U.S. ARMY CORPS OF ENGINEERS

STREAMFLOW RECORD EXTENSION

05 May 2022

Ryan Cahill, Portland District

(with much source material from Beth Faber and Chuck Parrett)

US Army Corps of Engineers

U.S.ARMY

OUTLINE

- Problem statement
- Regression basics
- Annual Peak extension
- Continuous extension (e.g. daily)
- Common pitfalls

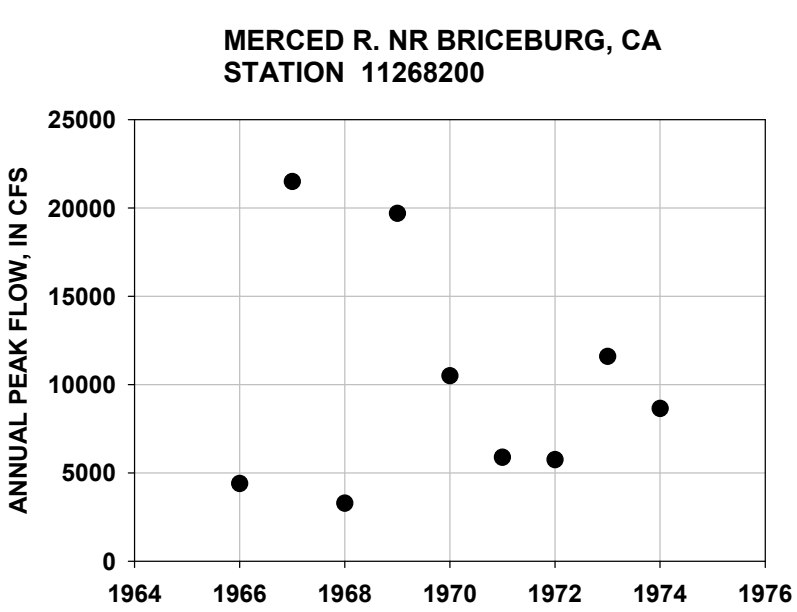
US Army Corps of Engineers

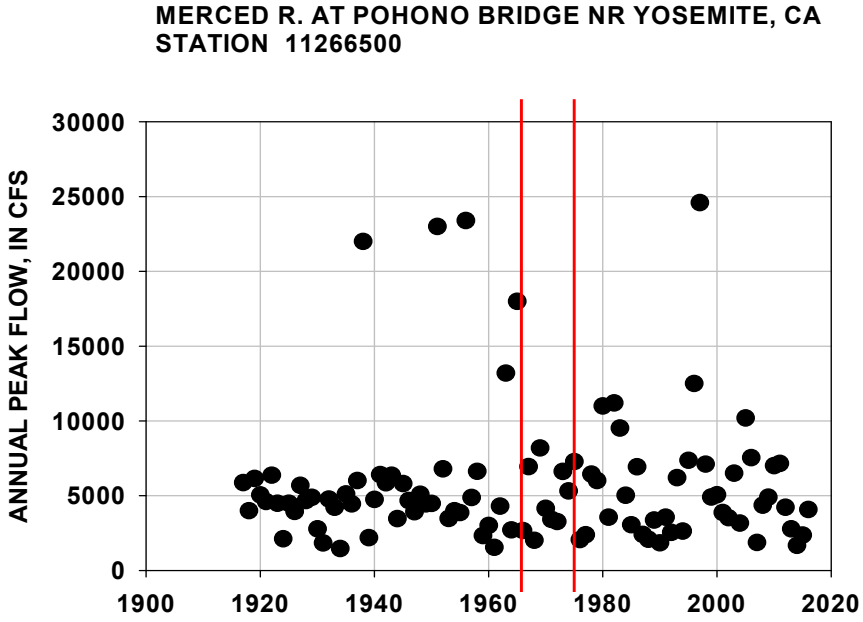
U.S.ARMY

PROBLEM STATEMENT

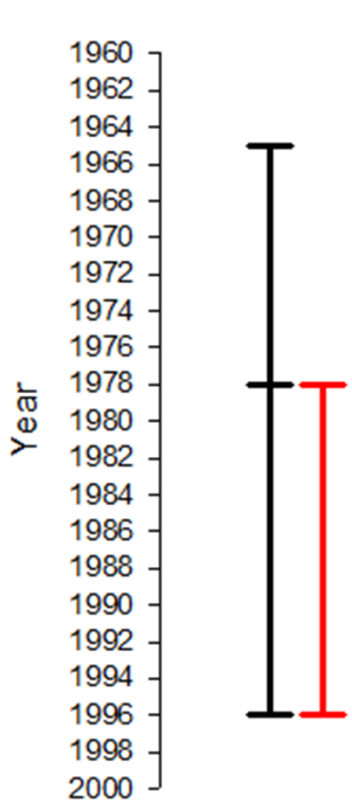
Problem: Not much data at a location of interest

Solution: Use a nearby long-term site to extend the record





ANOTHER VIEW



Use the relationship to fill in missing record at the short-term site (N_2=13)

Develop a relationship from the **concurrent** record (N_1=18)

US Army Corps of Engineers

OUTLINE

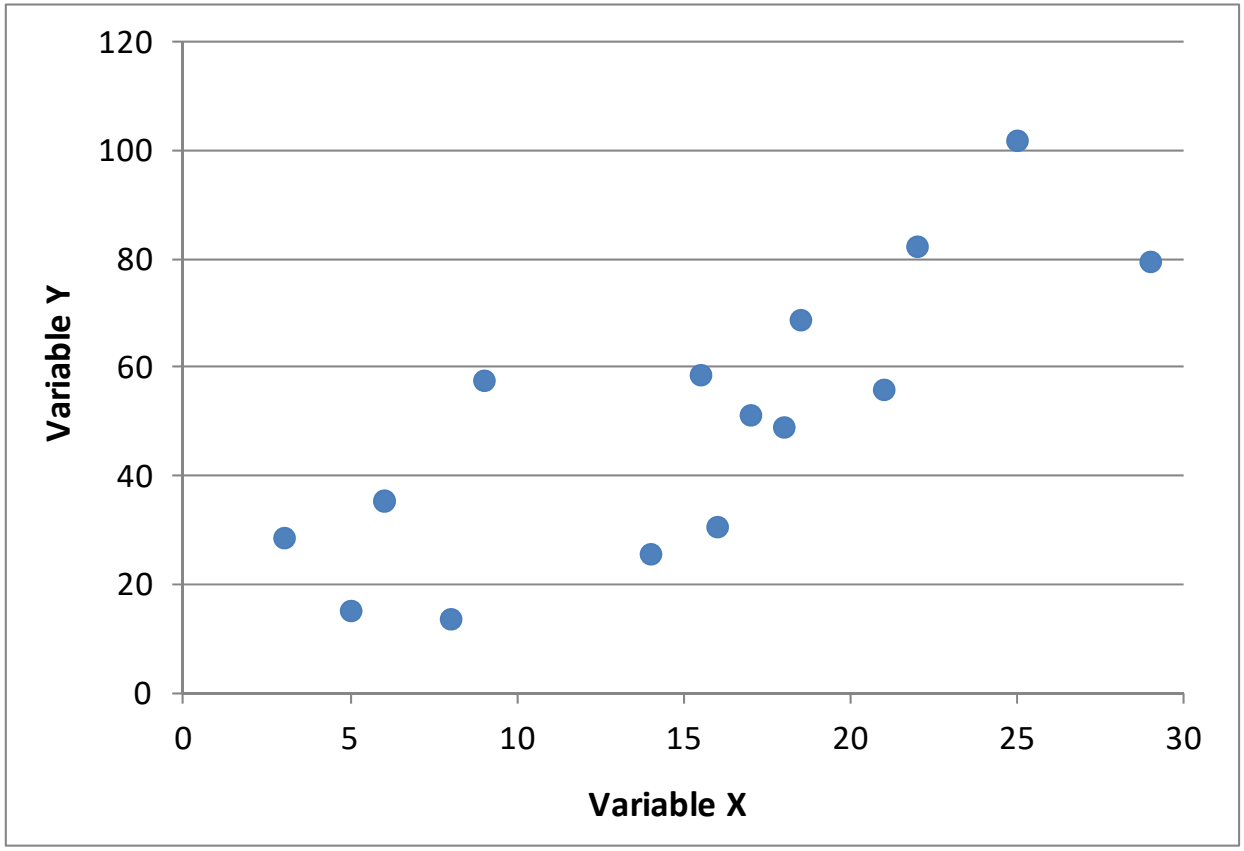
- Problem statement
- Regression basics
- Annual Peak extension
- Continuous extension (e.g. daily)
- Common pitfalls

US Army Corps of Engineers

U.S.ARMY

BASIC REGRESSION CONCEPT

- Can knowing the value of one variable help predict the value of another variable?

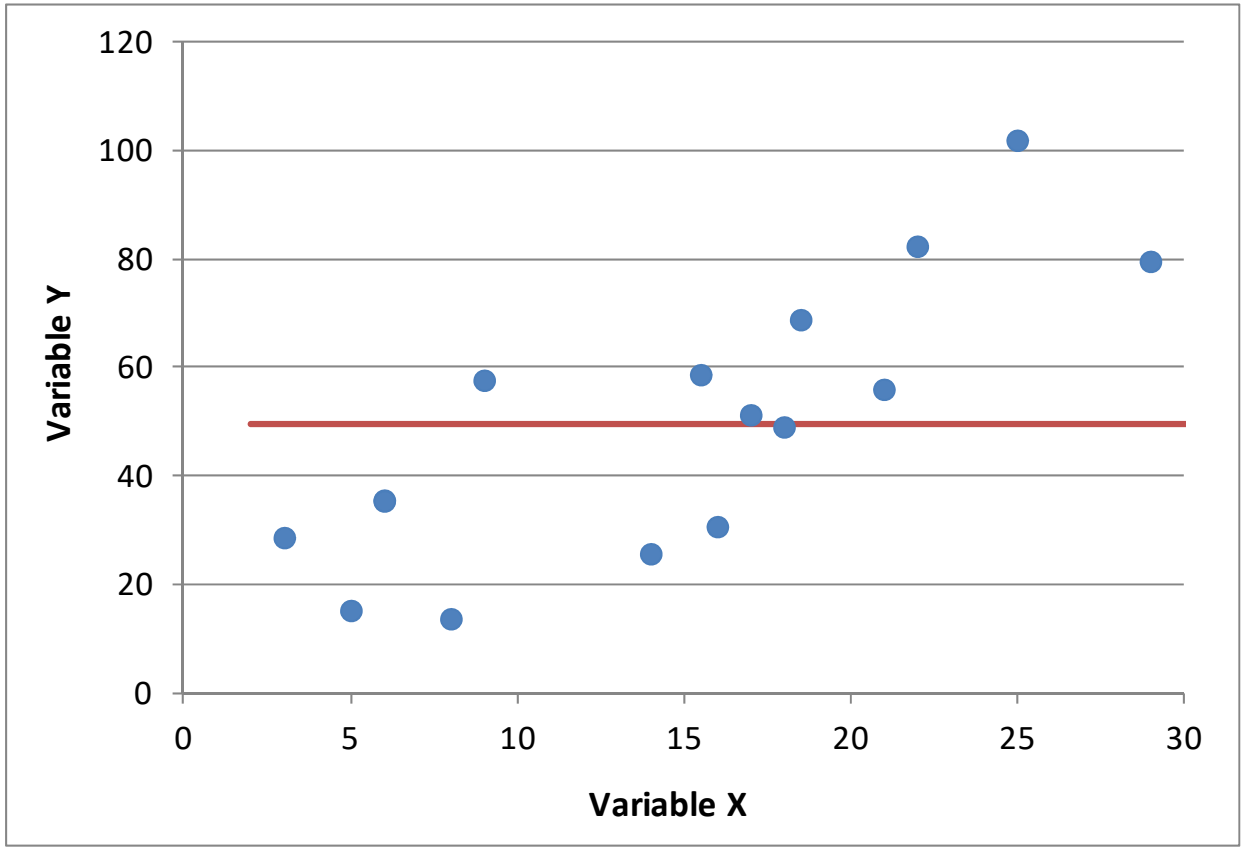


US Army Corps of Engineers

U.S.ARMY

BASIC REGRESSION CONCEPT

- If there is no relationship, the best prediction of variable Y is simply the mean of variable Y…

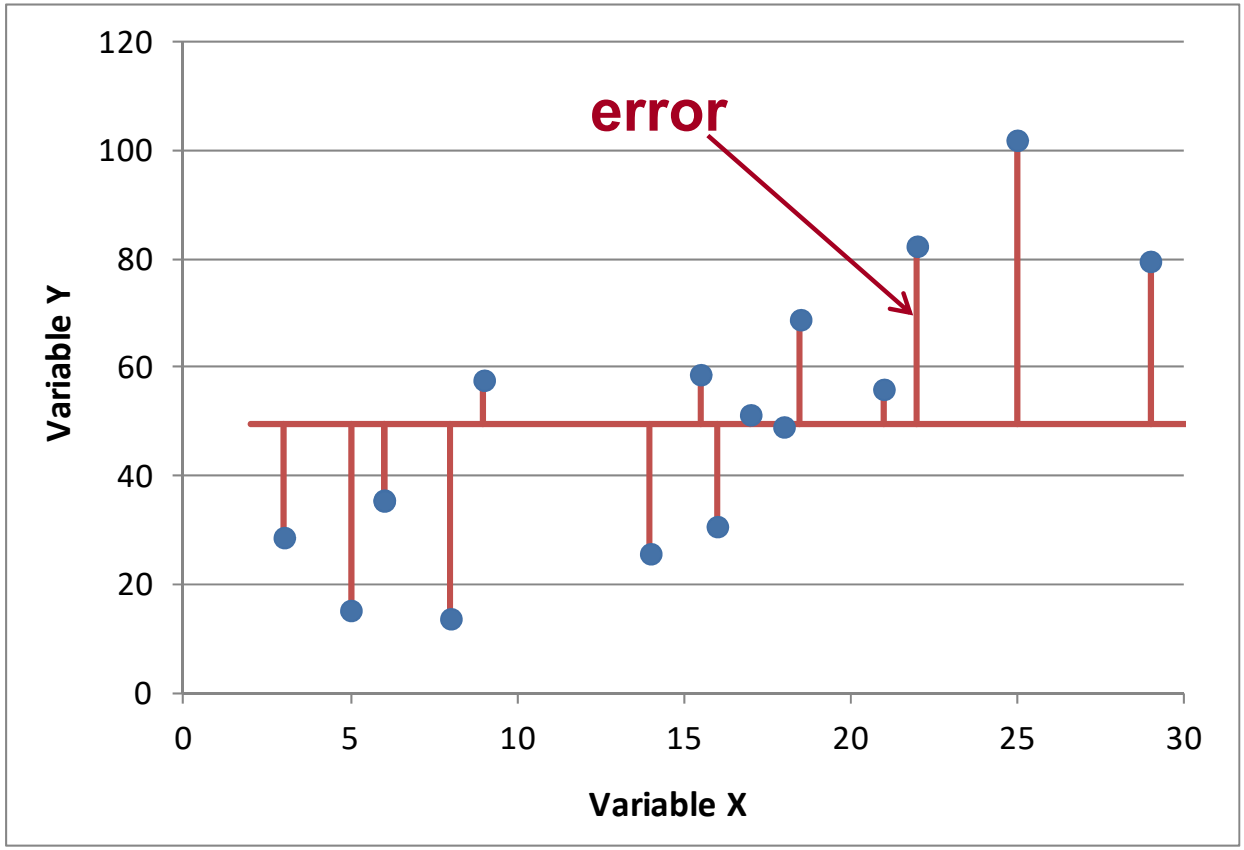


US Army Corps of Engineers

U.S.ARMY

BASIC REGRESSION CONCEPT

- If there is no relationship, the best prediction of variable Y is simply the mean of variable Y…

