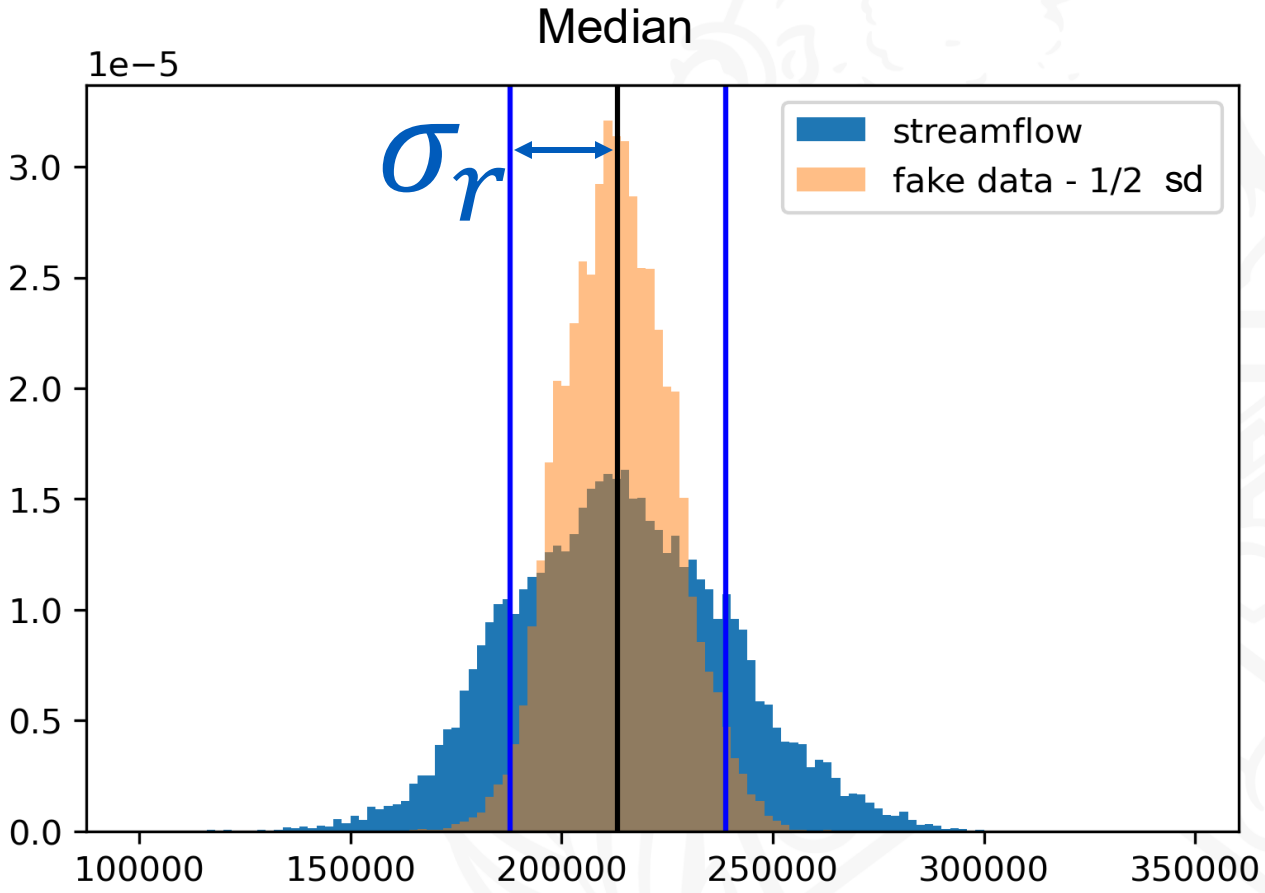


How do we evaluate the distribution of streamflow data?

- 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?

