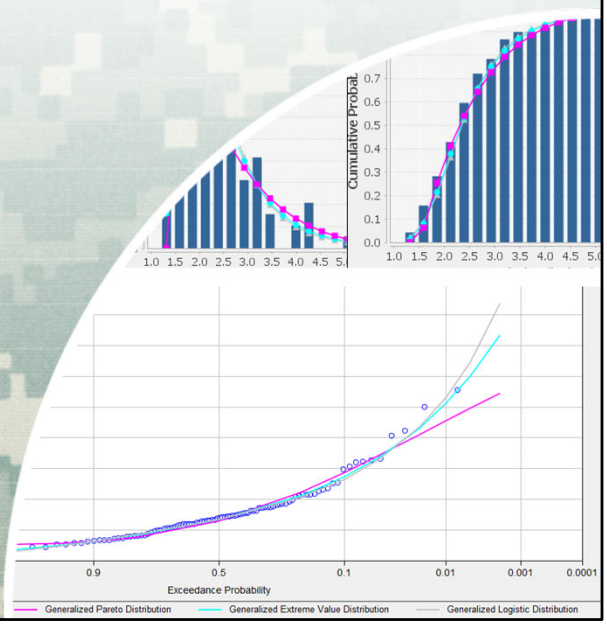
Introduction to Parametric Modeling

Mike Bartles, P.E.

Hydrologic Engineering Center

May 2022



US Army Corps of Engineers
BUILDING STRONG®

- Hey everybody, this is Mike Bartles from the Hydrologic Engineering Center.
- This lecture is devoted to parametric modeling from a hydrologic and hydraulic modeling perspective.

Overview

- Describe parametric modeling, its advantages, disadvantages, and steps.
- Define data requirements and how to develop a data set for analysis.
- Detail commonly used probability distributions.
- Outline fitting methods and parameter estimation.
- Explain multiple ways of validating goodness of fit.
- Briefly describe uncertainty.

- This lecture is meant to serve as an introduction and will not be an exhaustive, detailed look at all the various aspects of parametric modeling.
- We won't be deriving equations or even delving into too much math.
- Instead, we'll talk about the topic from a 30000 ft perspective and focus more on concepts.
- The next few lectures will delve into more detail.

What is a Model?

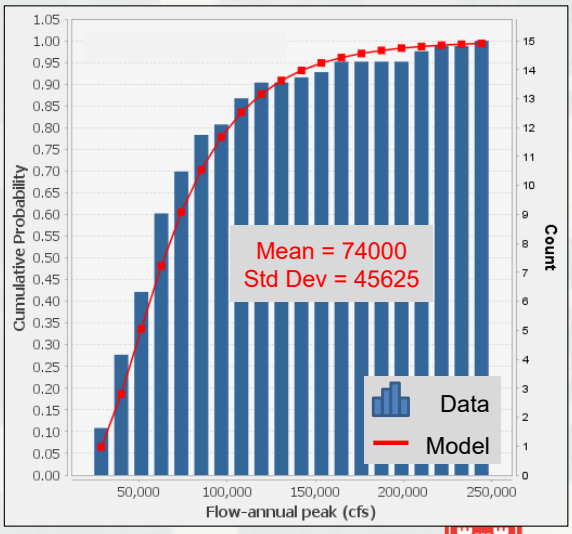
- A model is a "formal representation of a theory" (Ader, Bollen)
- All models are simplifications of the real world
  - Some are better than others…
- Analytical models usually include general principles and a set of statements
- Empirical models usually omit these principles and simply represent the data
  - Cannot extrapolate (at least easily)

Infiltration in the real world…

• First, let's discuss what we mean when we talk about a "model" in a hydrologic or hydraulic modeling context
• A model is a formal representation of a theory
    • Take, for instance, infiltration into soil
    • Infiltration in the real world is complicated, as shown in the image to the right. Water with blue dye was placed on top of this soil and allowed to infiltrate over a few hours. After a while, a backhoe dug this hole and allowed for us to see how far individual strands of dye had progressed downward.
    • A specific representation of infiltration would be the Green and Ampt method which simplifies this process but still preserves important aspects.
• Modeling allows engineers to estimate the behavior of a system that is not otherwise captured in time and/or space. Questions that could be answered with a model include:
    • What will the river flow be 3 days from now?
    • How much runoff will be generated given 20 inches of precipitation over three days?
    • How will the water surface elevation change if a levee were to be constructed?
    • What is the likelihood of streamflow exceeding 20,000 cfs?
• Analytical models, like the aforementioned Green and Ampt method, usually include general principles and a set of statements. This allows for analytical models to be extrapolated beyond what has been observed, which is of particular interest to us in this profession given that we're often times interested in extreme events, like the 1/100 annual exceedance probability for floodplain studies or the probable maximum flood for dam safety studies.
• Conversely, empirical models usually omit any principles and simply use the data. This is analogous to interpolating between known points. However, this leaves us high and dry when we want to extrapolate because you can't extrapolate with empirical models or at least it's not easy or straightforward.

What is Parametric Modeling?

- **Parametric Modeling** is equivalent to **Analytical Frequency Analysis**
- The result of this type of analysis is a **fully parameterized probability distribution** (with parameters than are in finite-dimensional space)
  - *Log Normal distribution with defined mean and std dev*
  - Provides median values

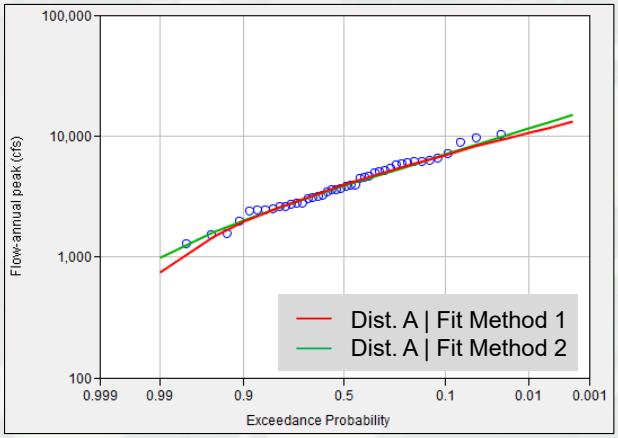


BUILDING STRONG®

• Parametric modeling and analytical frequency analysis are equivalent and can be used interchangeably
• The result of this type of analysis is a fully parameterized probability distribution and median values (additional steps can be taken to estimate confidence limits and mean estimates)
• For something to be parametric, the parameters must be located in “finite dimensional space”; they cannot be imaginary
• An example of an analytical model is shown on the right, which visualizes the cumulative distribution function of a fully parameterized log-normal distribution. The two parameters of this distribution are a mean and standard deviation. The median values are signified by the red line. Observed data is shown with the blue bars.

What is Parametric Modeling?

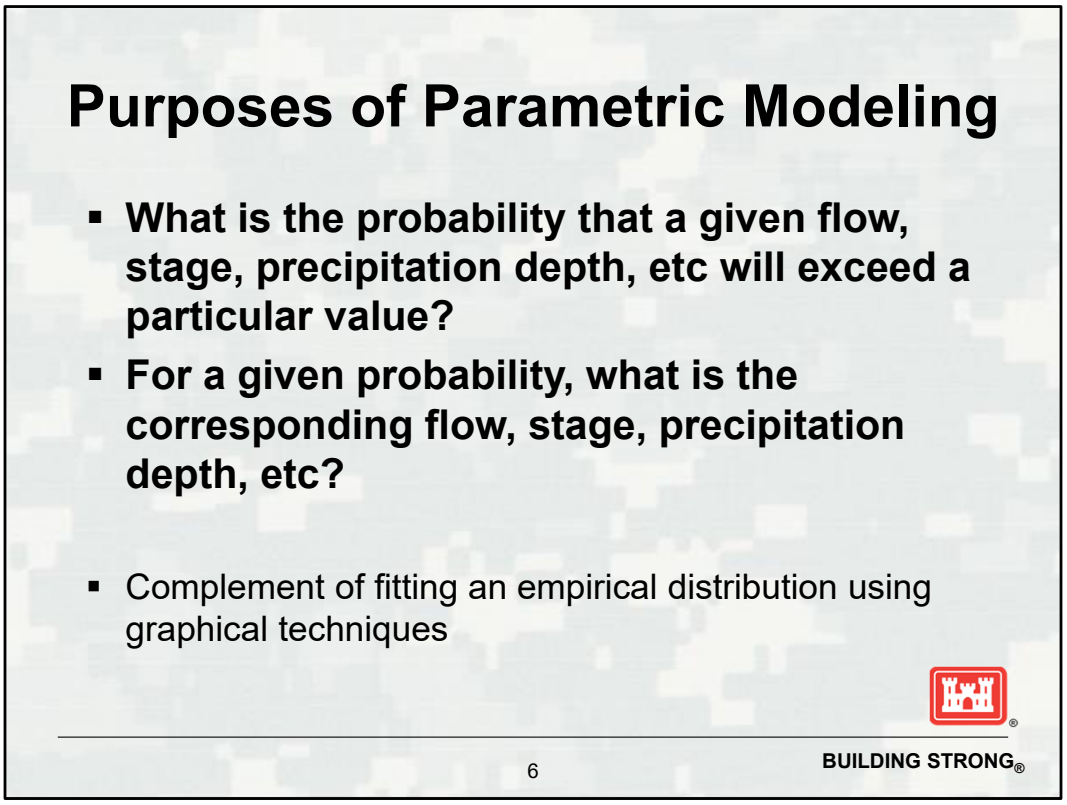
- Requires a combination of **analytical distribution** and **fitting method**
- Fitting method is used to parameterize the analytical distribution using data