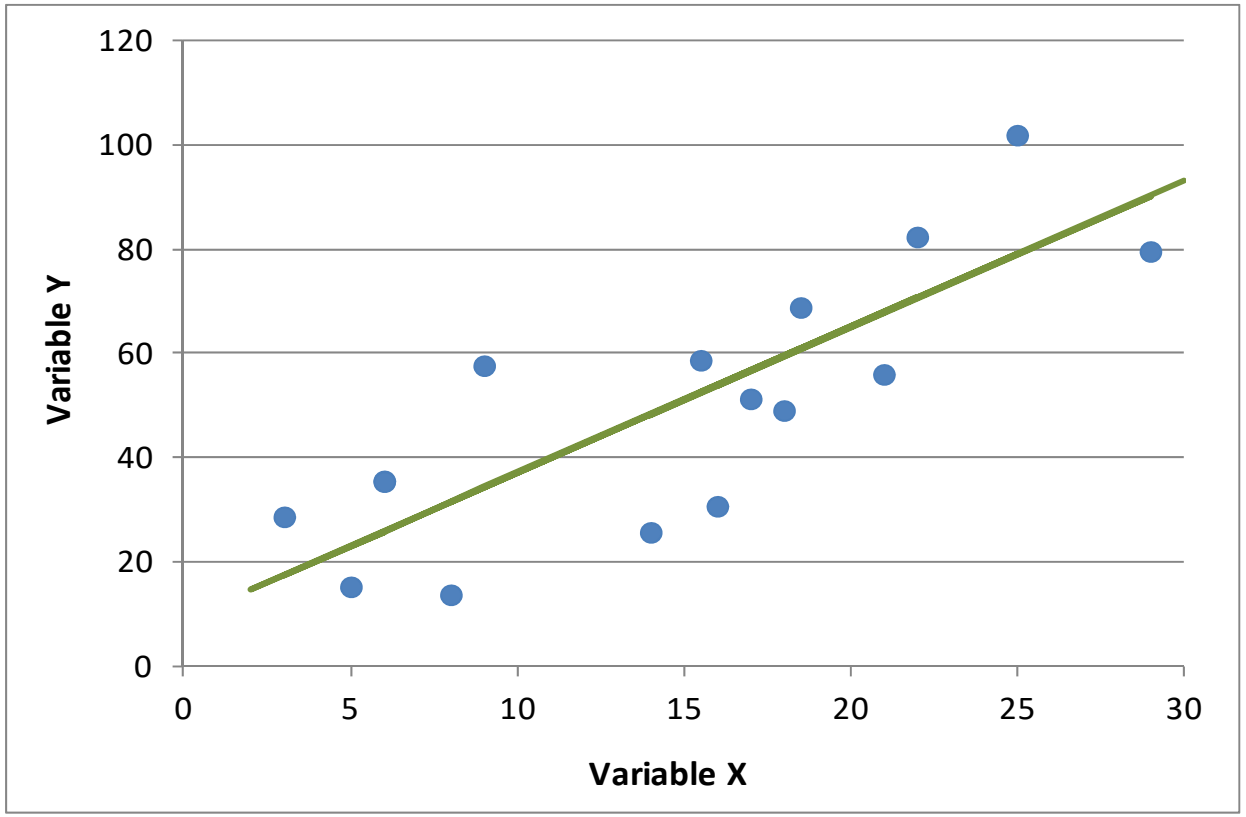


US Army Corps of Engineers

U.S.ARMY

BASIC REGRESSION CONCEPT

- If variable X has some ability to help predict variable Y, we seek a relationship between the two

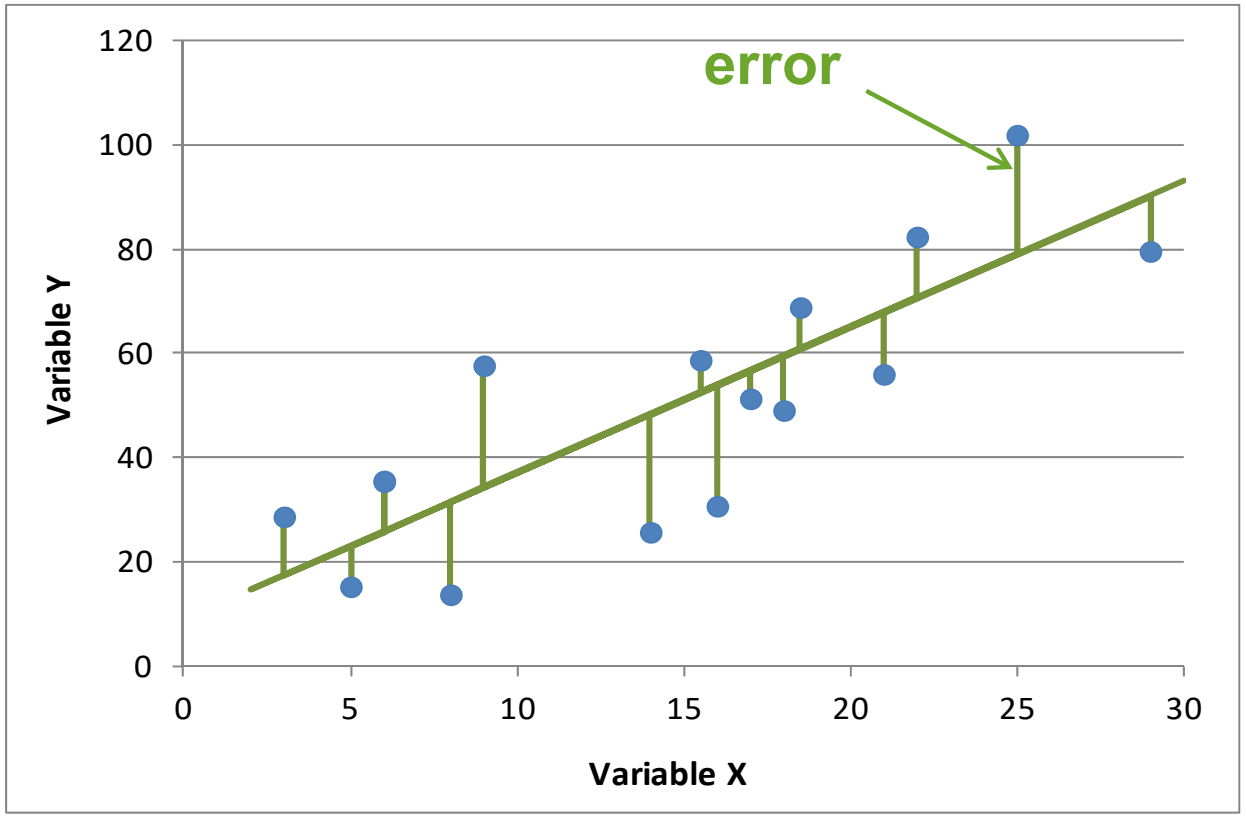


US Army Corps of Engineers

U.S.ARMY

BASIC REGRESSION CONCEPT

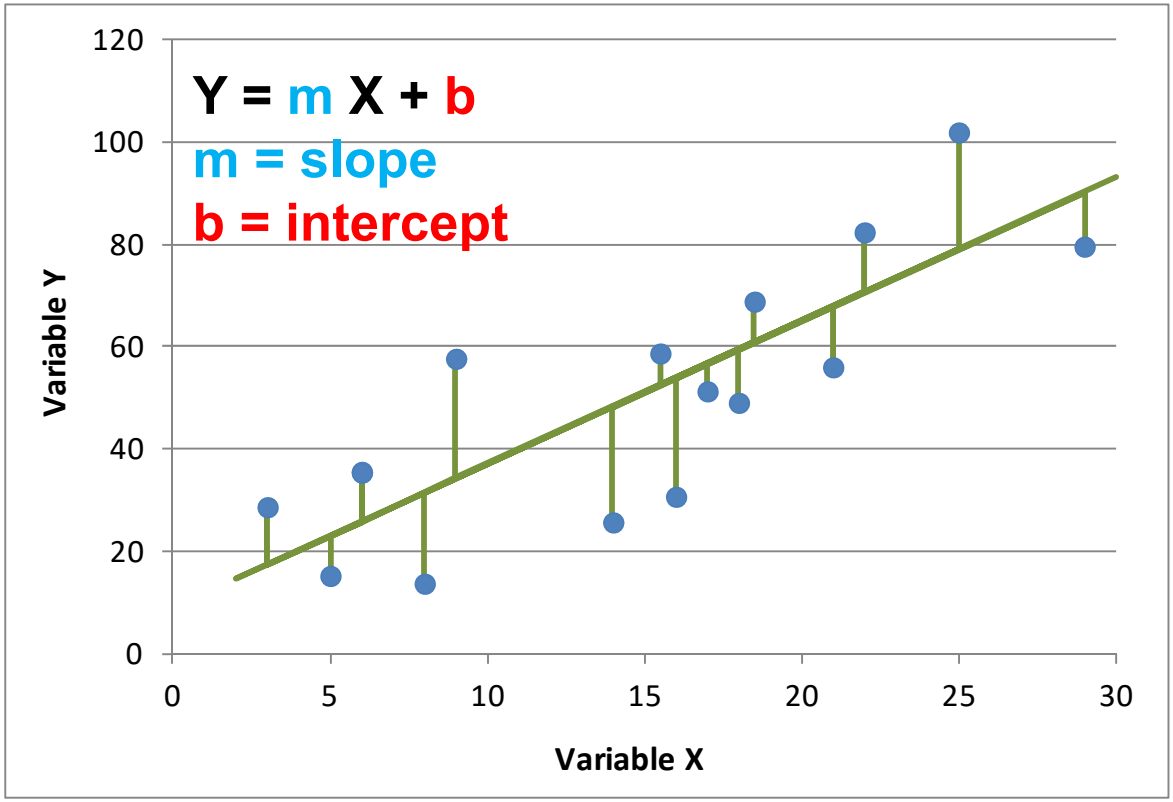
- If variable X has some ability to help predict variable Y, we seek a relationship between the two



US Army Corps of Engineers

ORDINARY LEAST-SQUARES REGRESSION

- In OLS, the “best” relationship between variable X and variable Y is one the minimizes the sum of squared errors



US Army Corps of Engineers

U.S.ARMY

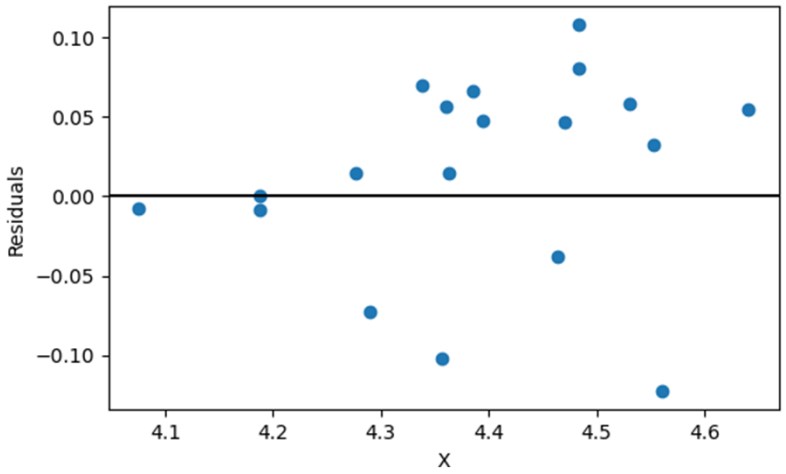
QUALITY AND ASSUMPTIONS

Goodness-of-fit metrics:

- R^2 = squared correlation
  - % of the variability in Y explained by the variability in X
- standard error = square root of the sum of squared errors

Assumptions:

- errors are homoscedastic = evenly distributed across X
- errors are normally distributed