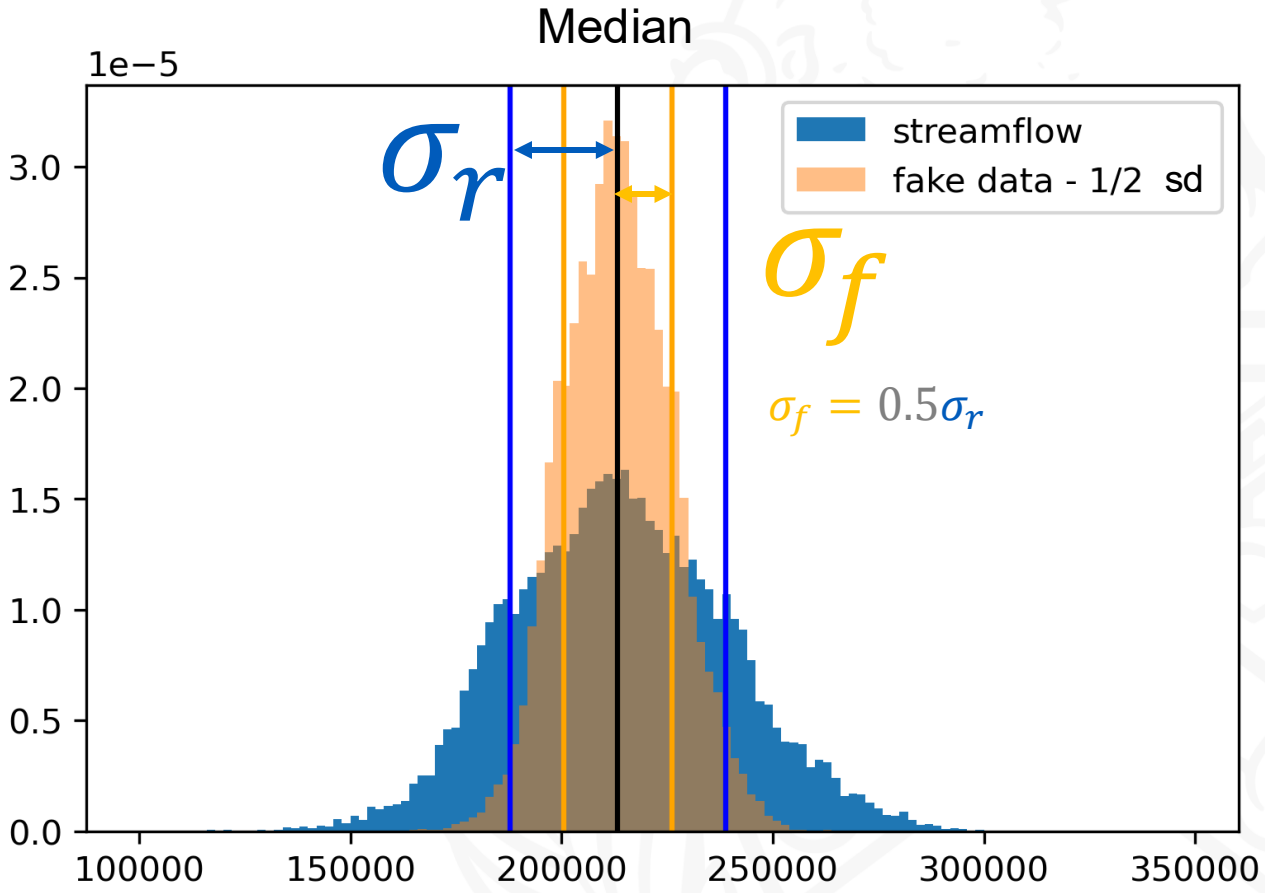


How do we evaluate the distribution of streamflow data?