5

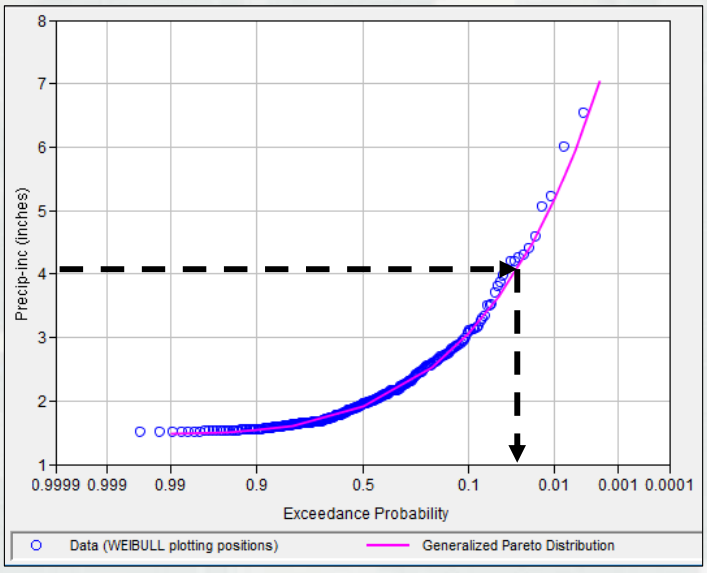
BUILDING STRONG®

• A “model”, in the sense that we’re interested in within this set of lectures, consists of both an analytical distribution and a fitting method.
• The same analytical distribution can be parameterized using two different fitting methods to produce a different parameterization. This is shown on the right. The green and red lines are both the same analytical distribution (for example, Generalized Extreme Values), but they have different parameterizations because they were fit to the blue observed data using different fitting methods.
• I’ll describe some of the more commonly used analytical distributions and fitting methods in a future lecture.
• The complement of parametric modeling is fitting an empirical distribution using graphical techniques. That would be the equivalent of drawing a line of “best fit” through the blue observed data in the image to the right. However, like I said before, you can’t easily extrapolate beyond the range of the observations.



• Parametric modeling is performed all the time by tons of hydraulic engineers and modelers all across the world.
• Two examples of a commonly asked questions that can be answered using parametric modeling are:
    • "what is the probability that a quantile will exceed some value?" and
    • "given some probability, what is the corresponding quantile?"

Purposes of Parametric Modeling

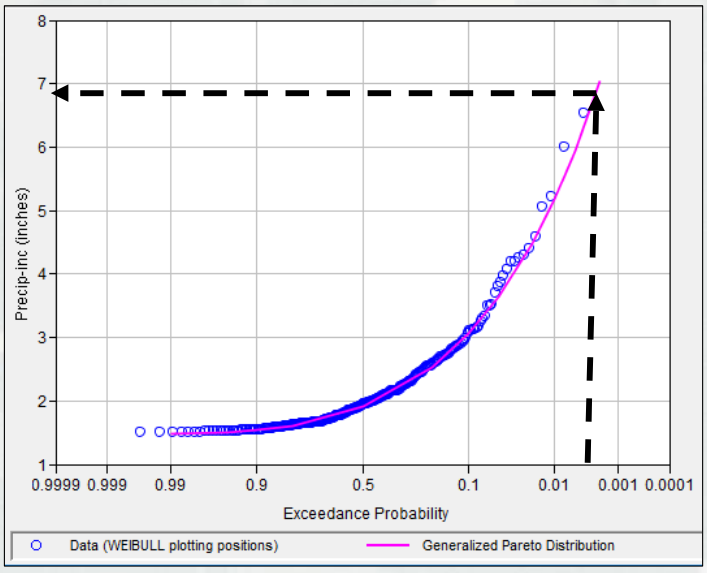


7

BUILDING STRONG®

• This image is an instance of the first question: “what is the probability that a quantile will exceed XXX?”
• Start on the y-axis at some important value. For example, 4 inches of precipitation. Then, trace a horizontal line until you intersect the model. Then, trace a vertical line down to estimate a corresponding probability.
• In this case, there is an approximate 1/50 annual chance of exceeding 4 inches of precipitation.

Purposes of Parametric Modeling



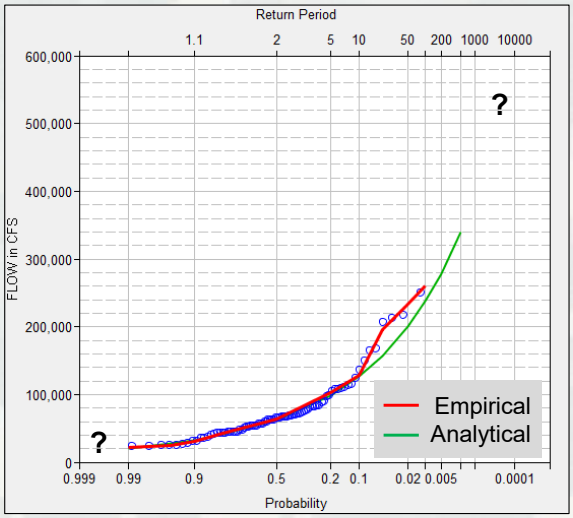
8

BUILDING STRONG®

• This image is an instance of the second question: “given YYY probability, what is the corresponding quantile?”
• This time, start on the x-axis at some important probability. For instance, the 1/500 exceedance probability. Then, trace a vertical line upwards until you intersect the model. Then, trace a horizontal line to estimate a corresponding quantile.
• In this case, the 1/500 annual exceedance probability precipitation is approximately 6.75 inches.

Advantages of Parametric Modeling

- **Extrapolation throughout the entire range of probability**
- Allows for regionalization of parameters
- Analytical means to compute confidence limits
- Process the data using numerical techniques
  - No “eyeballing” it
- Provides a consistent procedure and uniform estimates
  - Important for the NFIP!



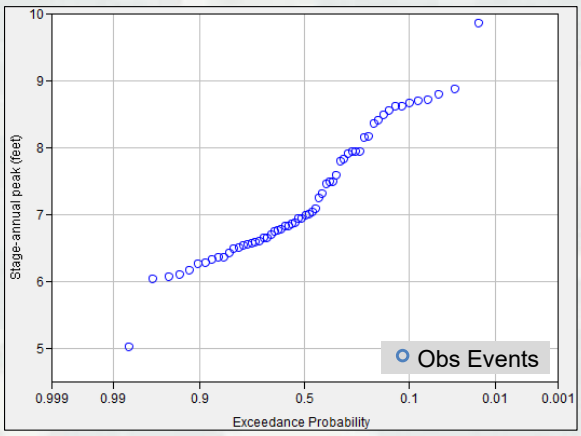
9

BUILDING STRONG®

• Parametric modeling has many advantages when compared to graphically fitting an empirical distribution.
• The biggest advantage is that a fully-parameterized analytical distribution allows the user to estimate quantiles throughout the entire range of probability from 1 -> 0.
• Also, parameters can be “pooled” and regionalized to improve estimates everywhere.
• Confidence limits can be computed.
• The calculations that are used to fit an analytical distribution are tractable and repeatable by other engineers. This isn’t the case when graphically fitting an empirical distribution.
    • This becomes incredibly important when comparing, say, 1/100 annual exceedance probability floodplains in the eastern United States against those developed on the west coast. They better mean the same thing in order to prioritize funding for infrastructure or management!