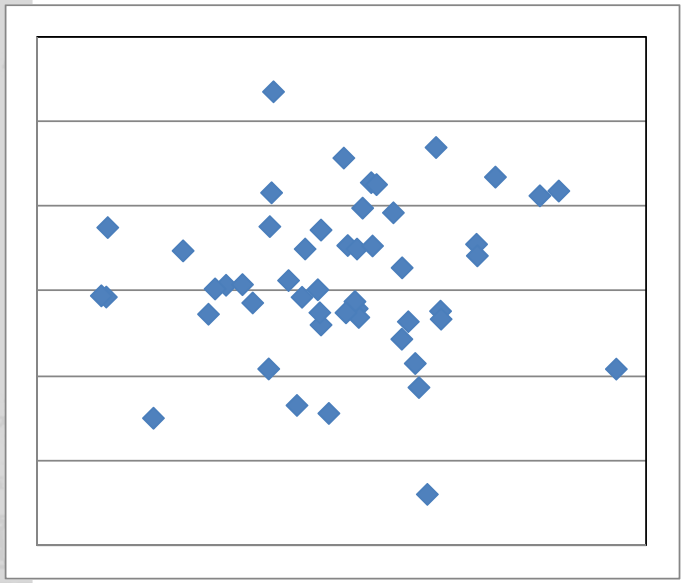


US Army Corps of Engineers

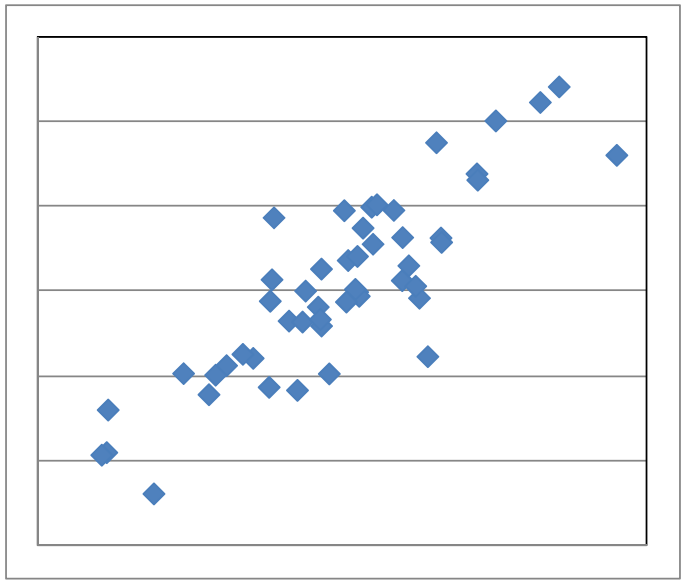
U.S.ARMY

CORRELATION

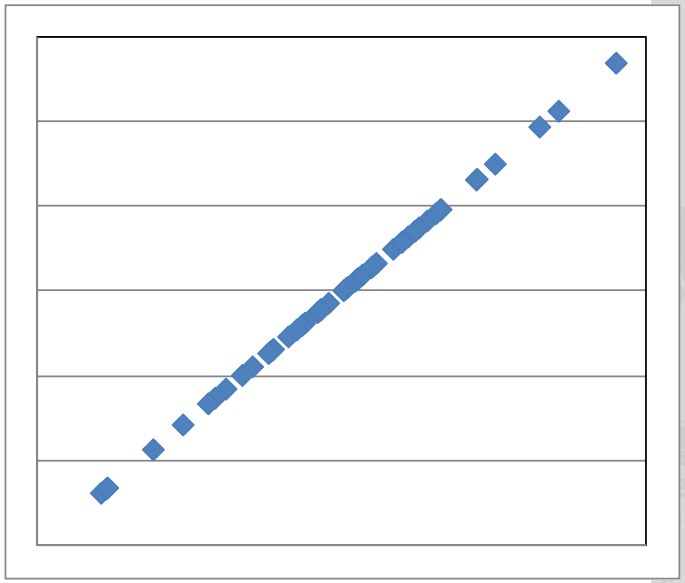
correlation = 0



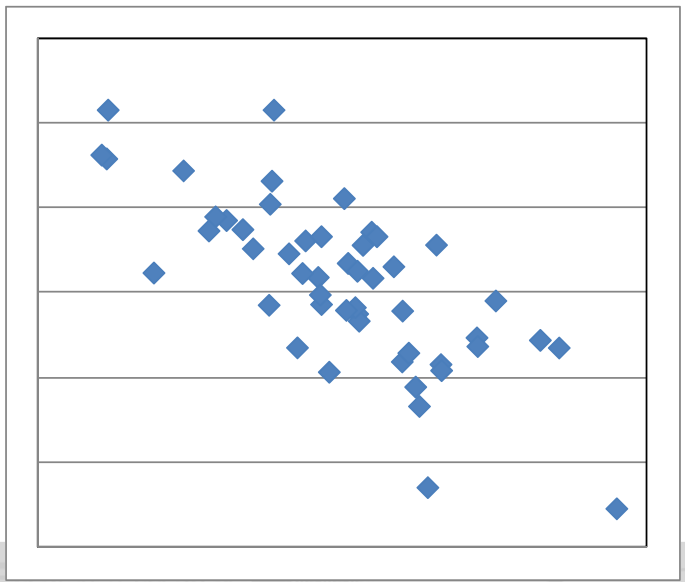
correlation = 0.7



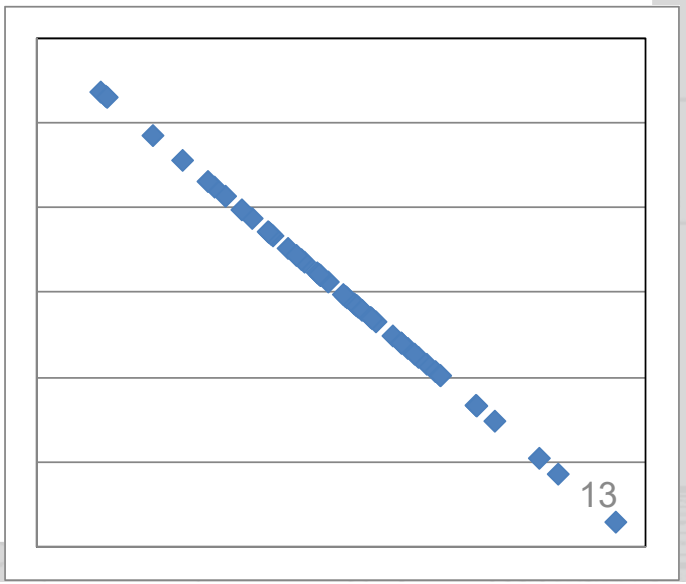
correlation = 1.0



correlation = -0.6

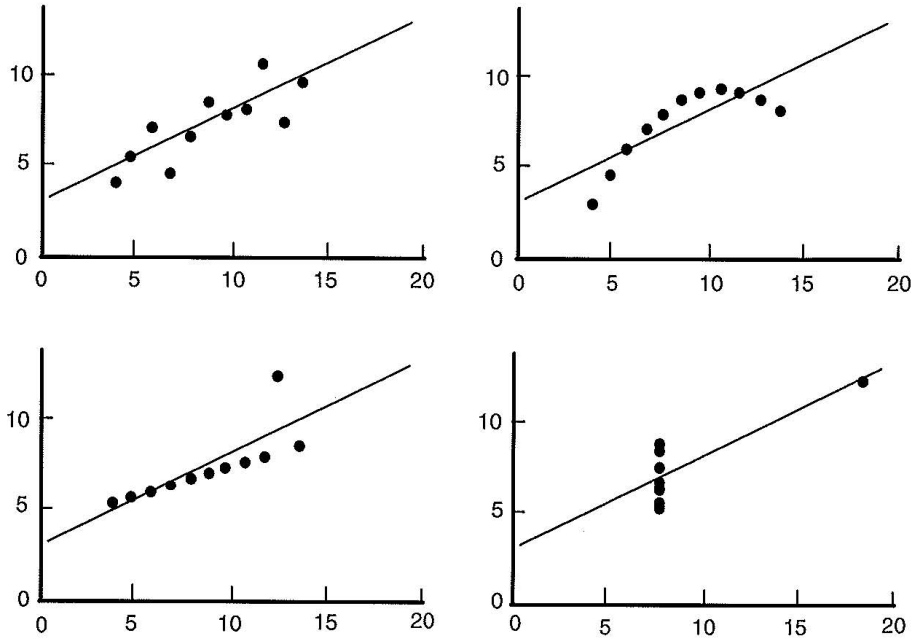


correlation = -1.0



REGRESSION DIAGNOSTICS

- R^2 and Standard Error can be misleading
- Plot the data
- Same R^2 and SE on all graphs (Anscombe 1973)



US Army Corps of Engineers

WHY LEAST-SQUARES?

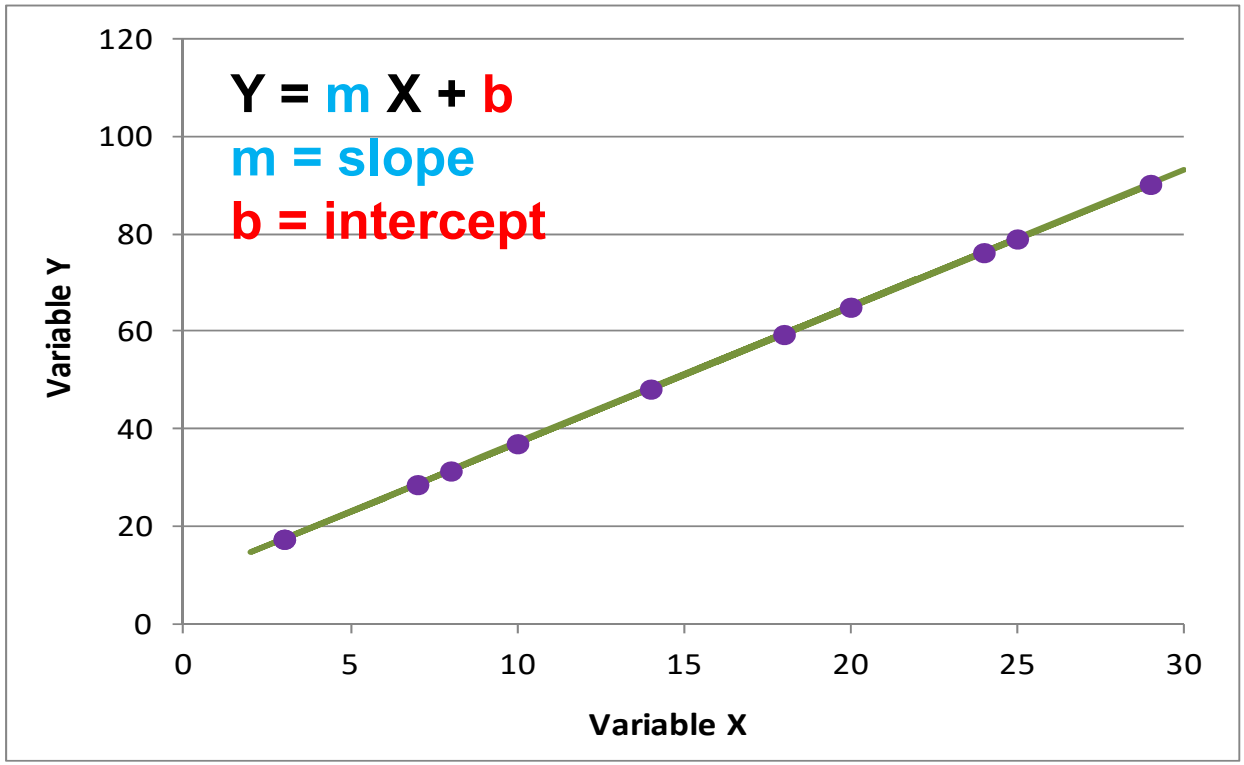
- Other measures of “closeness” could be used (e.g. absolute value of errors)
- It is mathematically convenient—there is no analytical solution to the absolute value method

US Army Corps of Engineers

U.S.ARMY

BASIC REGRESSION CONCEPT

- Estimates of variable Y, when only Variable X is available, will follow the regression line



US Army Corps of Engineers

U.S.ARMY

OLS REGRESSION

- OLS is preferred method of predicting a particular value of Y given a value of X

But…

- If |r| < 1 (nonperfect fit), then variance of predicted values of Y will tend to be less than variance of true values. That is,

$$E\left[S^2_{\hat{Y}}\right] < \sigma^2_Y$$

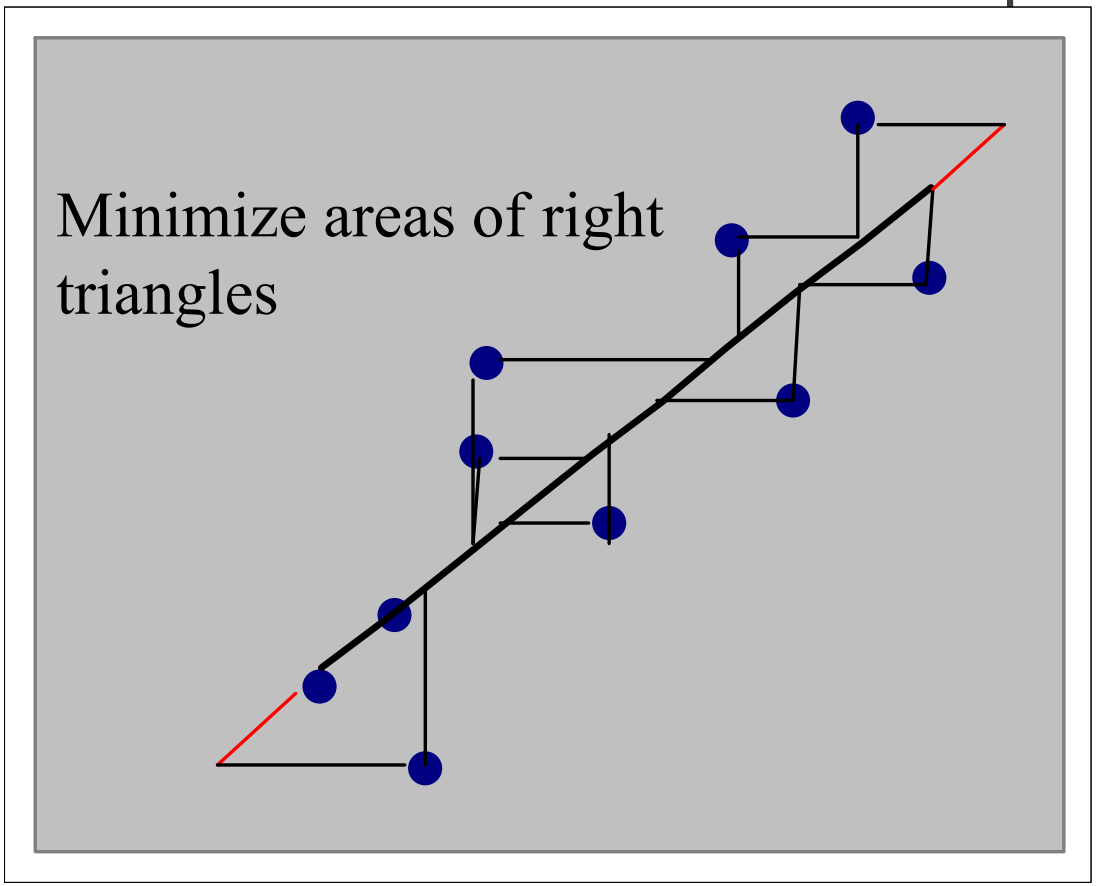
- Reduced variance for a series of estimates is a problem for record extension