- Probability Density Function

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

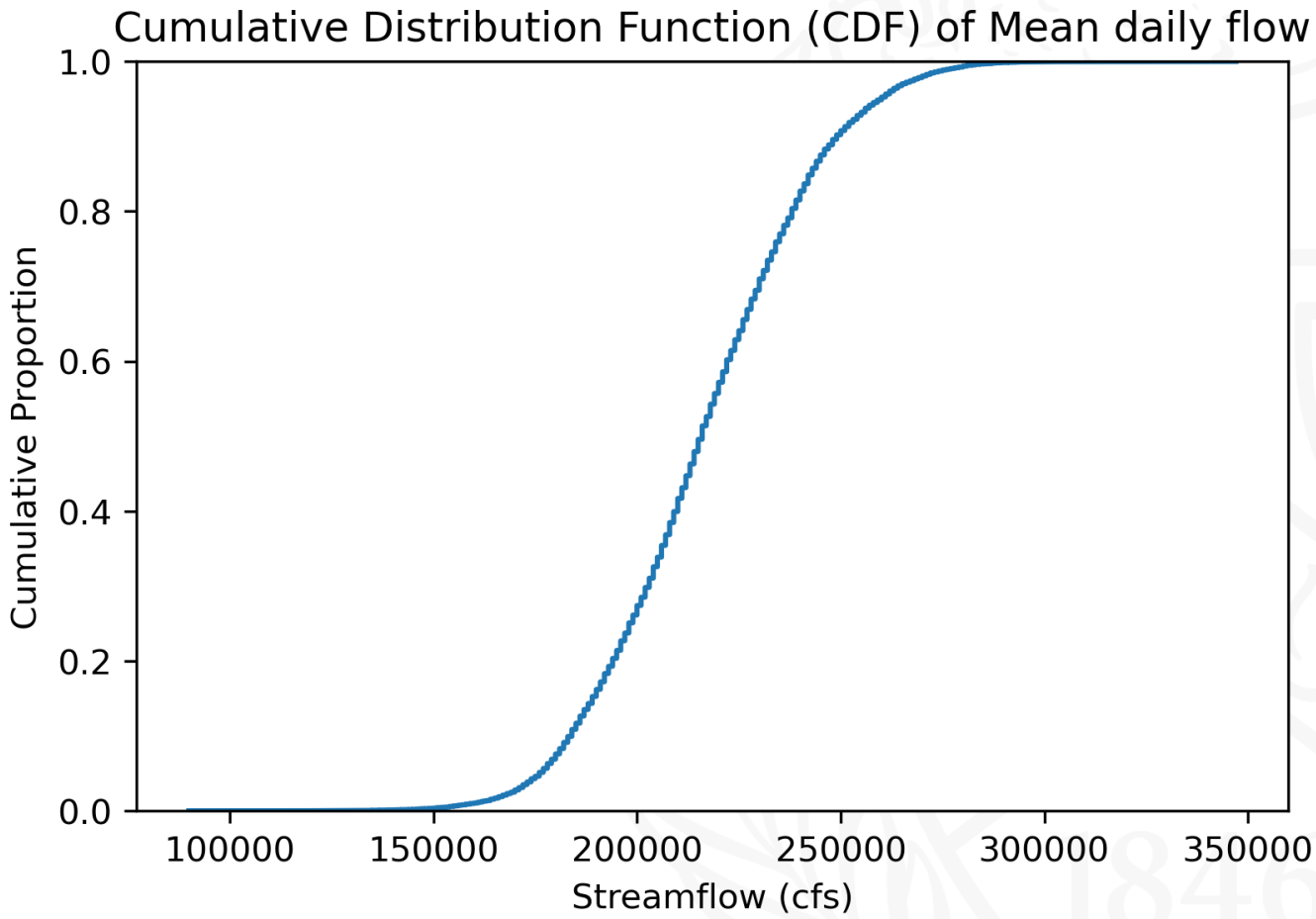
What does the PDF for our fake streamflow data look like?



How do we evaluate the distribution of streamflow data?