Disadvantages of Parametric Modeling

- **Must assume a model up front**
- **Some types of data will not be well fit**
- May provide a false sense of accuracy
  - Precision vs. Accuracy
- Can take more time than graphical techniques



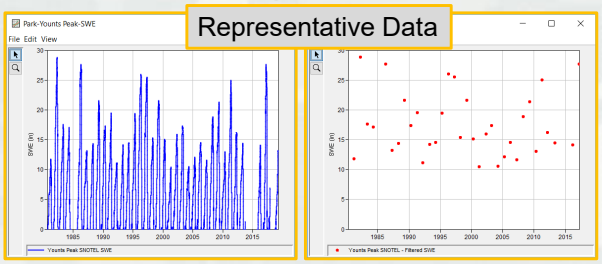
10

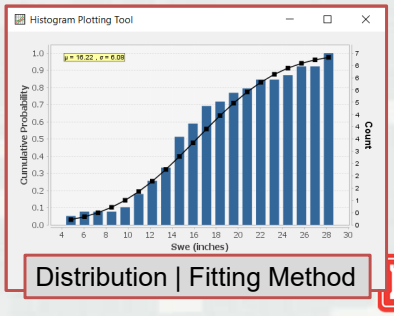
BUILDING STRONG®

• However, there are some disadvantages to parametric modeling when compared to graphically fitting an empirical distribution.
• First and foremost, a model must be assumed up front. We'll discuss the implications of this assumption in greater detail later.
• Some data will not be well fit using commonly employed models. This includes regulated data or stages, which are shown to the right. Notice the sharp discontinuities and multiple changes in slope throughout the full range of probabilities. It'll be really hard to fit a meaningful analytical distribution to this data.
• A false sense of accuracy can also be inadvertently implied when using parametric modeling. This is analogous to reporting precision to the millionth decimal place.
• Finally, computations can take more time than graphical techniques.

Parametric Modeling Steps

1. Develop a representative data set
2. Select a probability distribution and fitting method (“model”)
3. Compute parameters of the model using the data (i.e. fit the distribution)
4. Verify appropriateness of the model
5. Use model to predict quantiles of interest



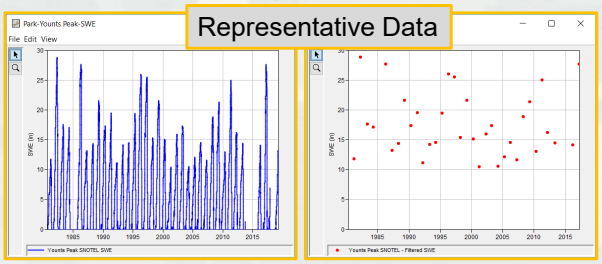


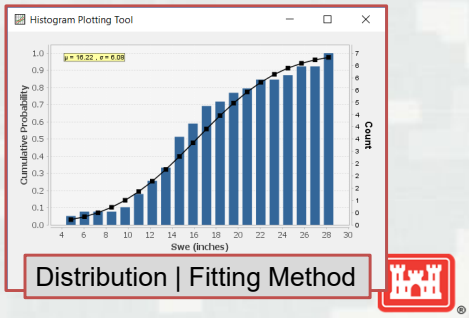
BUILDING STRONG®

•Now, let’s introduce the steps that are used when performing a parametric modeling exercise or fitting an analytical distribution.
• First, you must develop a representative data set.
• Second, you must select a probability distribution and fitting method, which constitutes a “model”.
• Third, you have to fit the selected distribution to the data in order to compute the parameters of the distribution.
• Fourth, you must verify the appropriateness of the model given the sample.
• And finally, you can then use the model to predict variables (i.e. quantiles) of interest.
• Each of these points will be discussed in greater detail within the next few slides.

Parametric Modeling Steps

1. Develop a representative data set
2. Select a probability distribution and fitting method (“model”)
3. Compute parameters of the model using the data (i.e. fit the distribution)
4. Verify appropriateness of the model
5. Use model to predict quantiles of interest

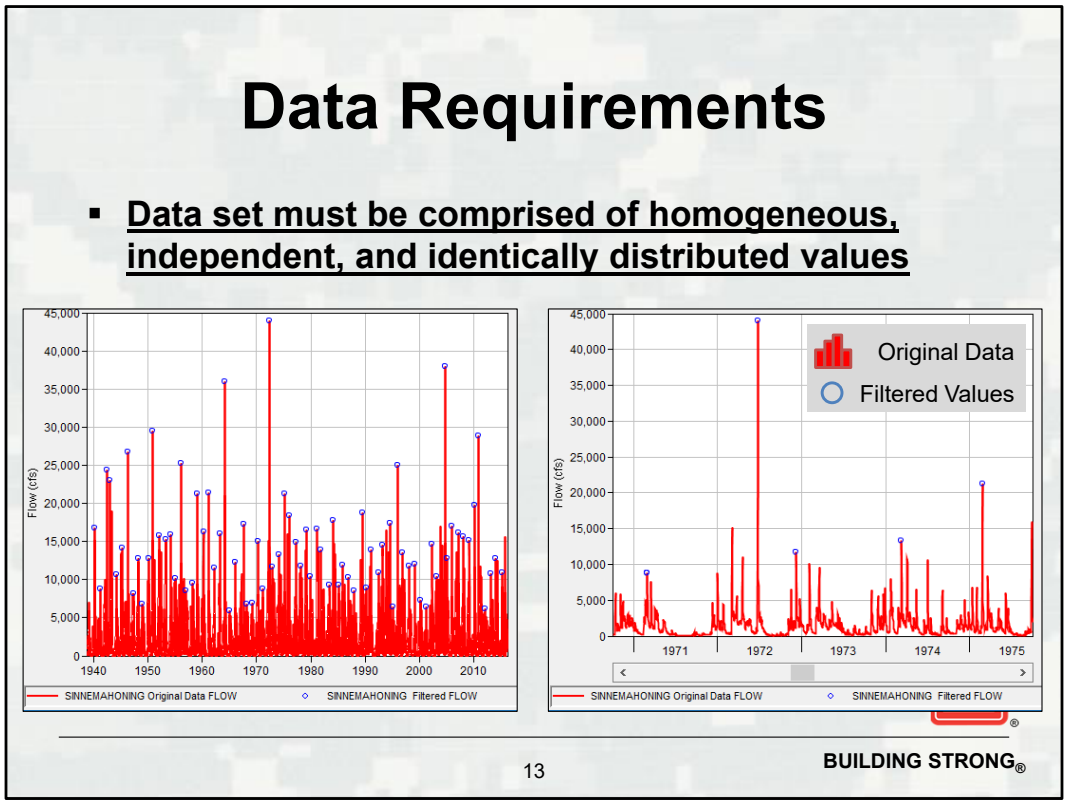




12

BUILDING STRONG®

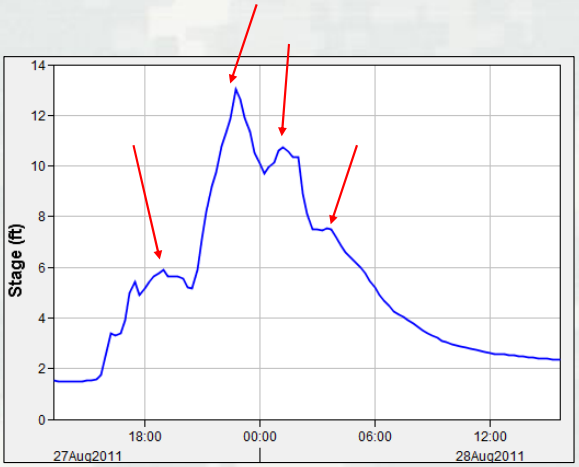
•Developing an adequate data set is foundational to any parametric analysis. You’ve heard the common phrase “garbage in, garbage out”. That can’t be more true with parametric analyses. If the data is junk, your conclusions will be as well.



• When performing a parametric analysis, the data must be a representative sample of the "parent" population and be comprised of homogenous, independent, and identically distributed values.
• Throughout these lectures, I'll refer to the representative sample, which is what we typically have to analyze, as a "child" population
• To be representative, the sample should be a random sample of possibilities from the parent population, which accounts for natural variability.
• The data should not include, or at least minimize, the effects of things like:
    • changing land use (including increasing/decreasing urbanization)
    • regulation, for example reduction in stages/flow due to upstream reservoir and/or diversions
    • large climactic variations/oscillations
• There are other data considerations, but these are the big ones that we typically run into as modelers and engineers.