US Army Corps of Engineers

U.S.ARMY

LINE OF ORGANIC CORRELATION (LOC)

- The LOC is the line that minimizes the sum of squared geometric distances in both the X and Y direction
- This method does not reduce variance of predictions
- Used for MOVE.1 record extension technique



US Army Corps of Engineers

U.S.ARMY

REGRESSION EQUATIONS

- OLS: $\hat{Y}_i = \overline{Y} + r\frac{S_Y}{S_X}(X_i - \overline{X})$

- LOC: $\hat{Y}_i = \overline{Y} + \frac{S_Y}{S_X}(X_i - \overline{X})$

No “r” term here

US Army Corps of Engineers

U.S.ARMY

STATISTICS (BASED ON LOGS) FOR CONCURRENT RECORD (N_1)

longer record station	shorter record station
$\overline{X} \equiv \frac{1}{N_1}\sum_{i=1}^{N_1} X_i$	$\overline{Y} \equiv \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i$
$S_X^2 \equiv \frac{1}{N_1-1}\sum_{i=1}^{N_1}(X_i-\overline{X})^2$	$S_Y^2 \equiv \frac{1}{N_1-1}\sum_{i=1}^{N_1}(Y_i-\overline{Y})^2$

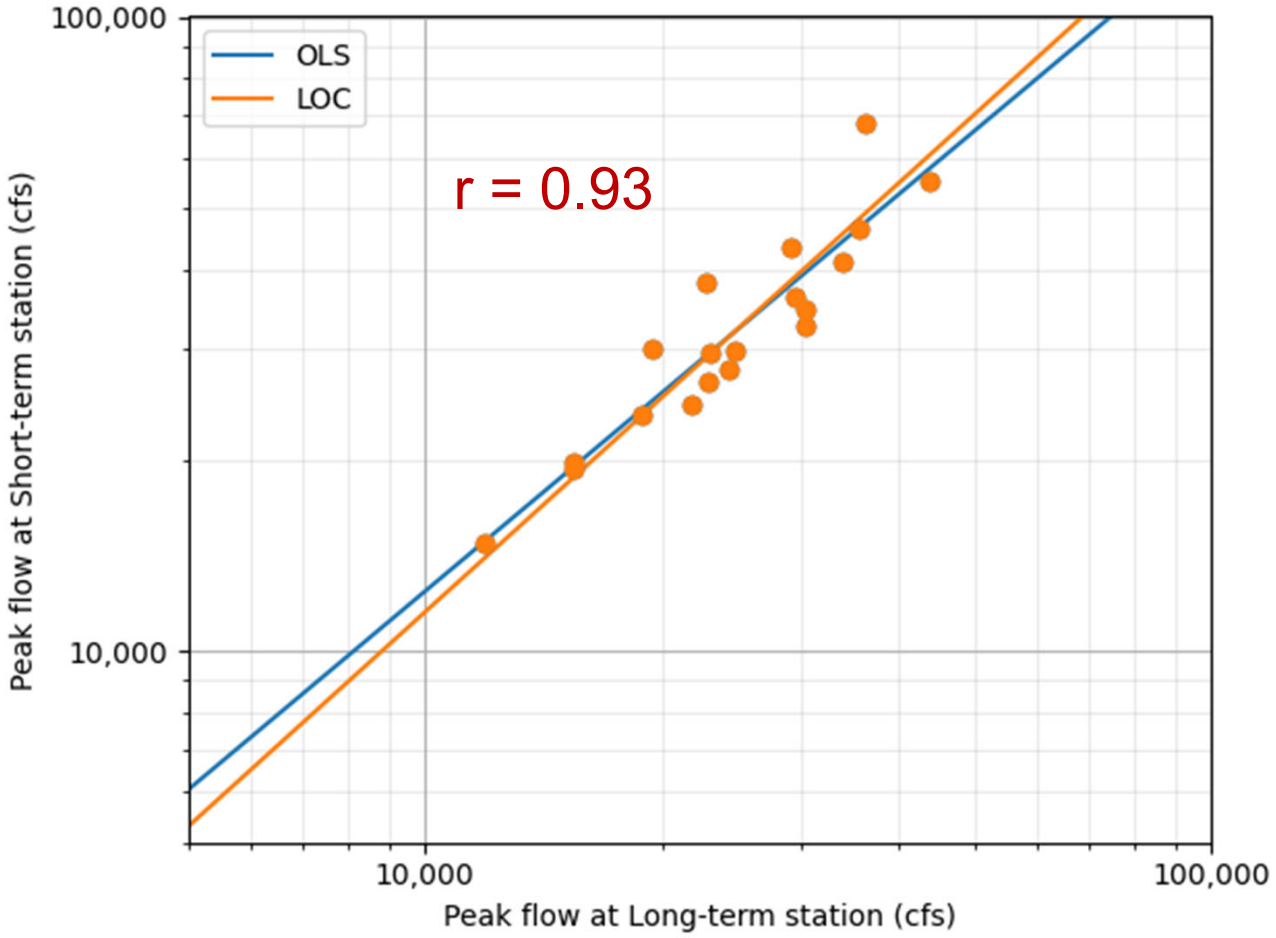
$$r \equiv \frac{\frac{1}{N_1-1}\sum_{i=1}^{N_1}(X_i-\overline{X})(Y_i-\overline{Y})}{S_X S_Y}$$

US Army Corps of Engineers

U.S.ARMY

OLS VS LOC

- For streamflow record extension applications, LOC produces higher estimates for large flows and lower estimates for small flows



US Army Corps of Engineers

U.S.ARMY

OUTLINE

- Problem statement
- Regression basics
- Annual Peak extension
- Continuous extension (e.g. daily)
- Common pitfalls

US Army Corps of Engineers

U.S.ARMY

ANNUAL PEAK FLOW EXTENSION

- Goal is typically a Bulletin 17C analysis
- All record extension techniques use logarithm of flow

Step 1: Develop a linear relationship between X and Y (the long and short record stations) using the concurrent record

Step 2: Use the relationship to estimate values for the short record station for times we only have values for the long record station (non-concurrent)

Step 3: Perform a frequency analysis on the extended dataset using the Expected Moments Algorithm

US Army Corps of Engineers

U.S.ARMY

HOW SHOULD WE SELECT A LONG-TERM SITE FOR RECORD EXTENSION?

- Various studies have recommended that the correlation coefficient (r) between short-term and long-term sites for the concurrent record be 0.8 or greater.
- Long-term sites with flow values in the non-concurrent record period that are substantially outside the range of values in the concurrent period may provide more information than other potential long-term sites.
- *More than one long-term site* can be used and results weighted (perhaps using r or record length) with results from another long-term site.
- Long-term sites should be near the short-term site with similar basin characteristics

US Army Corps of Engineers

PAST METHODS

Methods not typically used in current practice for annual peak extension:

- OLS Regression Plus Noise (RPN)
- Maintenance of Variance Extension (MOVE)
  - MOVE.1
  - MOVE.2
  - MOVE.3 (as originally formulated in Vogel and Stedinger 1985)
  - MOVE.4
  - GMOVE
- Two-station comparison used in Bulletin 17B (MOVE.2)

US Army Corps of Engineers

U.S.ARMY

MATALAS-JACOBS ESTIMATORS

- Updated mean and standard deviation of the shorter station, based on the full record of the longer station.
- Used in both MOVE.2 (two-station comparison in Bulletin 17B) and MOVE.3

X = Longer station
Y = Shorter station

$$\widehat{\overline{Y}_{all}} = \overline{Y}_1 + \frac{N_2}{(N_1 + N_2)} r \frac{S_{Y_1}}{S_{X_2}} (\overline{X}_2 - \overline{X}_1)$$

Matalas-Jacobs Estimators

$$\widehat{S_{Y_{all}}} = \sqrt{\frac{1}{(N_1 + N_2 - 1)} (A + B + C)}$$

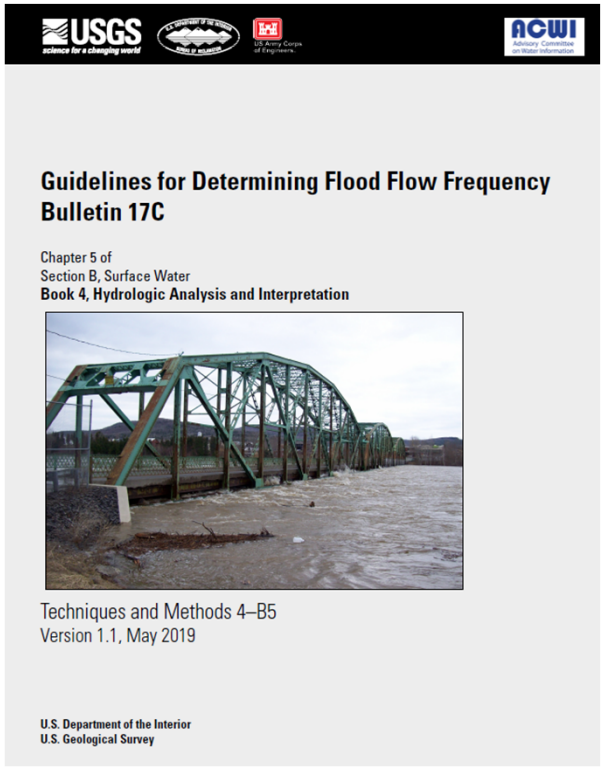
N_1 = concurrent record, N_2 = additional record

US Army Corps of Engineers

U.S.ARMY

CURRENT METHOD

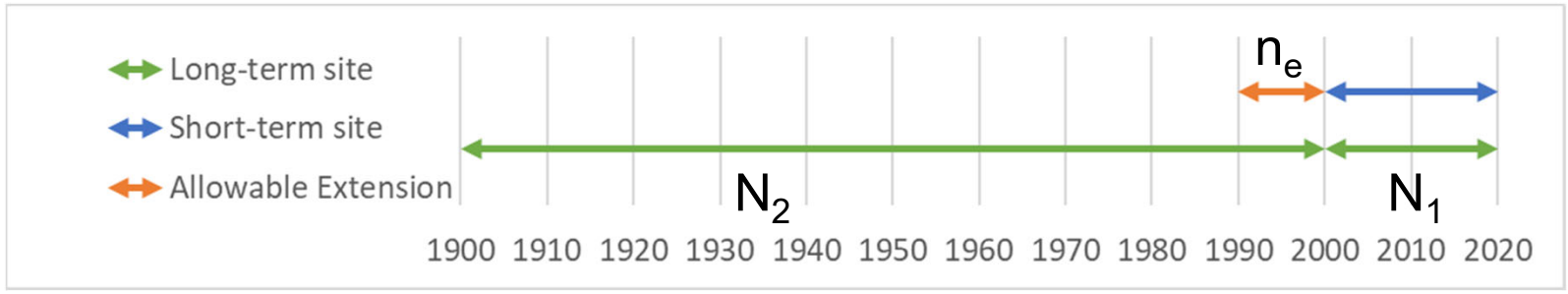
- Appendix 8 of Bulletin 17C contains guidance on record extension
  - Make sure to use Version 1.1 of Bulletin 17C, not the original published version
- Bulletin 17C recommends the MOVE.3 technique, but with a twist



US Army Corps of Engineers

MOVE.3 (BULLETIN 17C VERSION)

- Original MOVE.3 allows for extension to be performed for every non-concurrent value of the long-term site (N_2)
- If the long-term site has many more years than the short-term site, we would get a false sense of confidence in our estimates at the short-term site
- The "twist" adopted by Bulletin 17C:
  - Define **n_e**: the maximum number of years allowable for record extension.
  - Higher correlation = higher n_e
  - Modify MOVE.3 equations to use n_e instead of N_2



US Army Corps of Engineers

MOVE.3 (BULLETIN 17C VERSION)

n_e is calculated twice in Bulletin 17C:

1. Max allowable n_e for the mean
2. Max allowable n_e for the variance

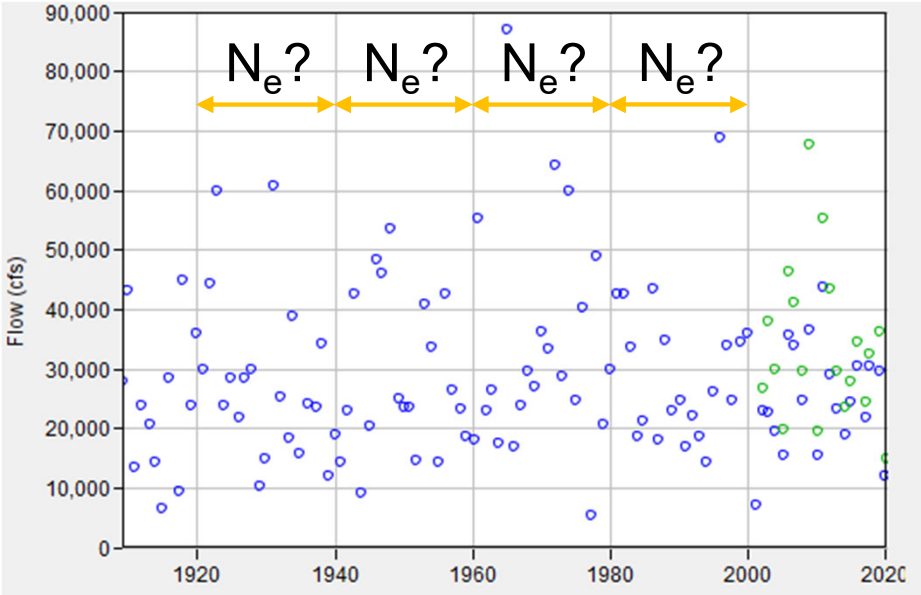
#2 is always smaller than #1, so it governs

US Army Corps of Engineers

U.S.ARMY

WHICH YEARS TO EXTEND?