- Cumulative Distribution Functions

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

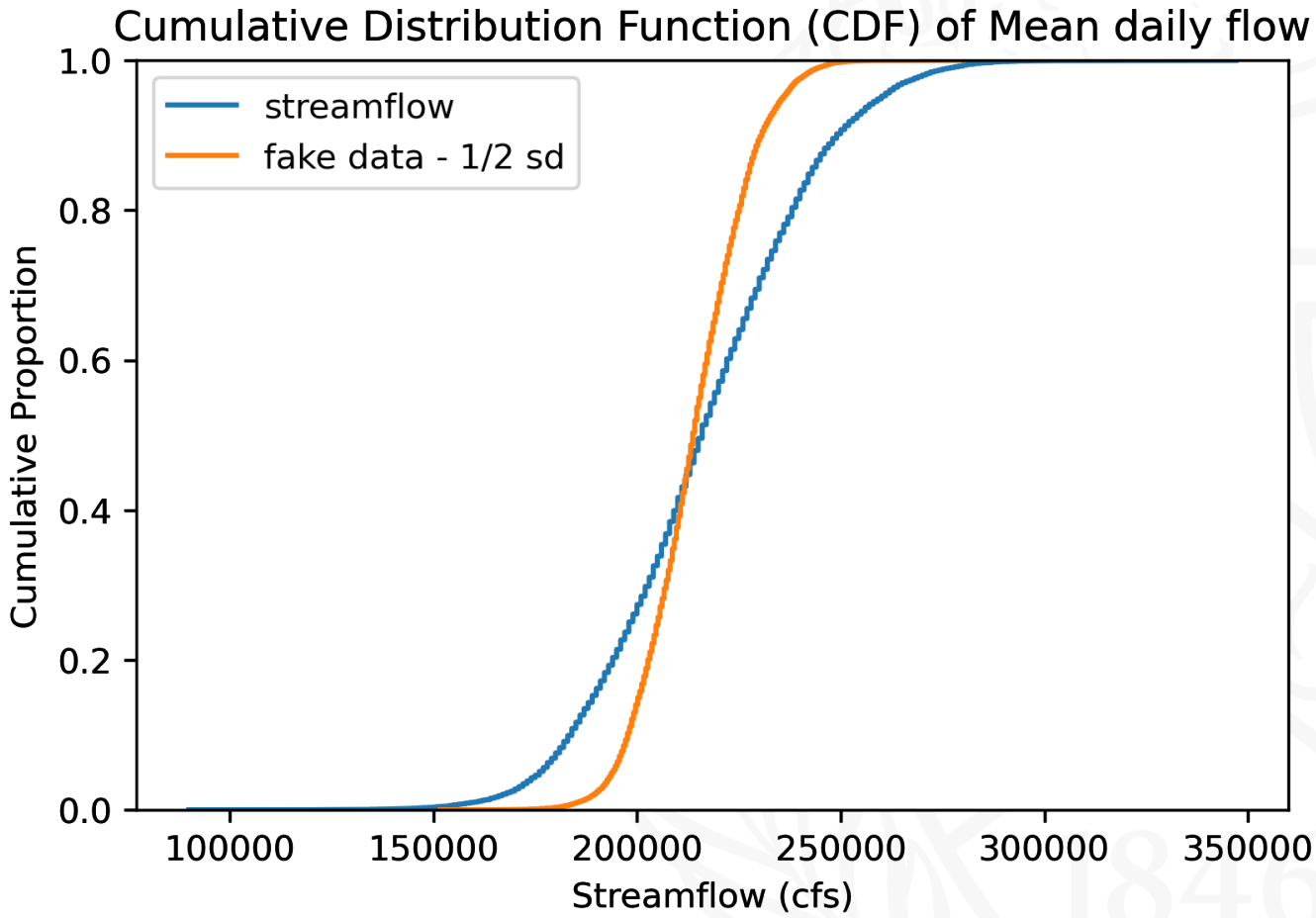


How do we evaluate the distribution of streamflow data?

- 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?