Independent and Identically Distributed (IID)

- **Independent**
  - Is the magnitude of one peak dependent upon another?
  - How many independent peaks are there in this stage hydrograph?
    - *Frankford Creek at Castor Ave, Philadelphia, PA*



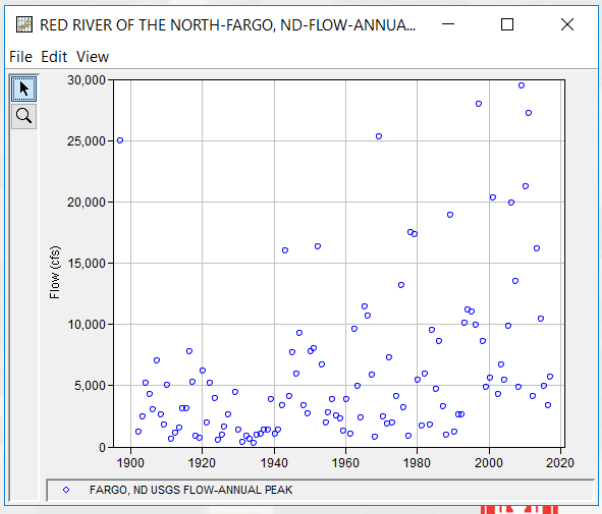
14

BUILDING STRONG®

• The previously mentioned independent and identically distributed, or IID, data assumption is super important. Let's talk about the first part, which is independence.
• When assessing independence, ask yourself: Is each peak unique and independent? Or are the magnitudes dependent upon one another?
• In the image to the right, certainly the first peak is unique, right? But are the others truly independent? The third and fourth peaks aren't. What about the largest peak? Is that independent?
• In the realm of hydrology and hydraulics, we typically look at the entire hydrograph shown here as a quote/unquote "event" and extract a peak for the entire thing.
• The largest peak for the event would be the second peak denoted in the image to the right.
• We typically extract maxima for the entire calendar year or water year in an effort to achieve independence because most watersheds settle back to a quote/unquote "normal state" within a year.
    • But, be careful with peaks that are close to your year demarcation. For instance, when using a water year which begins on October 1st and ends on September 30th, be vary cautious to avoid using peaks at the end of September or beginning of October because you might not be able to ensure that they're independent.

Independent and Identically Distributed (IID)

- **Identically distributed**
  - The sample (child population) is randomly drawn and is representative of the parent population
    - *Red River of the North at Fargo, ND*



15

BUILDING STRONG®

• Similarly, the child population must be an identically distributed sample of the parent population.

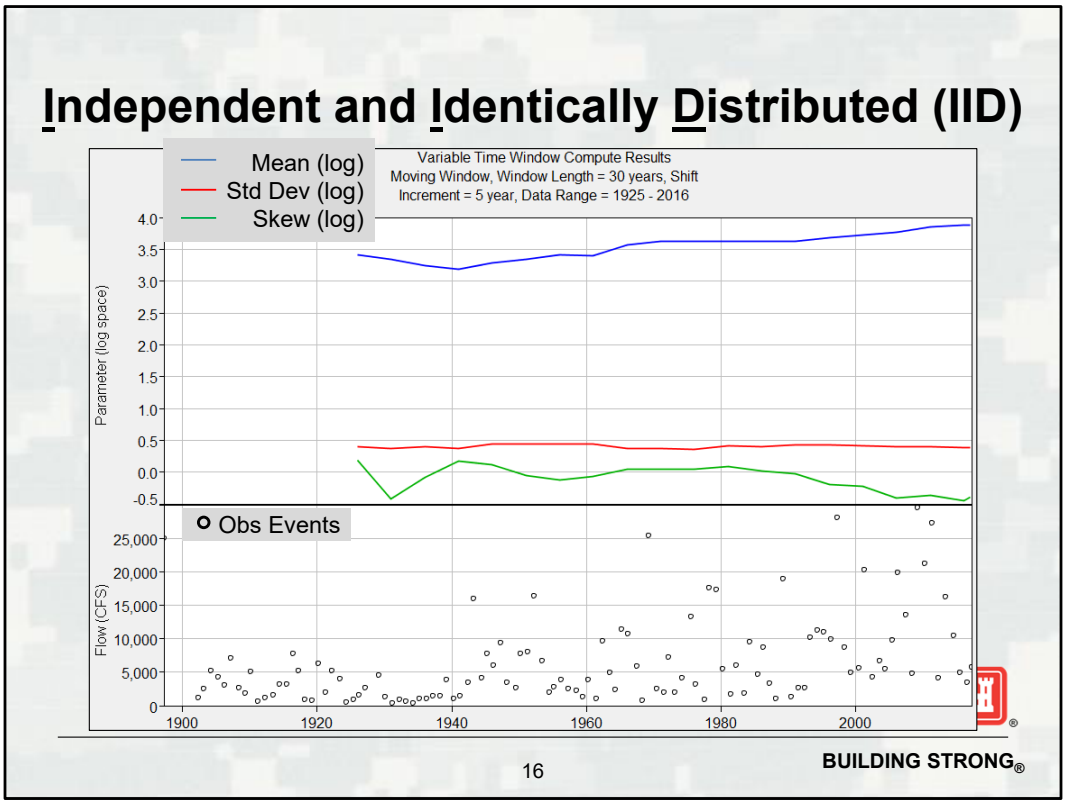
• This means that the sample should be randomly drawn from the parent population and not, or at least minimize, sampling bias

  • If climate varies over time (which it can sometimes do) and the sample is pulled from one time period that isn't representative of the entire range of climactic possibilities, then the data will likely not be identically distributed

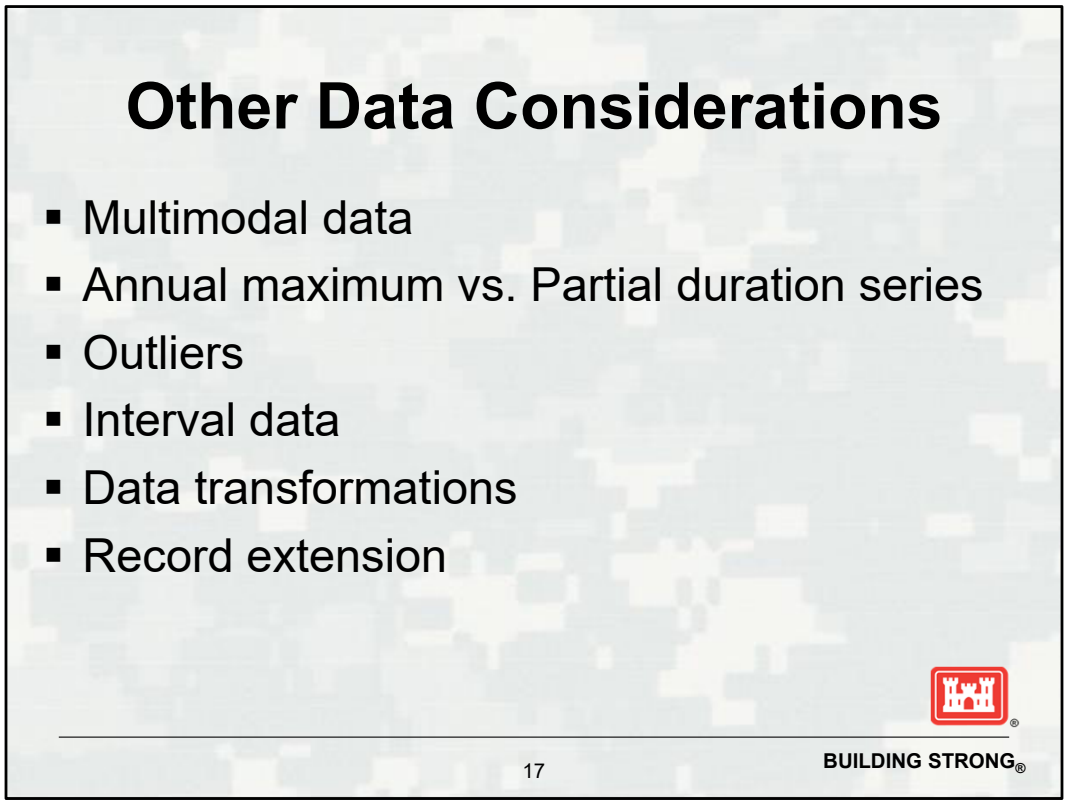
• A classic example of this phenomena is exemplified in the image to right, which is the annual maximum flow time series for the Red River of the North at Fargo, ND

  • Climactic variations are evident in the sample. Visualize the average of the data. You can see that the average, or mean, appears to be increasing over time. Also, look at the variability, or highs and lows. They also appear to increase over time.

  • How do you fit a model to this data set? That's a very difficult question to answer adequately.



• An analysis which used a moving time window was computed for the Red River of the North at Fargo, ND and shown here.
• In this example, a window length of 30 years was used. After the statistics of the first 30 year time window were computed, the time window was moved forward by 5 years and the next 30 year time window was computed. This process was repeated until the entire data set was analyzed.
• Notice how the mean, which is the blue line, standard deviation, which is the red line, and skew, which is the green line change pretty dramatically over time.
• Is the entire 1925 – 2019 time window homogeneous and identically distributed over time? Probably not.
• But, a follow-up question for you to ponder is, how many violations of the IID assumption are you willing to live with when doing parametric modeling? It's not too hard to nit pick and find little violations of that assumption in most analyses. In fact, many of the most-commonly used parametric modeling approaches, for instance Bulletin 17C procedures, includes some tools that are intended to minimize the negative consequences of this common violation.