- Only n_e years of record extension are allowed.
- Which years should we pick?
- Different year selections will not affect the mean or variance, but will affect the skew
- Bulletin 17C allows for judgment of the analyst on year selection to ensure the skew isn't misrepresented

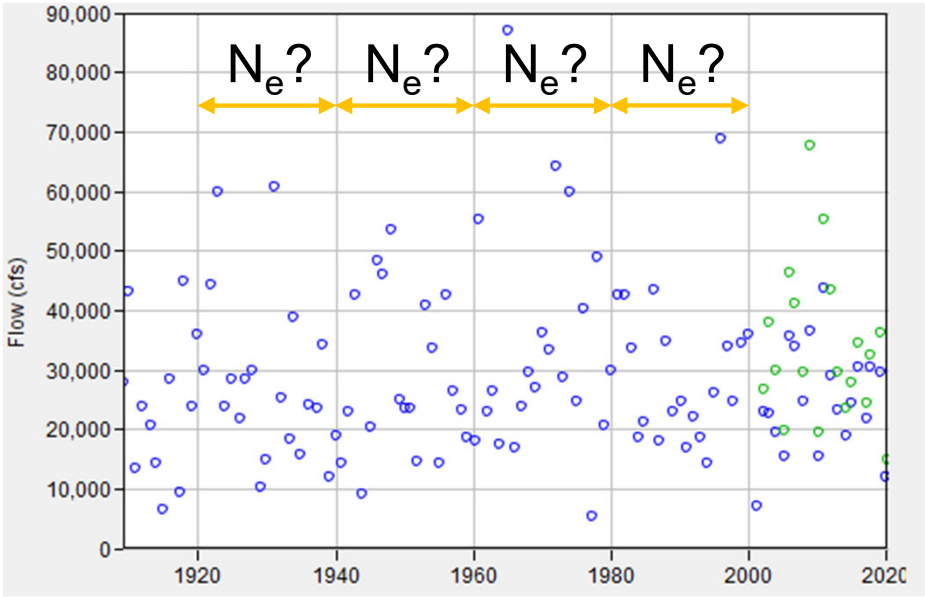


US Army Corps of Engineers

U.S.ARMY

WHICH YEARS TO EXTEND?

- **Default**: use the most recent years. Usually just fine
- But if a sequence of unusually big floods or small floods is in n_e, may need to adjust:
  1. Compute the skew using a record extension for the entire period of record (original MOVE.3 technique using N_2, not limited to n_e)
  2. Select a sequence of n_e years that results in a similar skew value

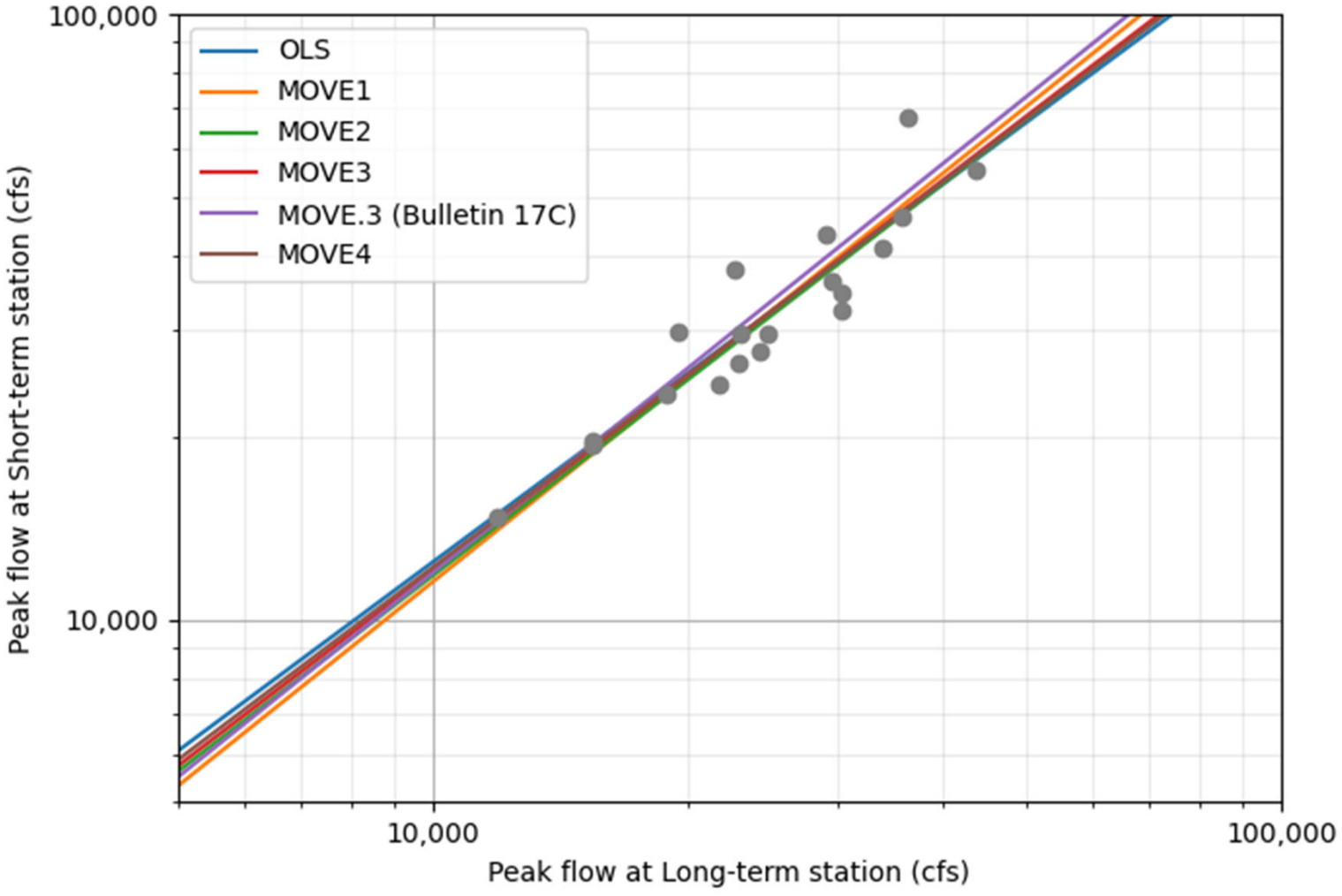


US Army Corps of Engineers

U.S.ARMY

RECORD EXTENSION METHOD COMPARISON

- For many real-world datasets, the various record extension techniques produce similar regression lines



US Army Corps of Engineers

U.S.ARMY

MOVE.3 LINE (BULLETIN 17C)

- Refer to Bulletin 17C for the full equations

$$\hat{Y}_i = a + b(X_i - \overline{X}_e)$$

X = Longer station
Y = Shorter station

$$a = \frac{(N_1 + N_e)\widehat{\overline{Y}_{all}} - N_1\overline{Y}_1}{N_e}$$

N_1 = concurrent record,
N_e = additional record

$$b^2 = \frac{(N_1 + N_e - 1)\widehat{S^2_{Y_{all}}} - (N_1 - 1)S^2_{Y_1} - N_1(\overline{Y}_1 - \widehat{\overline{Y}_{all}})^2 - N_e(a - \widehat{\overline{Y}_{all}})}{[(N_e - 1)S^2_{X_2}]}$$

US Army Corps of Engineers

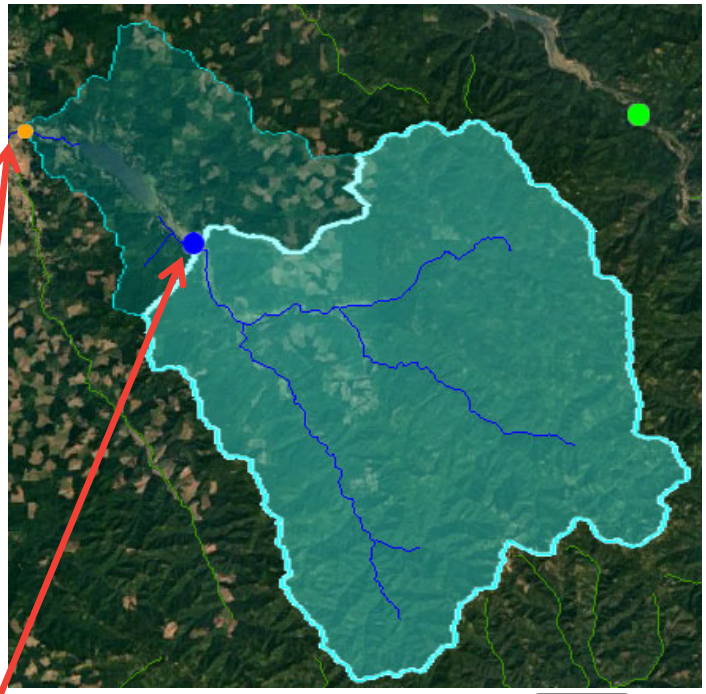
U.S.ARMY

SIDEBAR: DRAINAGE AREA RATIO

- Uses a ratio of the drainage area between two locations to estimate periods of missing flow.

$$Y = X\left(\frac{A_y}{A_x}\right)^{\phi}$$

Y = flow estimate at missing station
X = known flow at long-term station
Ay = Drainage area of missing station
Ax = Drainage area of long-term station
Φ = 1, unless there is a regional regression study



Missing station (downstream):
Drainage Area = 270 sq. miles

Long-term station (upstream):
Drainage Area = 211 sq. miles

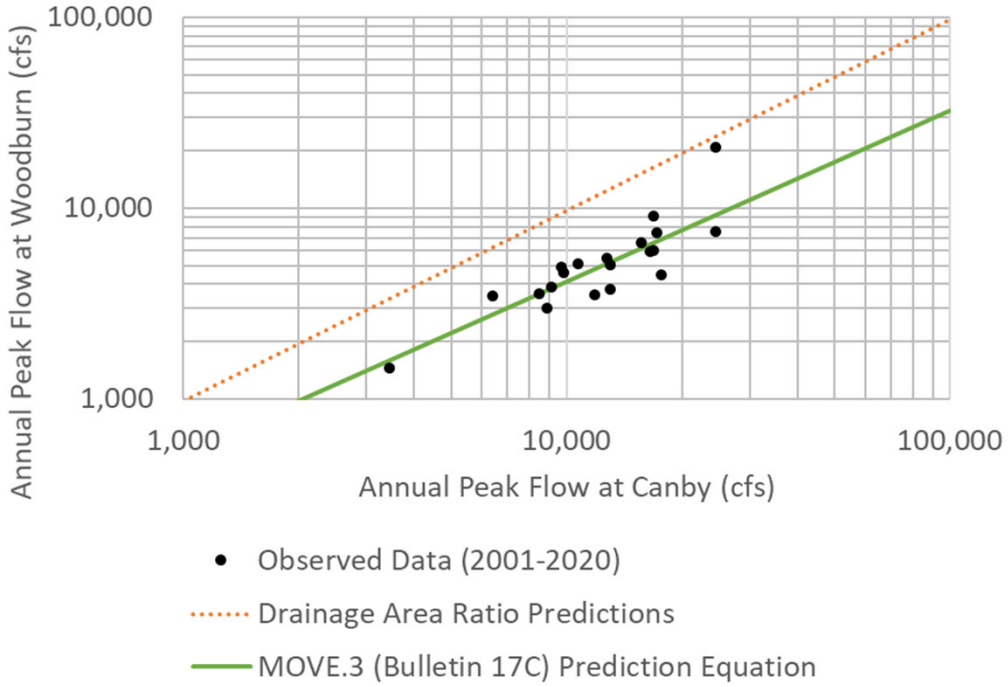
DAR = 270/211
= 1.28

US Army Corps of Engineers

U.S.ARMY

SIDEBAR: DRAINAGE AREA RATIO

- Generally works well when two sites are on the same river, with drainage areas within 50% of each other
- Always use a record extension technique (e.g. MOVE.3) instead of a drainage area ratio when concurrent record is available.
- Can produce poor results if different streams are used.