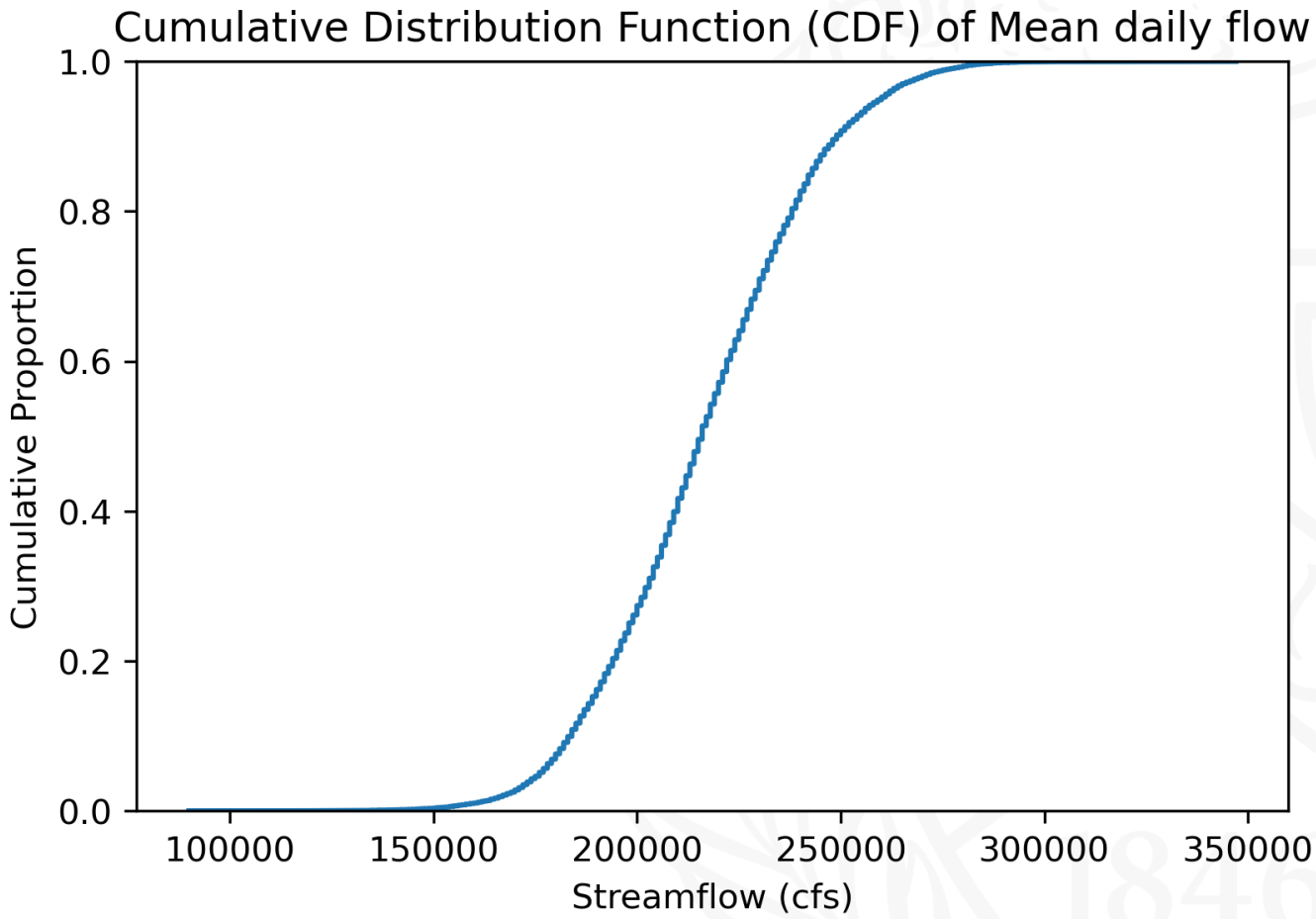


How do we evaluate the distribution of streamflow data?