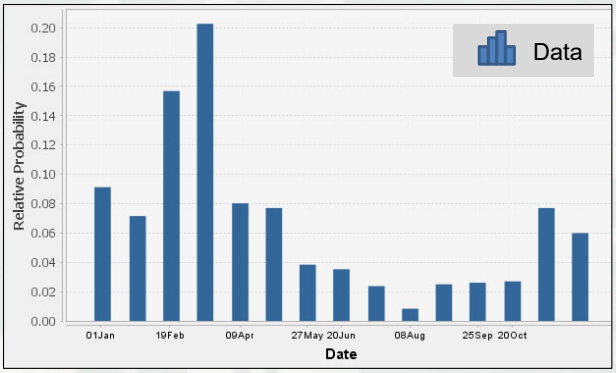


• Some other considerations to keep in mind when analyzing data for a parametric analysis include the following:

- Multimodal data
- Annual maximum vs. Partial duration series
- Outliers
- Interval data
- and Data transformations. All of the aforementioned topics will be discussed in this video.
- However, record extension is such a voluminous topic that this warrants its own lecture/video series.

Multimodal Data

- Could be caused by multiple mechanisms:
  - Rain and snow floods
  - Early and late season floods
  - Etc
- Usually indicative of a mixed population
- Should be separated into different mechanisms and combined in a mixed population analysis

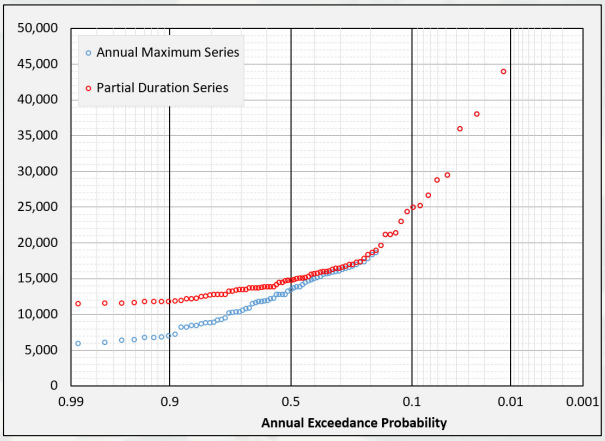


BUILDING STRONG®

• First up, is multi-modal data. An example of a multi-modal data set is shown on the right as a probability density function. Notice the multiple peaks in this plot? They're all separate modes.
• Multimodal data is likely caused by a mixed population. If this situation is encountered, you definitely should work on separating out the various causal mechanisms into individual data sets.
• Mixed populations can be caused by, for example, a location being subjected to floods emanating from rain-on-snow events, summer thunderstorms, extratropical storms (i.e. nor'easters), and tropical storms.
• It's unreasonable to expect that a single model could fit and accurately predict the probability distribution of floods due to all of these mechanisms.
• To best predict the exceedance probabilities of flood quantiles of interest (i.e. what is the likelihood of exceeding XXX cfs in any given year?), it is best to split the data set into the individual mechanisms, fit a model to each data set, and combine the results into a single probability distribution using a mixed population analysis.
• We'll talk more about how to do that in a later lecture.

Annual Maximum vs Partial Duration Series

- One event per year (AMS) or multiple events per year (PDS)
- Fitting an analytical distribution to a partial duration series can be difficult
  - Usually leads to the skew parameter being overestimated
- **95% of the time, use an annual maximum series**



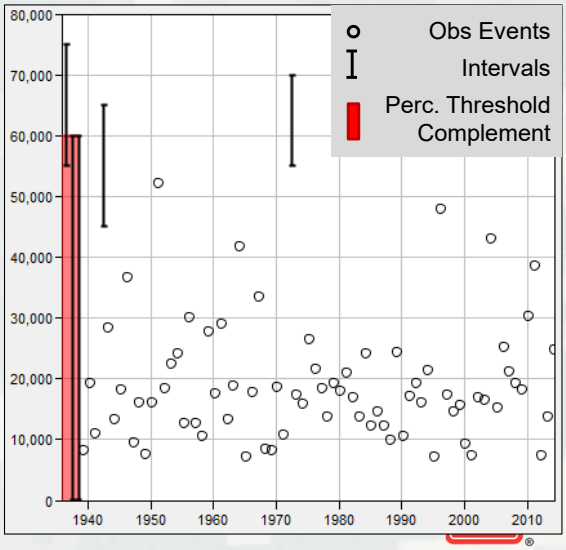
19

BUILDING STRONG®

• An annual maximum series contains one and only one value per year. This is the most commonly utilized type of data set when estimating flow-frequency with, say, Bulletin 17C procedures.
• Conversely, a partial duration series can contain more than one value per year. This type of data is commonly used to estimate precipitation-frequency.
• When plotted together on a normal probability axis, these two types of data, for the same location and time window, commonly merge somewhere between the ½ and 1/50 annual exceedance probability.
• Just to reiterate, the 95% use case when analyzing large or extreme things will be to use an annual maximum series.

Interval Data

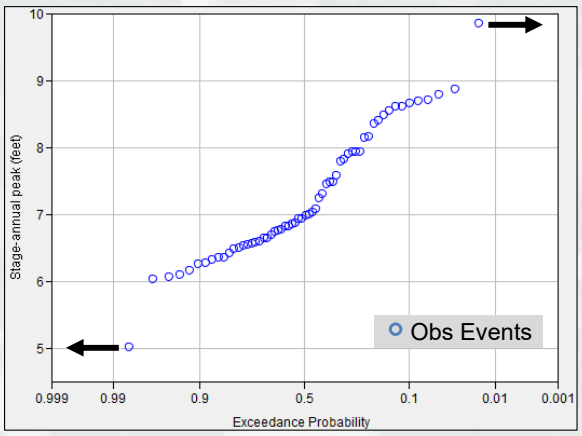
- Instead of using a specific value, use a range of possible values
- Expected Moments Algorithm and Maximum Likelihood Estimation can incorporate this type of data



• Up to this point, we've been discussing point data.
• However, there are situations in which interval data will be encountered and must be incorporated.
• In the figure to the right, the open, black circles represent point data where there is not uncertainty in the actual magnitude.
• The black bars and red rectangle represent interval data where there is uncertainty in the actual magnitude.
• The Expected Moments Algorithm (EMA) contained within B17C procedures can natively incorporate both of these data types. Also, other fitting methods like Maximum Likelihood Estimation can as well.
• More on this to come.

Outliers

- High and low outliers
- If encountered, search for additional data
  - Is the high outlier(s) rarer than indicated by the sample?
  - Is the low outlier(s) more common than indicated by the sample?
- Bulletin 17 procedures give them special treatment



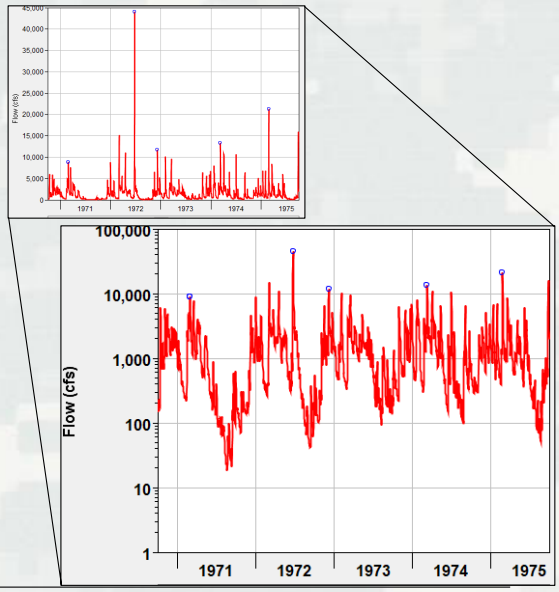
21

BUILDING STRONG®

• Outliers are a common occurrence in most data sets
• Sometimes they're on the high side and sometimes they're low.
• However, their inclusion, without special treatment, can violate the IID assumption, specifically the identically distributed part.
• When you encounter them, search for additional data and ask yourself these questions:
  • Is the high outlier(s) rarer than indicated by the sample? In the figure to the right, this would be analogous to moving the highest point to a smaller, or rarer, exceedance probability.
  • Is the low outlier(s) more common than indicated by the sample? In the figure to the right, this would be analogous to moving the smallest point to a larger, or more common, exceedance probability.
• Bulletin 17B and C procedures include tools that can be used to identify and treat these outliers.

Data Transformations

- Why transform data?
  - Analyze relative changes rather than absolute changes
  - Obtain a symmetrical distribution about zero
  - Remove heteroscedasticity
- Common transforms:
  - Log base 10
  - Log base e (i.e. natural log)
  - Normsinv()
  - Square or square root
  - Inverse (i.e. 1/x or x^{-1})



22

BUILDING STRONG®

• Sometimes data must be transformed to analyze relative changes instead of absolute changes. Other times, engineers and modelers are more interested in obtaining a symmetrical distribution about zero. Also, engineers may be interested in removing the effects of heteroscedasticity, which refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

- Think of flow vs stage or something similar where the variability of flow changes drastically as the stage increases in a typical, incised cross section with a very wide flood plain.
- For a small change in stage, the corresponding change in flow could be small or extremely large depending upon whether the flow is in or out of bank.