US Army Corps of Engineers

U.S.ARMY

OUTLINE

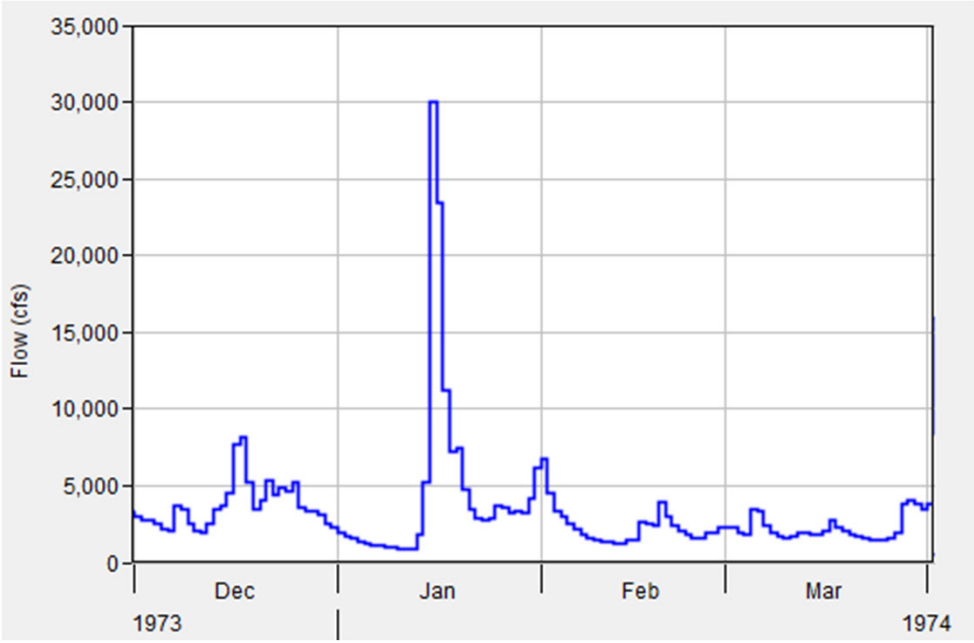
- Problem statement
- Regression basics
- Annual Peak extension
- Continuous extension (e.g. daily)
- Common pitfalls

US Army Corps of Engineers

U.S.ARMY

DAILY RECORD EXTENSION

- MOVE.1 is usually used (Line of Organic Correlation)
- Why do daily extensions?
  - Missing very short periods of time
  - Long-term simulations of reservoir operations
  - Hydropower modeling
- If flood-frequency is the goal, use MOVE.3 from Bulletin 17C instead



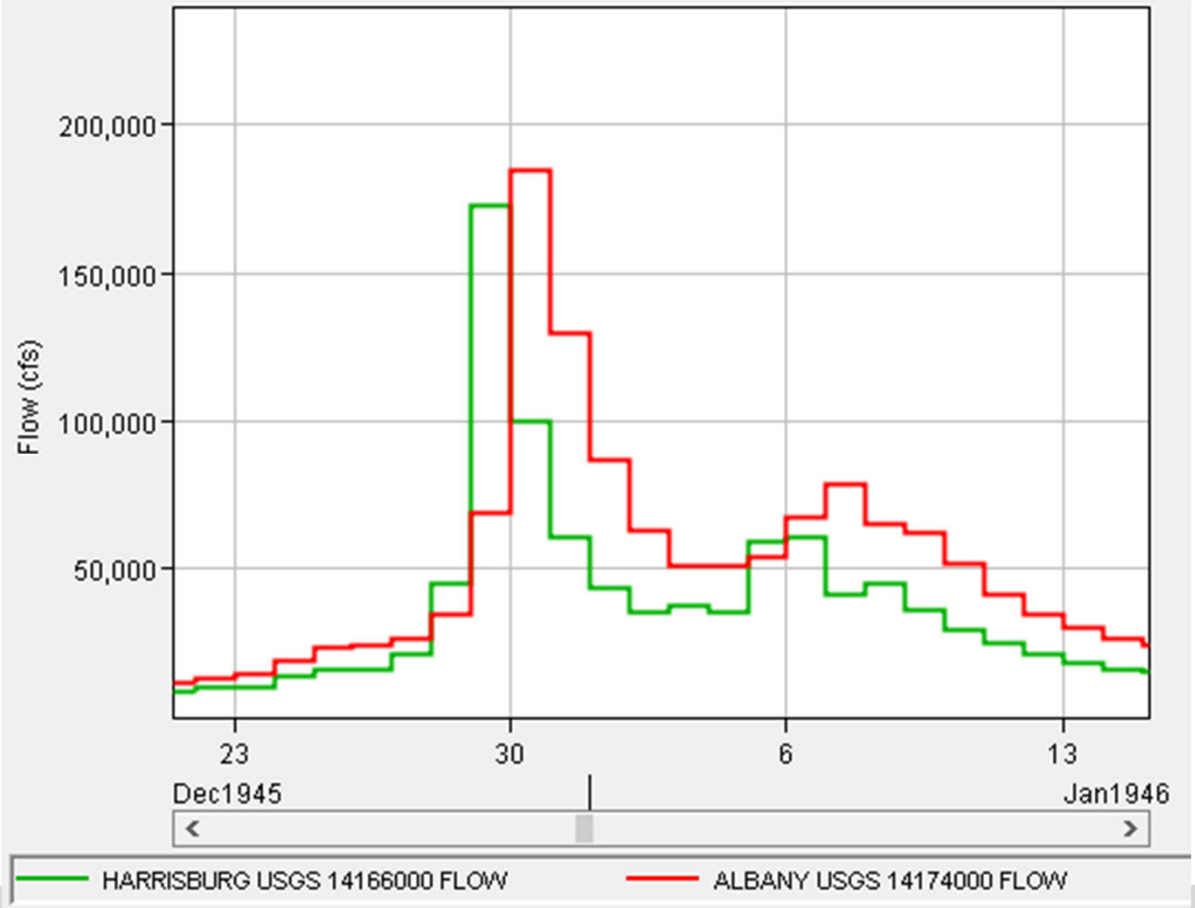
US Army Corps of Engineers

SERIAL CORRELATION

- In the context of time series, the error in a period may influence the error in a subsequent period
- If there are factors (other than the independent variables) making the observation at some point in time larger than expected, (i.e., a positive error), those same factors may linger, creating a positive bias in the error term of a subsequent period.
- Known as positive first-order serial correlation

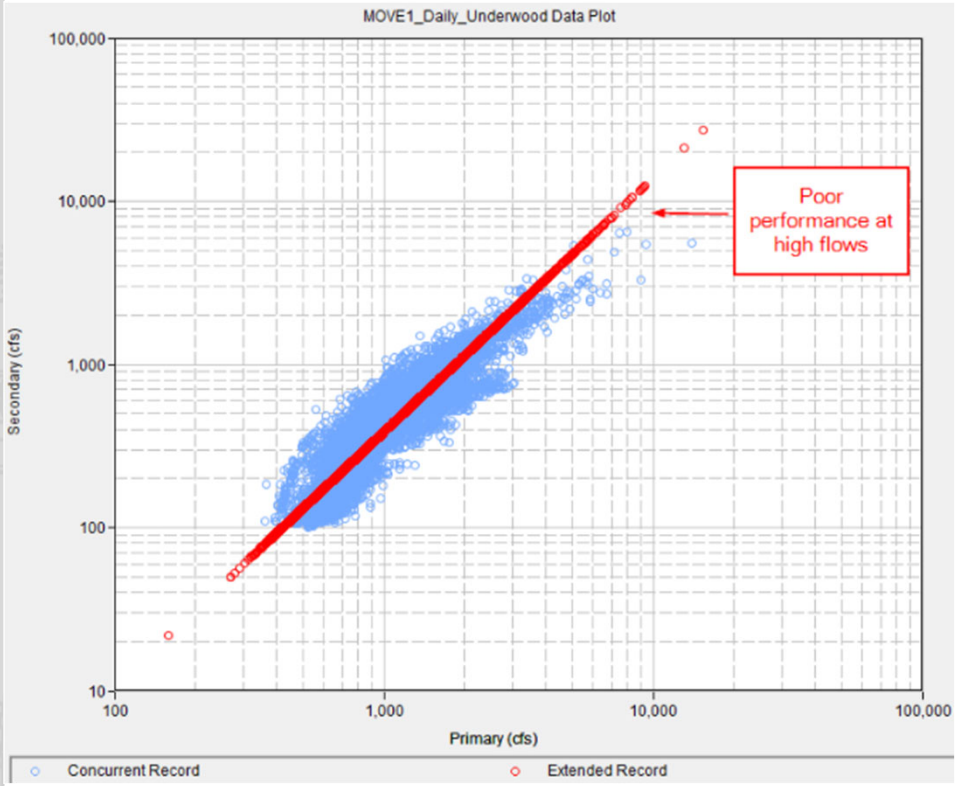
DEALING WITH SERIAL CORRELATION

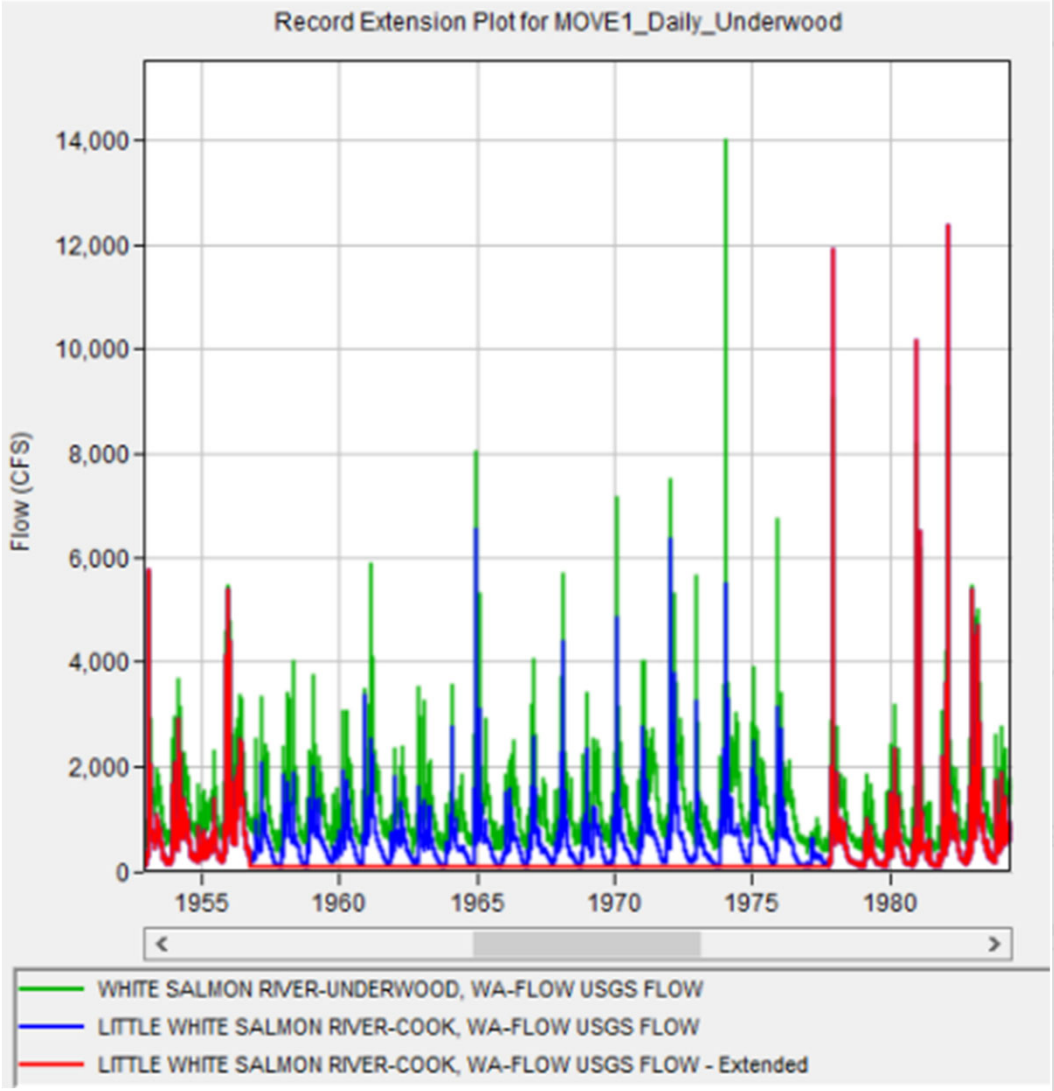
- Test it, using Durbin-Watson
- Break the regression up seasonally
- Can route/lag the data before doing the regression to account for routing effects



US Army Corps of Engineers

DAILY FLOW EXTENSION EXAMPLE





US Army Corps of Engineers

U.S.ARMY

OUTLINE

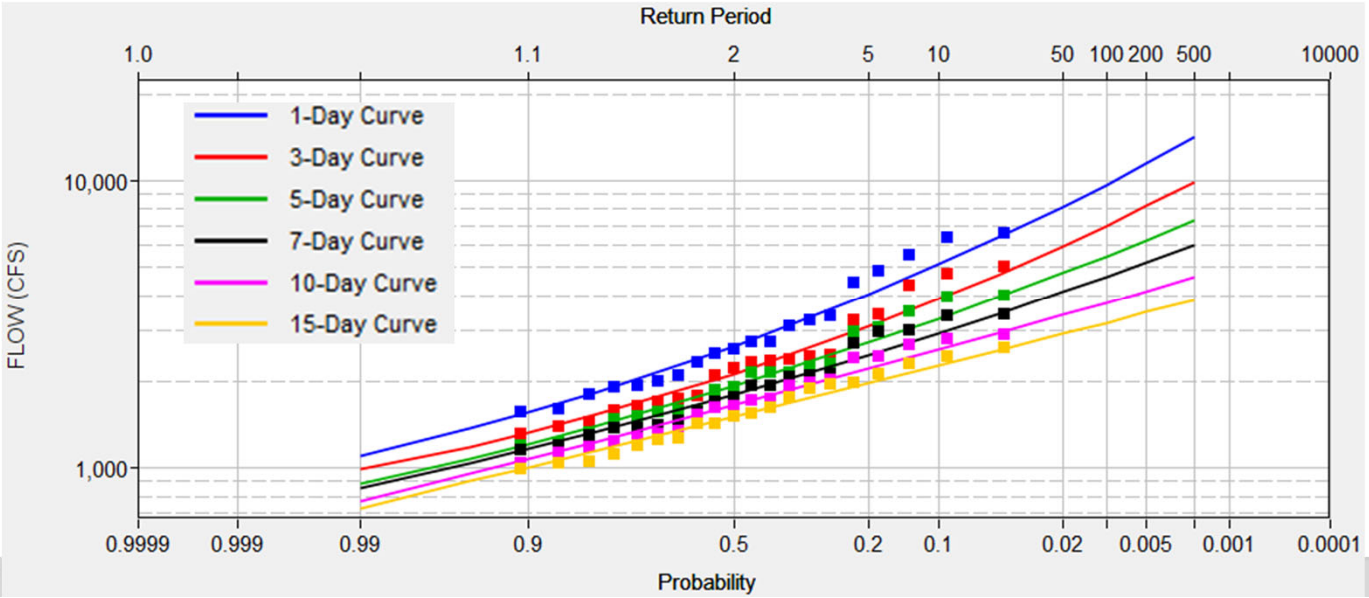
- Problem statement
- Regression basics
- Annual Peak extension
- Continuous extension (e.g. daily)
- Common pitfalls

US Army Corps of Engineers

U.S.ARMY

PITFALL: VOLUME FREQUENCY CURVES

- **Pitfall**: Do a MOVE.1 extension on daily data, and then calculate volume-frequency curves directly from this extended record
  - Regression is not focused on flood events
  - Ignores the concept of limiting the extension by only n_e years
- **Better idea**: use a separate MOVE.3 (Bulletin 17C) record extension for each duration.