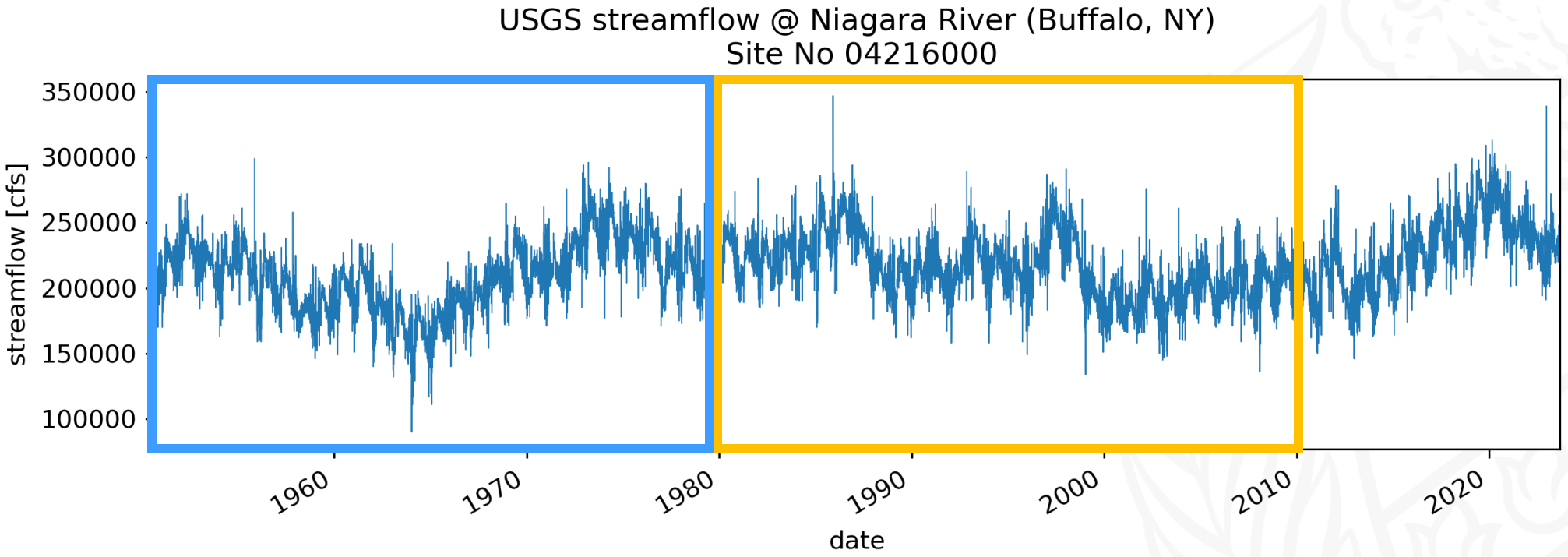
- Cumulative Distribution Functions

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

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



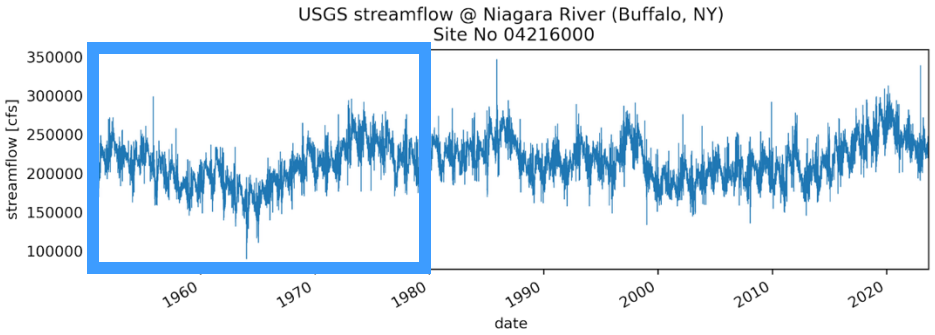
Use CDF to visualize the shift in hydrologic regimes