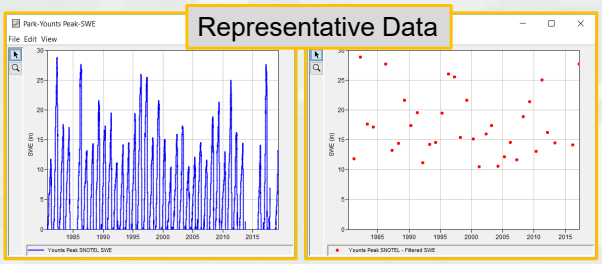
• In terms of common transforms, Log base 10 is probably the most commonly used transform when dealing with flow.

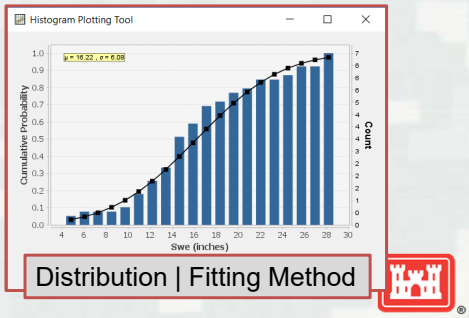
- This transformation allows for the user to analyze relative changes rather than absolute changes and an example is shown to the right.
- Remember that the mean of the log transformed values is NOT the same as log of the mean of the values.

• However, this is by no means the only data transform that is used. I list out a few common transforms at the bottom of the slide.

Parametric Modeling Steps

1. Develop a representative data set
2. Select a probability distribution and fitting method ("model")
3. Compute parameters of the model using the data (i.e. fit the distribution)
4. Verify appropriateness of the model
5. Use model to predict quantiles of interest





23

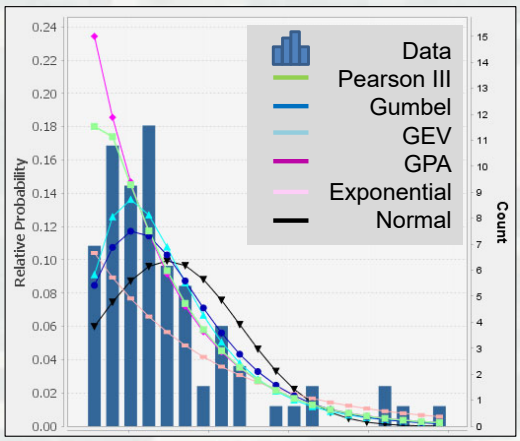
BUILDING STRONG®

•Step 2 is the selection of a distribution and fitting method. When put together, a distribution and fitting method create a model.

• Step 3 focuses on computing parameters of the model.

Examples of Commonly-Used Probability Distributions

- Exponential
  - 1 parameter
- Gumbel
  - 2 parameters
- Normal
  - 2 parameters
- Pearson Type III
  - 3 parameters
- Extreme Value Distributions
  - Hosking and Wallis (1996) parameterizations
  - Generalized Extreme Value (GEV), Generalized Pareto (GPA), and Generalized Logistic (GLO)
  - 3 parameters

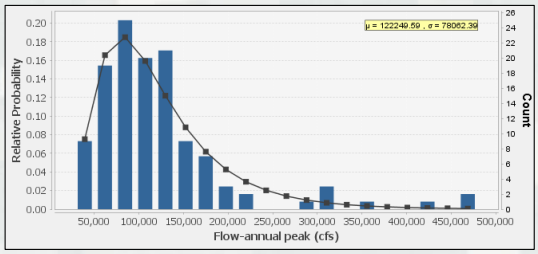


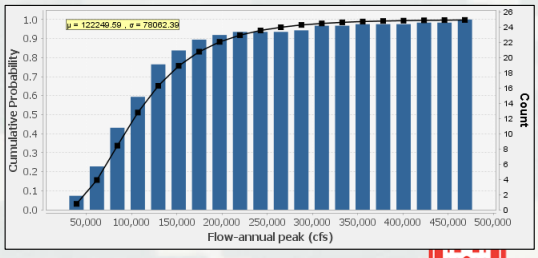
BUILDING STRONG®

• Here are some common continuous probability distributions that are utilized within water resources applications.
• Aside from the familiar Exponential, Gumbel, Normal, and Pearson Type III, I'd like to highlight the extreme value distributions noted at the bottom of the list.
  • These distributions are part of the family of three-parameter distributions for which parameterization schemes were devised by Hosking and Wallis.
  • Each has location ξ, scale α, and shape κ parameters.
• First is the Generalized Extreme Values, or GEV, distribution.
  • This distribution subsumes the three extreme value distributions: Gumbel (EV type I), Frechet (EV type II), and Weibull (EV type III).
  • Which distribution it represents is dependent upon the shape parameter. For instance, GEV = Gumbel when shape = 0.
  • The GEV distribution is commonly used for precipitation-frequency studies within the U.S. and many other applications like flow-frequency outside of the U.S.
  • The GEV distribution was derived by taking the maximum of repeated independent samples from a homogeneous population, which is what is commonly used to create annual maximum series.
• Next up is the Generalized Pareto, or GPA, distribution.
  • An underlying assumption of the GPA distribution is that subsamples exceeding a sufficiently high threshold from repeated samples of a homogeneous population will converge to the GPA distribution.
  • In other words, if repeated samples are taken from a population, and only the values in those samples that are greater than a selected value are retained, then those retained values will follow the GPA distribution.
  • This type of data set is called a partial duration series, which is something that we discussed in the previous lecture.
• Finally, I'd like to mention the Generalized Logistic, or GLO, distribution.
  • Like the GEV distribution, the GLO distribution is commonly used in precipitation-frequency studies.

What is a continuous probability distribution?

- Continuous probability distributions are the most commonly used distributions within water resources applications
  - Data is comprised of a continuous random variable
- Area under PDF must equal 1
- CDF spans probability between 0 -> 1
  - Cannot decrease either





25

BUILDING STRONG®

• Discrete probability distributions are used to describe the expected results from experiments where only a certain number of outcomes can be realized, like flipping a coin or rolling a die.
• These distributions aren't used as much within water resources applications, but they are still used from time to time to model things like will it rain tomorrow, if it is going to rain, what type of storm will it be, etc?
• Continuous probability distributions are used much more commonly within water resources applications. This arises from the fact that we tend to model phenomena that is comprised of continuous random variables, like streamflow, precipitation, or stage.
• Similar to a unit hydrograph, the area under the probability density function, or PDF, must sum to 1. An example of a PDF is shown in the upper right image.
• Also, the cumulative distribution function, or CDF, must span 0 -> 1 and cannot decrease. An example of a CDF is shown in the lower right image.

Probability Distribution Fitting Methods

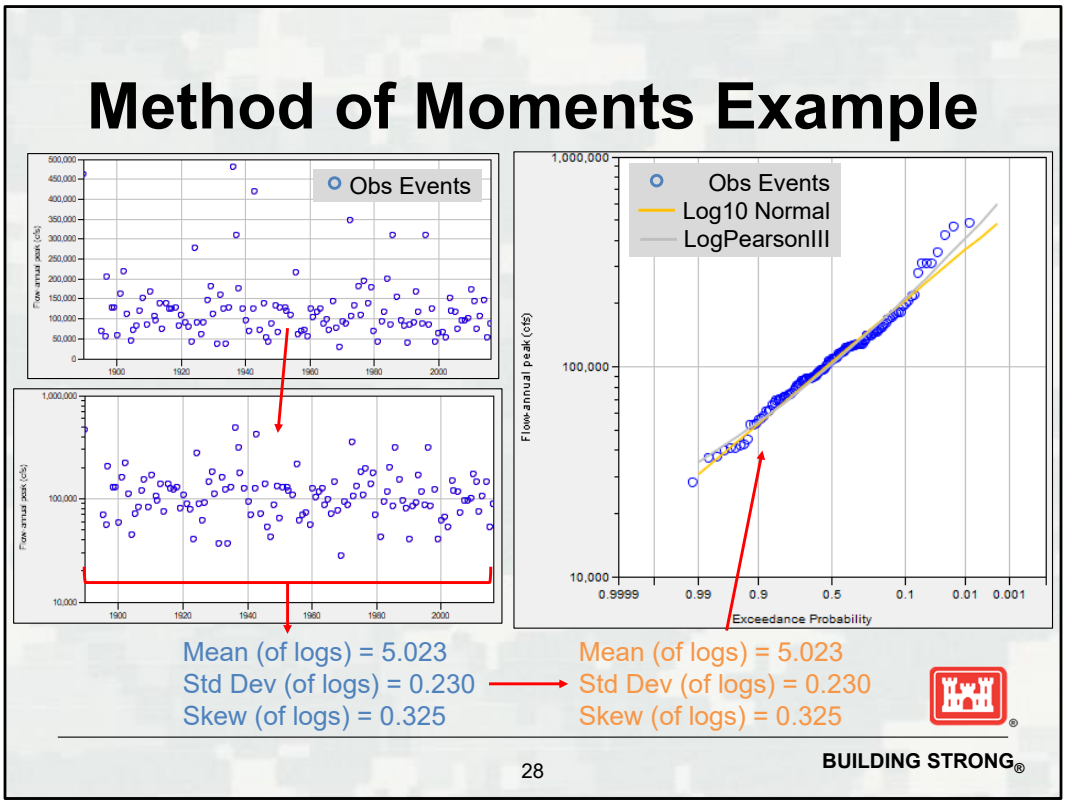
- **Fitting** a **probability distribution** to an **unknown and unknowable parent population** through the use of a **representative sample** (child population)
- Commonly used fitting methods:
  - Method of moments
    - Linear moments
    - Expected Moments Algorithm (EMA)
  - Maximum likelihood