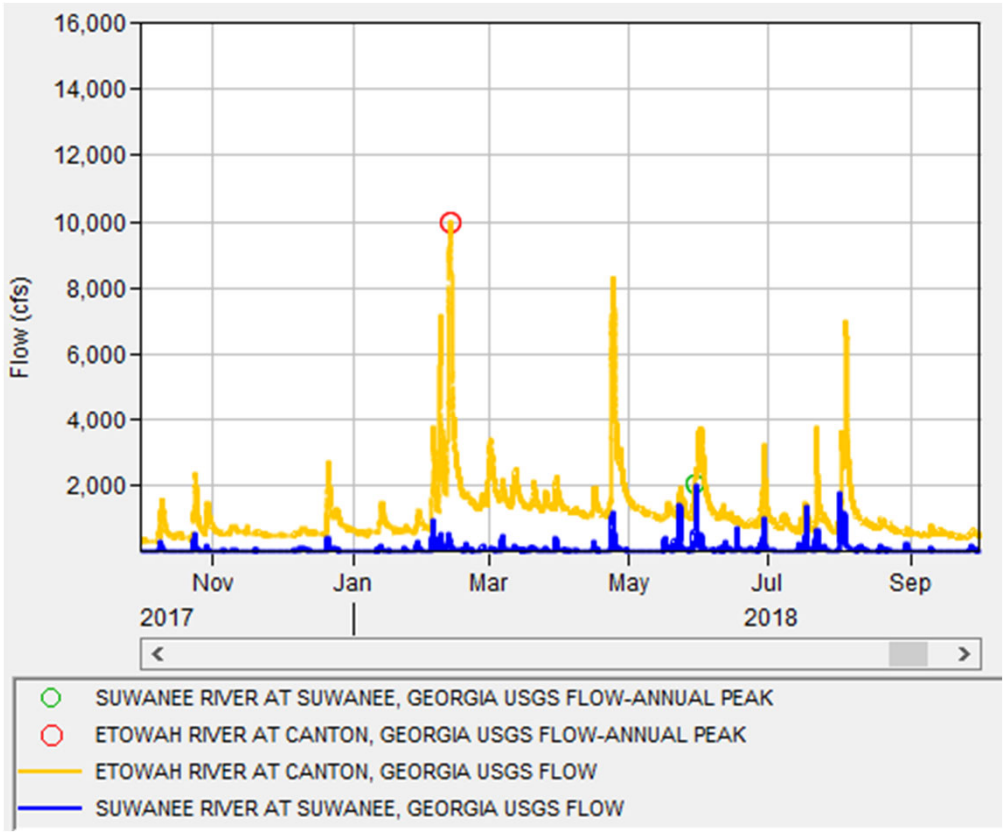


US Army Corps of Engineers

U.S.ARMY

PITFALL: INPUT DATA

- **Pitfall**: Use MOVE.3 record extension with annual peak flow records without examination
  - Peak flow may be generated by completely different storms months apart, hydrologically unrelated

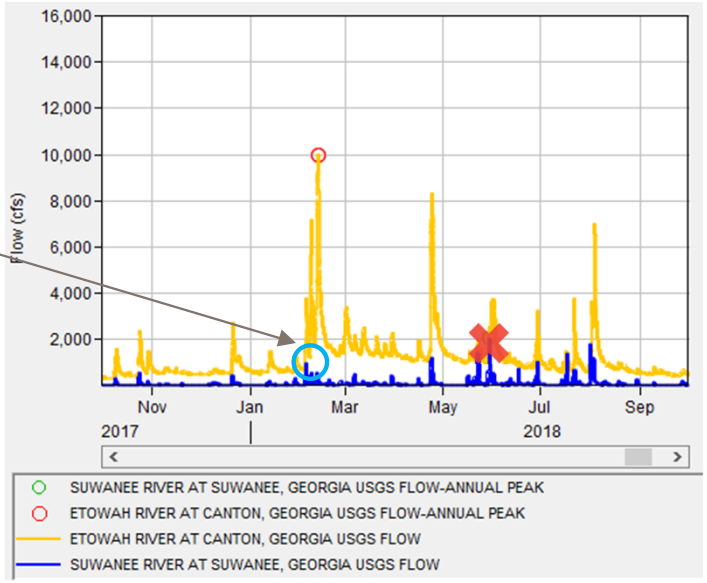


US Army Corps of Engineers

PITFALL: INPUT DATA

- **Better idea**: Examine the input data. Ensure the date of peak flow is from the same storm event. Peaks should be within a few days. Can use a correlation analysis in SSP.
- What about years with date of peak far apart? Can either:
  - Drop the year altogether (less preferred)
  - Try to find short-interval streamflow data from the same storm event and use it in the regression instead.

Use max flow value here instead

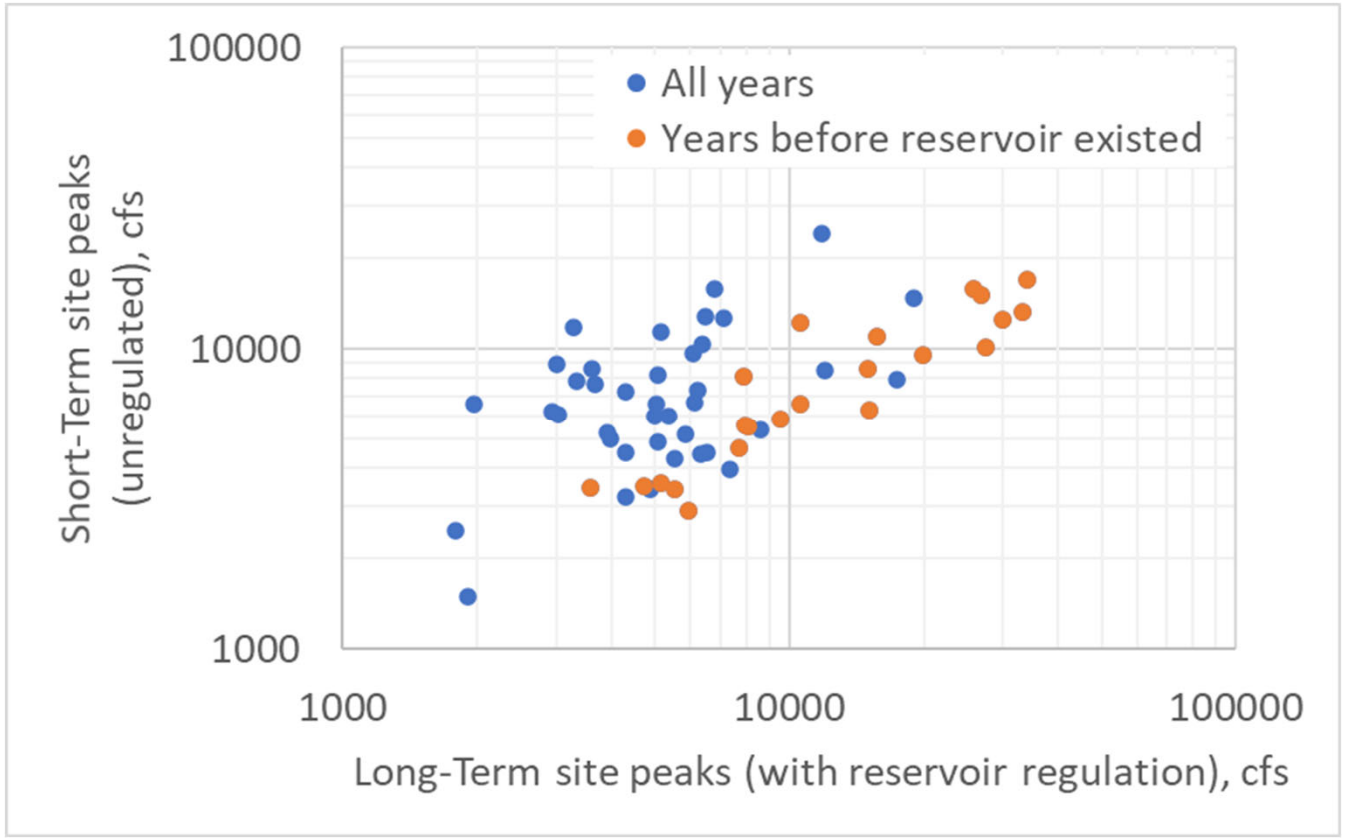


US Army Corps of Engineers

U.S.ARMY

PITFALL: REGULATED DATA

- **Pitfall**: Use record extension techniques at sites heavily affected by upstream reservoirs
- **Better idea**: Only do record extension on unregulated data (both pre-dam and reconstructed post-dam records)

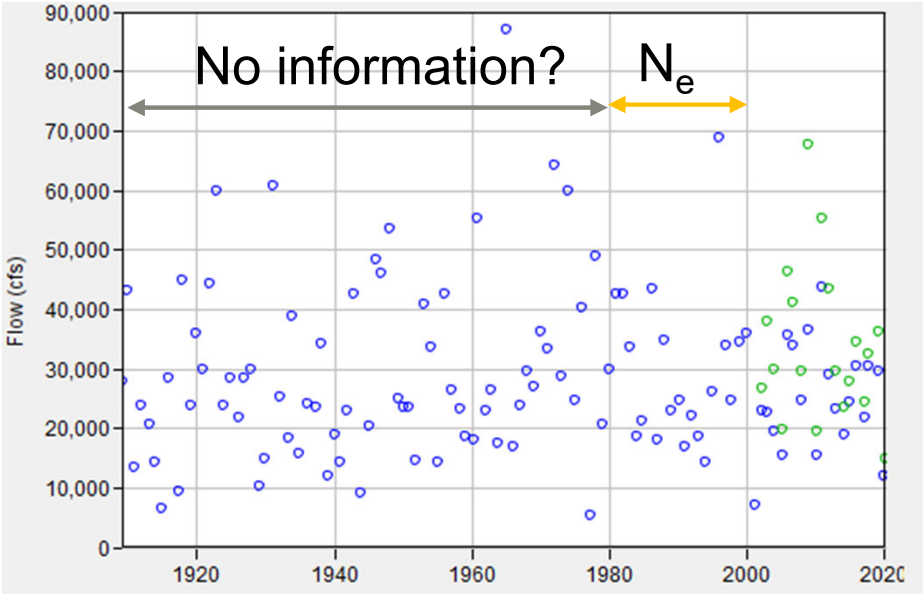


US Army Corps of Engineers

U.S.ARMY

PITFALL: RECORD EXTENSION AS FLOW INTERVALS

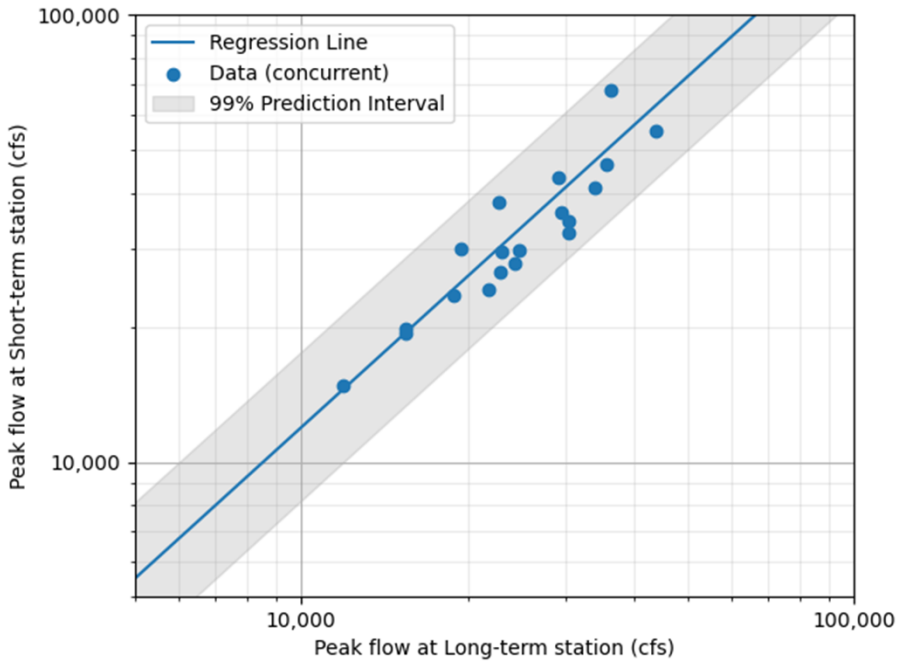
- The n_e years of extended data are input as systematic data to an EMA analysis.
- What about the other years?