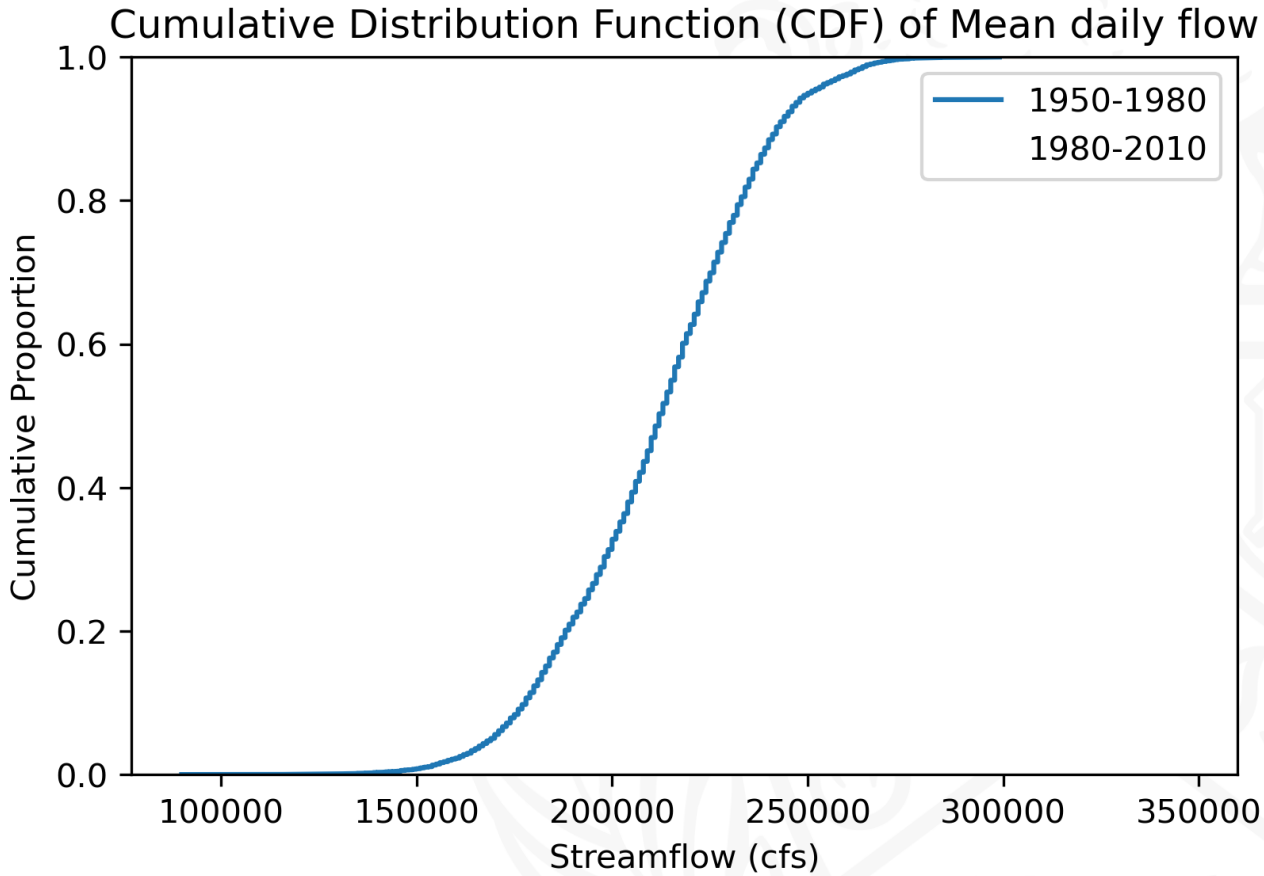
1950-1980

1980-2010

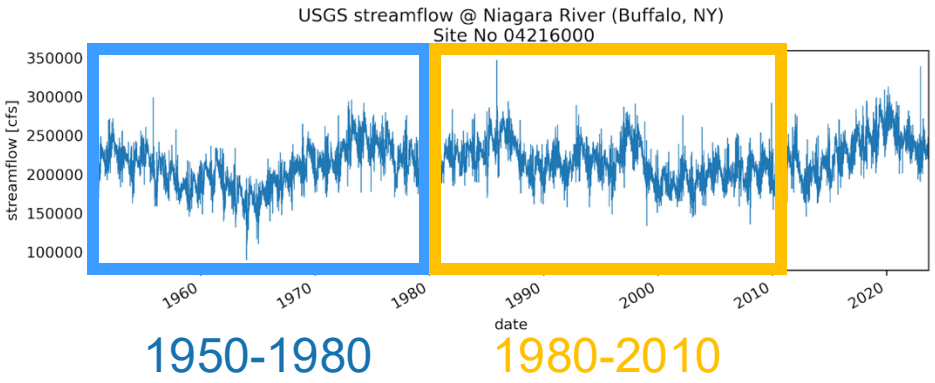
Use CDF to visualize the shift in hydrologic regimes