• Now, let's discuss fitting methods.
• As mentioned before, the same analytical distribution can be fit to the same data set using different fitting methods to obtain different parameterizations and, as such, different quantile estimates.
• When we fit a distribution, we are using the child population because we don't know the true parent population. In fact, within all real-world applications, we can't know the parent population. We can only know the parent population in contrived examples.
• Some commonly used fitting methods are method of moments and Maximum Likelihood Estimation.
• Some special extensions of the method of moments are Linear Moments and the Expected Moments Algorithm, or EMA.
• Greg will go into more detail regarding Linear Moments and their applications while Beth will present several videos detailing EMA and it's uses.

Method of Moments

- **Simple and widely used**
- Assume that the parameters of the sample (child population) are the same as the parent population
- Equal weights are given to the transformations of the observations
- Used within Bulletin 17 methods
  - ► EMA also uses method of moments (but iterates)

• The method of moments fitting method is widely used within water resources applications because it's fairly simple and robust.
• The underlying assumption of this method is that the parameters of the child population are the same as the parent population.
• Remember last lecture when I stressed how important it was to develop an IID and representative data set? This assumption is the reason why a fair amount of time should be spent processing your data set.
• Within this method, equal weights are given to the transformations of the observations.
• As I said before, the method of moments is used within both Bulletin 17B and Bulletin 17C procedures. However, Bulletin 17C makes use of a generalization called EMA that allows for the use of some unique types of data like intervals.



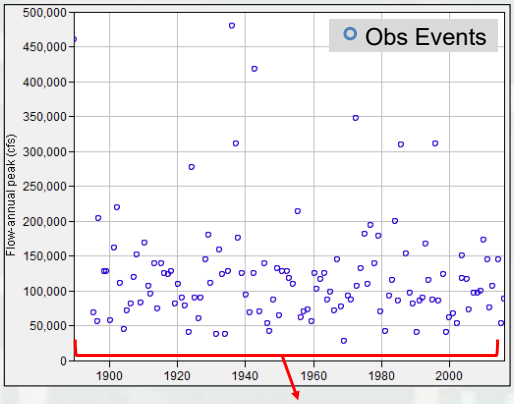
- Now, let's step through an example.
- In the upper left image, I have an annual maximum series of streamflow.
- In the lower left image, this annual maximum series has been transformed to log10 values.
- Next, the (log) parameters of the sample are computed and shown in the blue text.
- Then, those parameters are used to fit the both the Log Normal and Log Pearson Type III distributions in the orange text.
    - The Log10Normal does not use the skew parameter; only the LPIII distribution does (in this case).
    - Notice that the parameters of the sample match the parameters of these two distributions? That's the consequence of assuming that the parameters of the child population are the same as the parameters of the parent population.
- Finally, the fitted distributions are plotted along with the plotting positions of the data in the right hand image.

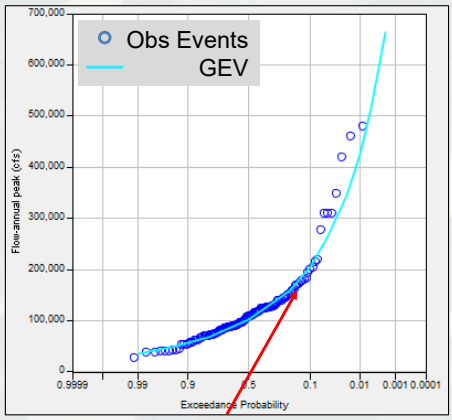
Linear Moments

- **A little more complicated but still widely used**
- Still assume that the parameters of the sample (child population) are the same as the parent population
- Used heavily within regional and precipitation frequency analyses
- Unequal weights given to order statistics of observations based upon their ranks
- Less weight is given to the tails of the distribution
- Hosking and Wallis (1990 and many, many others)

• Linear, or L-moments are widely used in precipitation frequency studies.
• This becomes especially useful when regionalizing parameters.
• This fitting method is an extension of the method of moments in that it still assumes the parameters of the child population are the same as the parent population
• However, unequal weights are given to order statistics of observations based upon their ranks
• In particular, less weight is given to the tails of the distribution
• Hosking and Wallis revolutionized the use of this fitting method along with the three aforementioned distributions.
• More on this fitting method will be presented within other videos.

Linear Moments Example





Linear Moments
L1 = 122249.6
L2 = 36781.3
L3 = 12273.5
L4 = 10397.4

Linear Moment Ratios
L-Mean = L1 = 122249.6
L-CV = L2 / L1 = 0.301
L-Skew = L3 / L2 = 0.334
L-Kurtosis = L4 / L2 = 0.283

30

BUILDING STRONG®

- Here's an example of L-moments.
- An annual maximum series of flow is used to compute the L-moments denoted as L1, L2, L3, and L4.
- Then, those L-moments are used to compute the L-moment ratios: L-Mean, L-CV, L-Skew, L-Kurtosis.
- Using these linear moments, the three parameter GEV distribution can be parameterized (using a scheme put forth by Hosking and Wallis) and is shown on the right.

Maximum Likelihood Estimation

- **Much more complicated and not as widely used in water resources as previous methods**
- Though, this method is widely used in other statistical applications
- Can incorporate point, censored, and interval data
- Can integrate arbitrarily complex distributions
- Requires log (actually natural log) likelihood functions (since ln() is a continuous strictly increasing function)
  - Calculate the first derivative and set equal to 0 to maximize

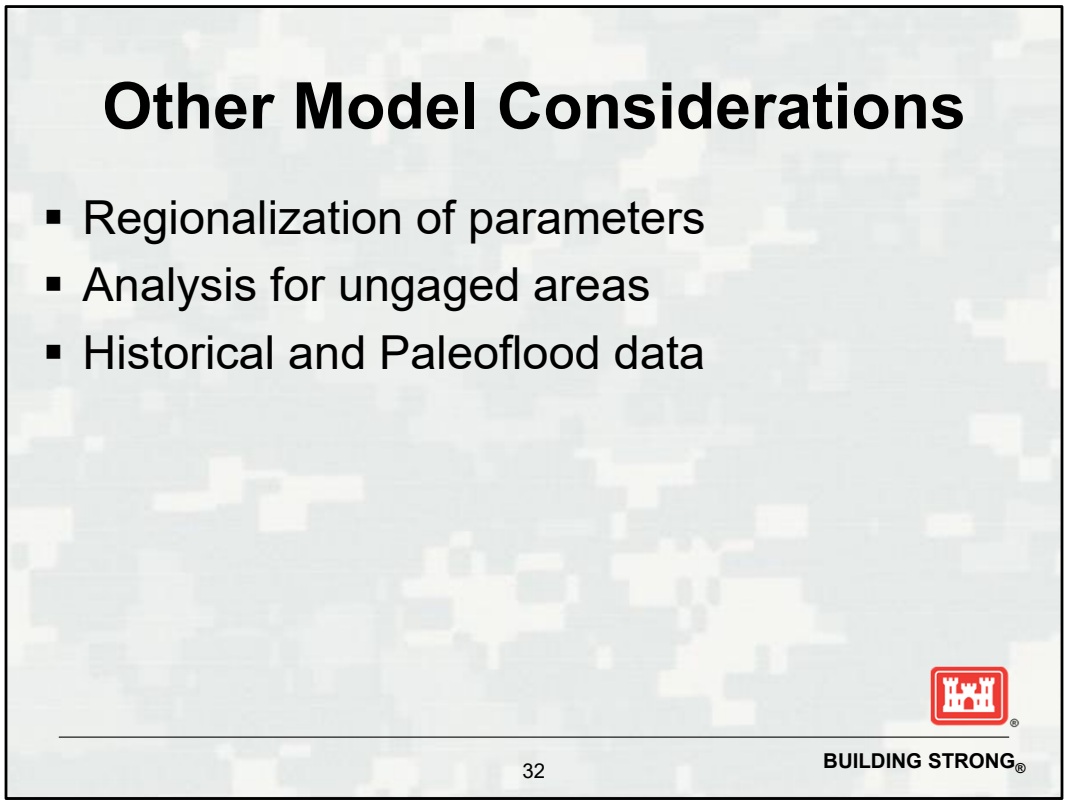
$$\frac{1}{\alpha}\sum_{i=1}^{S}\left[\frac{1-\kappa-(y_i)^{1/\kappa}}{y_i}\right]=0$$

$$-\frac{S}{\alpha}+\frac{1}{\alpha}\sum_{i=1}^{S}\left[\frac{1-\kappa-(y_i)^{1/\kappa}}{y_i}\left(\frac{x_i-\xi}{\alpha}\right)\right]=0$$

$$-\frac{1}{\kappa^2}\sum_{i=1}^{S}\left\{\ln(y_i)[1-\kappa-(y_i)^{1/\kappa}]+\frac{1-\kappa-(y_i)^{1/\kappa}}{y_i}\kappa\left(\frac{x_i-\xi}{\alpha}\right)\right\}=0$$