US Army Corps of Engineers

PITFALL: RECORD EXTENSION AS FLOW INTERVALS

- **Pitfall**: Represent them as flow intervals
  - Seems appealing, since the selection of which years to use as n_e becomes less important
  - Uncertainty bounds shrink significantly when this is done, which is inappropriate
  - Mean and std deviation also will differ from the Matalas-Jacobs estimators



PITFALL: RECORD EXTENSION AS FLOW INTERVALS

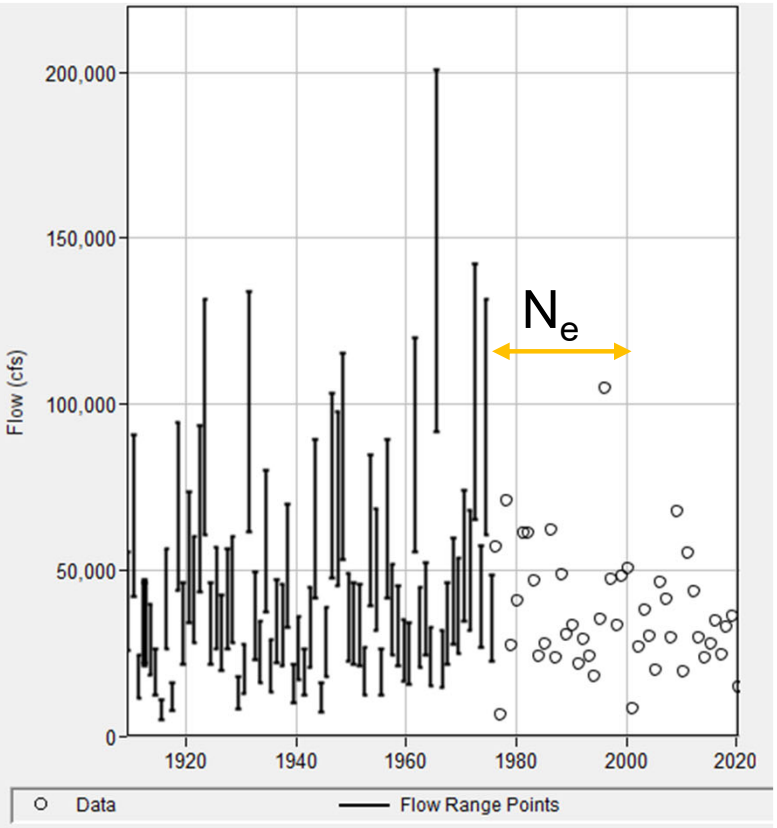
Perception Thresholds

Start Year	End Year	Low Threshold (cfs)	High Threshold (cfs)	Comments
1909	2020	0.0	inf	Total Record

Apply Thresholds

Flow Ranges

Year	Peak (cfs)	Low Value (cfs)	High Value (cfs)	Data Type
1909	37842.0	25816.2	55470.8	Historical
1910	61679.0	41993.9	90591.1	Historical
1911	16604.0	11301.4	24394.7	Historical
1912	31804.0	21699.3	46615.1	Historical
1913	27062.0	18460.4	39671.9	Historical
1914	18002.0	12259.6	26434.6	Historical
1915	7482.0	5047.5	11089.6	Historical
1916	38452.0	26231.6	56366.4	Historical
1917	10963.0	7433.6	16168.5	Historical
1918	64418.0	43845.9	94642.6	Historical
1919	31655.0	21597.3	46396.1	Historical
1920	50073.0	34131.6	73460.7	Historical

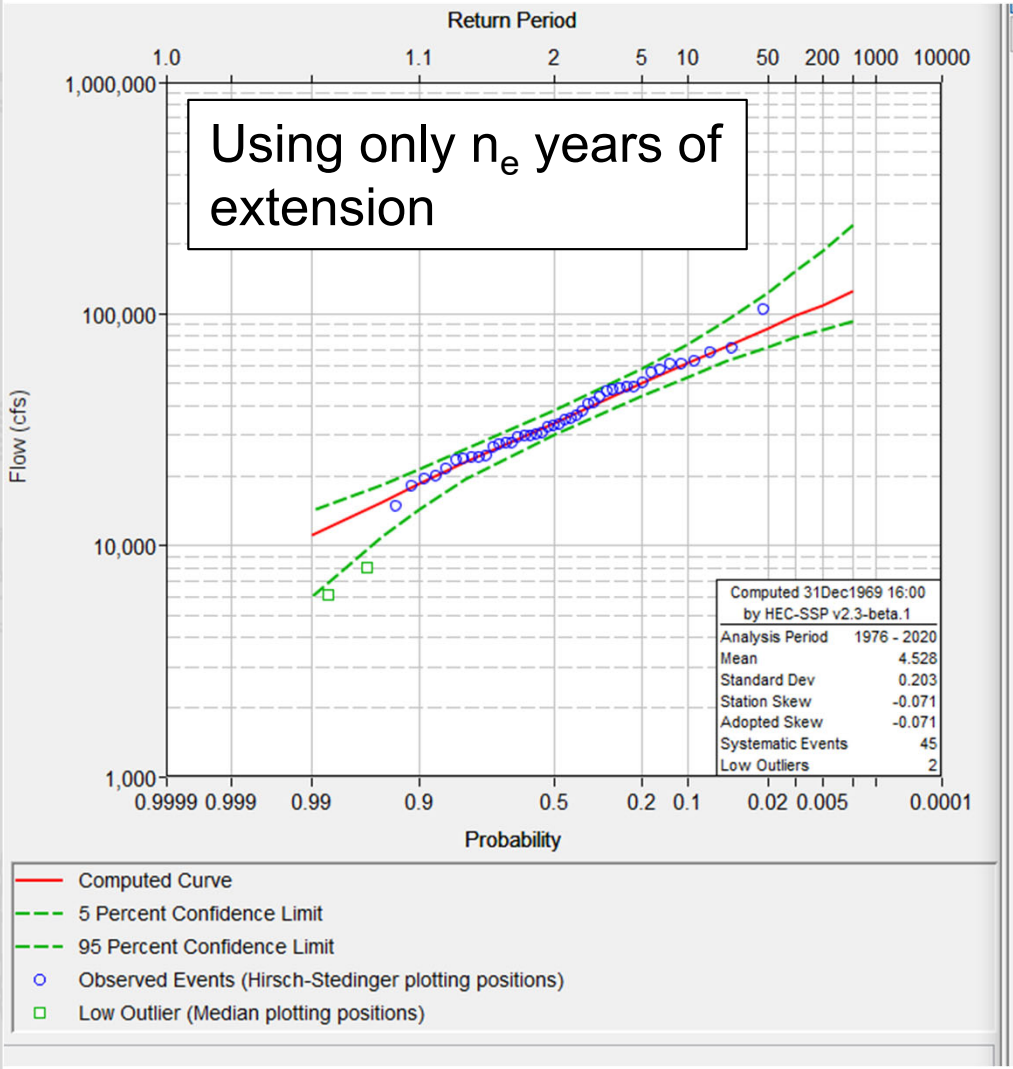


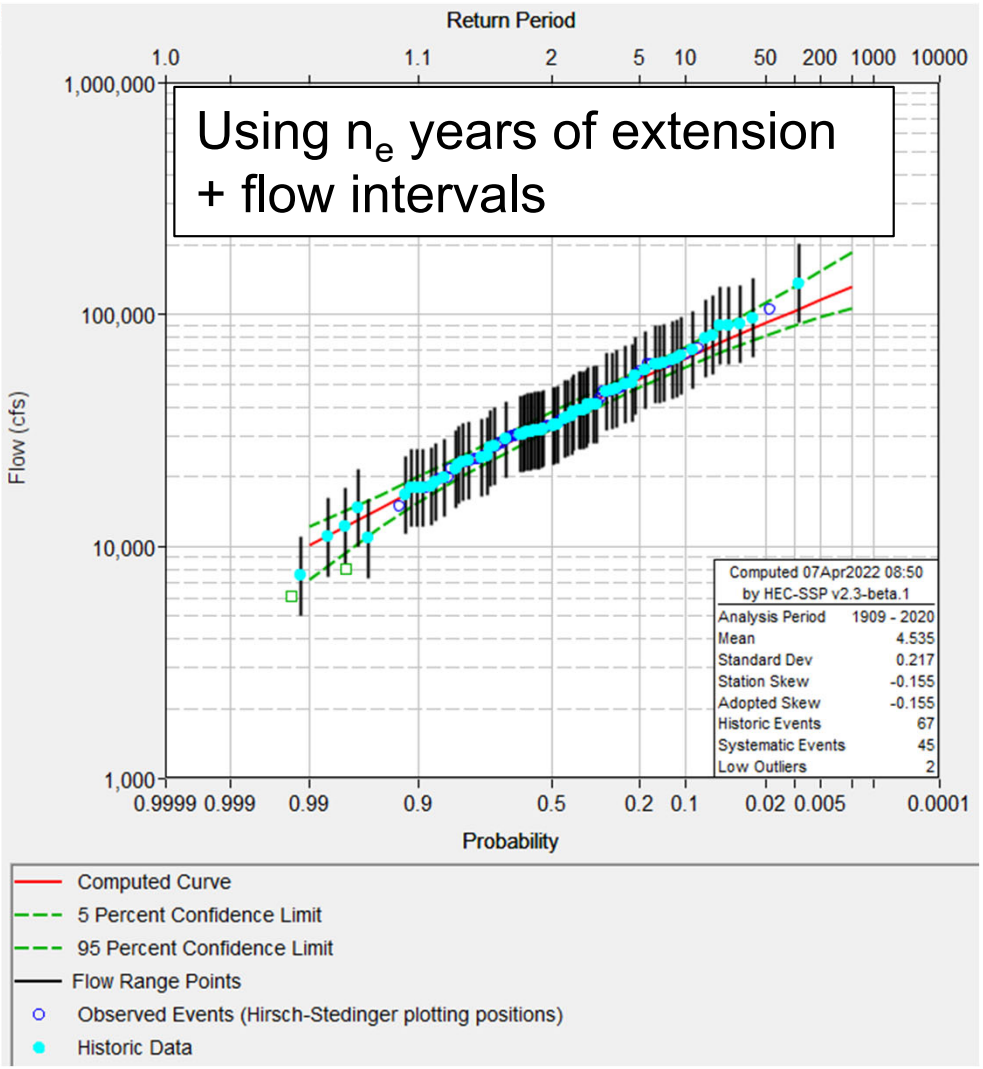
Lower/Upper bounds of 99% prediction interval

US Army Corps of Engineers

U.S.ARMY

PITFALL: RECORD EXTENSION AS FLOW INTERVALS



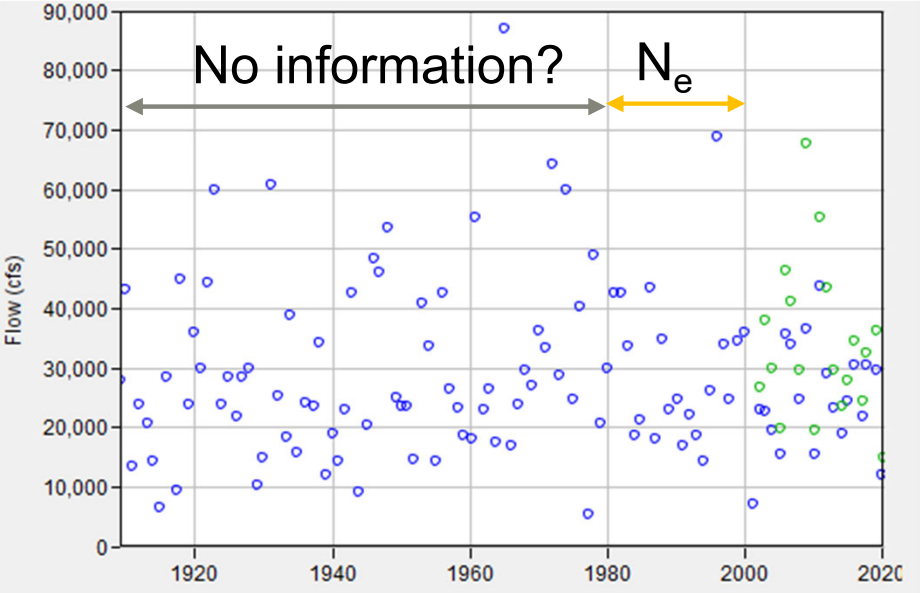


US Army Corps of Engineers

U.S.ARMY

PITFALL: RECORD EXTENSION AS FLOW INTERVALS

- **Better idea**: take care that selection of which n_e years to extend matches the skew from an extension using the full record length



US Army Corps of Engineers

U.S.ARMY

SUMMARY OF ESTIMATION TECHNIQUES

Method	Purpose	Uses
Drainage Area Ratio	Approximate analysis is good enough	Two gages are very close together, minimal effort
Ordinary Least Squares (OLS) Regression	Best individual flow estimates	Someone wants to get the best estimate peak flow for one particular year
Maintenance of Variance Extension (MOVE.1)	Filling in daily flows in an extended period	Water resources planning and management models; reservoir design and operation
Maintenance of Variance Extension (MOVE.3, Bulletin 17C)	Estimate flood peaks for years with missing data	Flood-Frequency Analysis

US Army Corps of Engineers

U.S.ARMY

QUESTIONS

US Army Corps of Engineers

U.S.ARMY

REFERENCES

- Alley, W. M., and A. W. Burns, 1983. Mixed-station extension of monthly streamflow records, J. Hydraul. Div. Am. Soc. Civ. Eng., 109(10), p. 1271-1284.
- Bulletin 17b, 1982. Guidelines for Determining Flow Frequency Analysis: Interagency Advisory Committee on Water Data, USGS, Office of Water Data Collection, Reston, Virginia.
- England, J.F. Jr., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Jr., Veilleux, A.G., Kiang, J.E., and Mason, R.R., Jr., 2018, Guidelines for determining flood flow frequency—Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p https://doi.org/10.3133/tm4B5.

US Army Corps of Engineers

U.S.ARMY

REFERENCES

- Granato, G. 2008, Streamflow Record Extension Facilitator (SREF), USGS Open-File Report 2008–1362.
- Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: Techniques of Water-Resources Investigations of the U.S. Geological Survey, Book 4, Chapter A3. , New York, Elsevier, 510 p. https://pubs.usgs.gov/twri/twri4a3/pdf/twri4a3-new.pdf
- Hirsch, R.M., 1982, A comparison of four streamflow record extension techniques: Water Resources Research, v. 18, no. 4, p. 1081-1088.
- Matalas, N.C. and Jacobs, B., 1964, A correlation procedure for augmenting hydrologic data: U.S. Geological Survey Professional Paper 434-E.

US Army Corps of Engineers

U.S.ARMY

REFERENCES

- Parrett, Charles, and others, 2011, Regional skew for California, and flood frequency for selected sites in the Sacramento–San Joaquin River Basin, based on data through water year 2006, Appendix A: U.S. Geological Survey Scientific Investigations Report 2010–5260, p. 41-85.
- Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis 1—Ordinary, weighted, and generalized least squares compared: Water Resources Research, v. 21, no. 9, p. 1,421-1,432. [with correction , *in* Stedinger, J.R., and Tasker, G.D., 1986, Water Resources Research, v. 22, no. 5, p. 844].
- Vogel, R.M., and Stedinger, J.R., 1985. Minimum Variance Streamflow Record Augmentation Procedures: Water Resources Research V21(5), p715-723.

US Army Corps of Engineers

U.S.ARMY