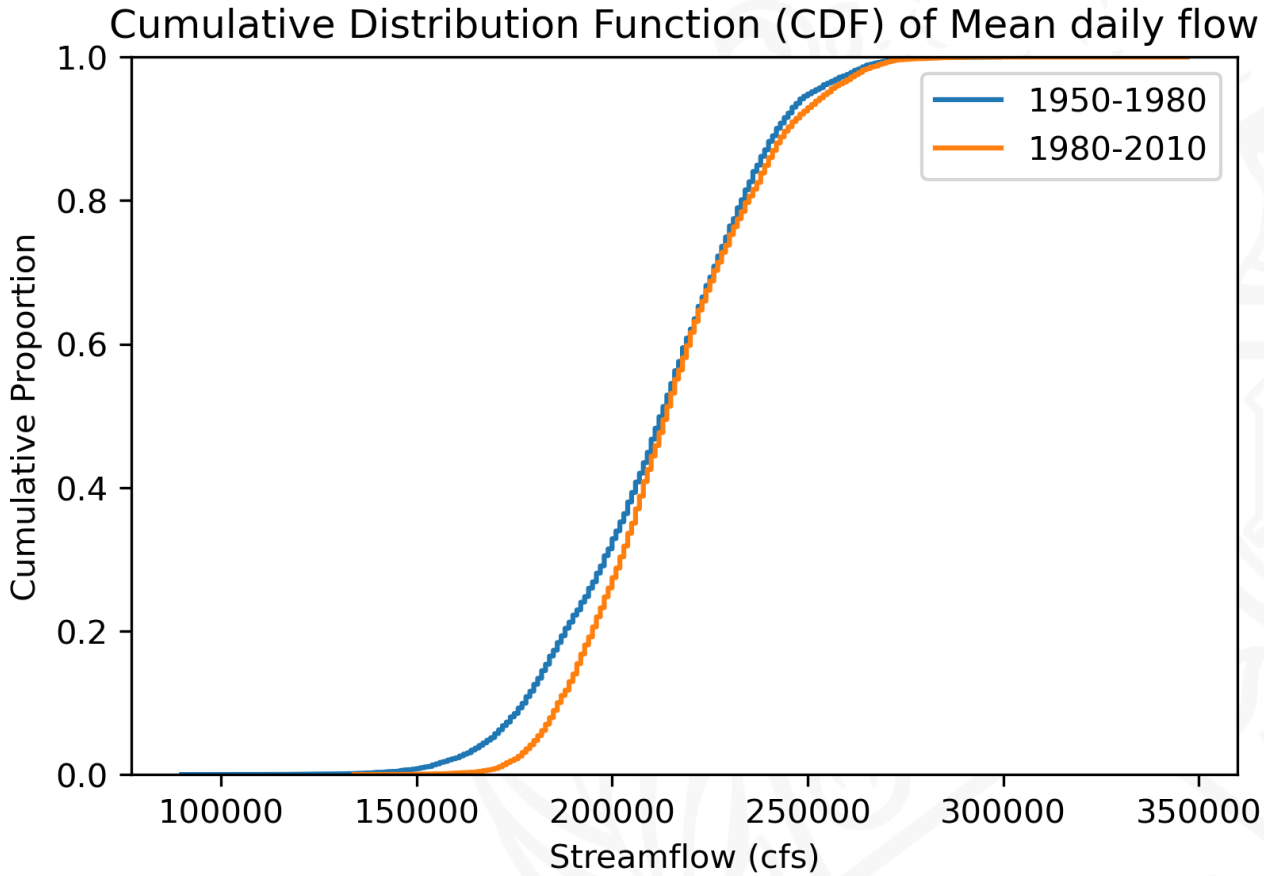
1950-1980 1980-2010



Use CDF to visualize the shift in hydrologic regimes



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



Thanks!