GEV log-likelihood functions

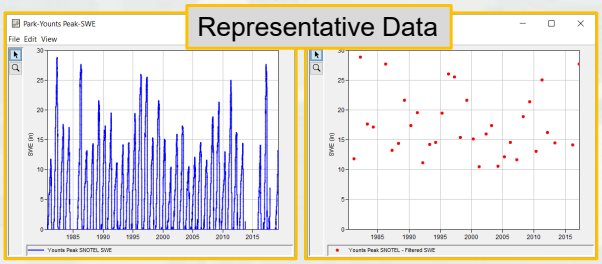
• Maximum Likelihood Estimation, or MLE, is a much more complicated fitting method than the previous two fitting methods.
• As such, this method has not been as widely used in water resources applications.
• But, it is commonly used in other arenas, like the financial and actuarial applications.
• This fitting method, like EMA, is able to incorporate point, censored, and interval data. However, this fitting method is able to integrate arbitrarily complex distributions.

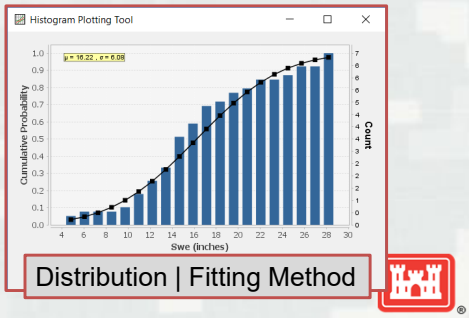


• Other topics related to model fitting include regionalizing data, analyzing ungaged areas, and incorporating historical and/or paleoflood data. Historical and paleoflood data are oftentimes missing direct observations and are best represented by interval or censored observations.
• These topics will be covered in additional videos.
• We also have another class dedicated to these topics. If you're interested, please check out our Flood Frequency Analysis for more information as well.

Parametric Modeling Steps

1. Develop a representative data set
2. Select a probability distribution and fitting method (“model”)
3. Compute parameters of the model using the data (i.e. fit the distribution)
4. Verify appropriateness of the model
5. Use model to predict quantiles of interest



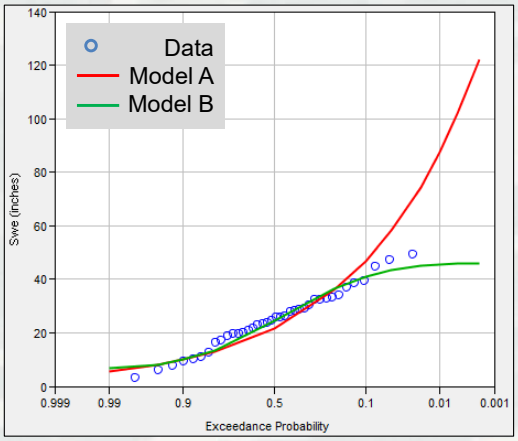


BUILDING STRONG®

- Step 4 emphasizes verifying that the model is appropriate for use with the data in question.
- Step 5 uses the parameterized model to make predictions.

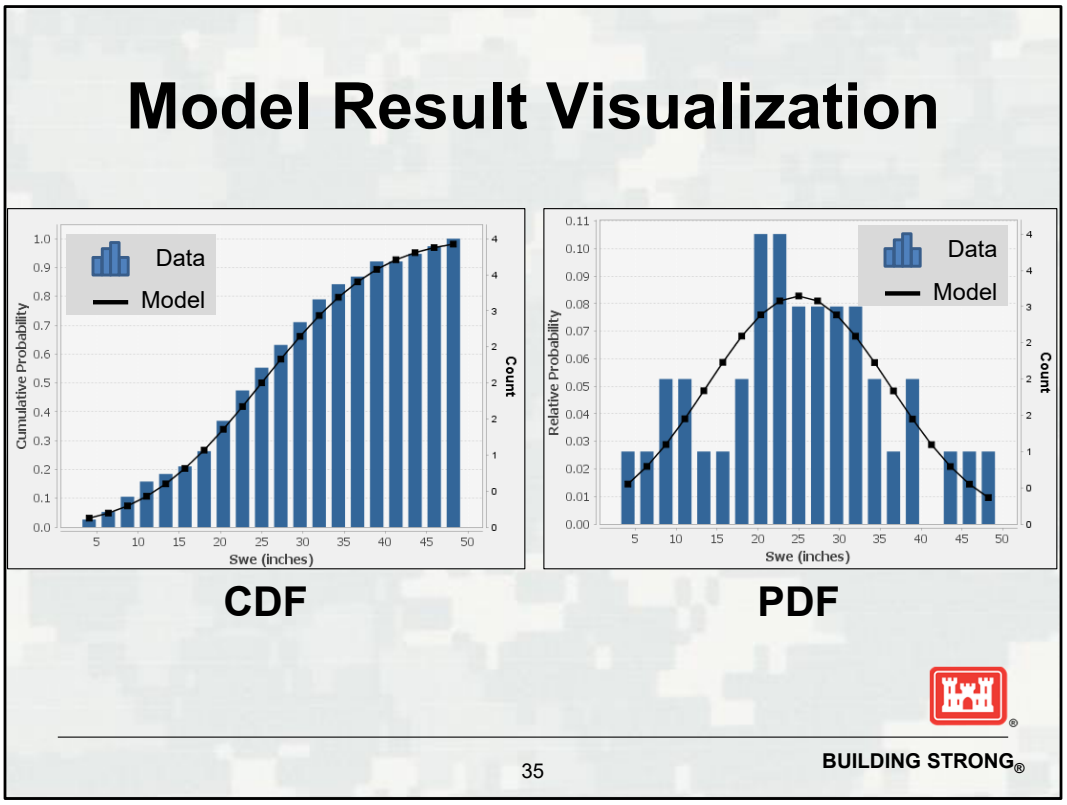
Model Result Visualization

- There are multiple ways to ascertain goodness of fit
  - **Visualize the model results against the data (multiple ways to do this)**
  - Goodness of fit tests
    - Kolmogorov-Smirnov
    - Chi-Square
    - Anderson-Darling
  - Split sample tests
- All should be used!



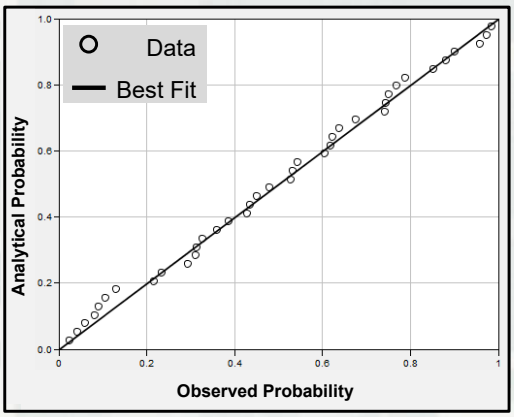
BUILDING STRONG®

• There are multiple ways in which the model and data can be visualized together in order to form a qualitatively comparison.
• This includes things like visualizing the model against the plotting positions of the data, as shown to the right.
  • This is the most commonly-used visualization within water resources applications.
  • This type of plot can provide a quick means to weed out bad models.
  • For instance, both of the models visualized here are probably not good choices to represent this data because the tail behavior of both don't adequately represent the data, which is usually where we're most interested in obtaining estimates.
• However, there are other ways to visualize the data, which we'll describe on the next few slides.
• Also, there are numerous quantitative tests, called goodness of fit tests, that can be used to ascertain the appropriateness of the model.
• Instead of relying on just a single visualization or quantitative test, all should be used to justify the selection of a model.

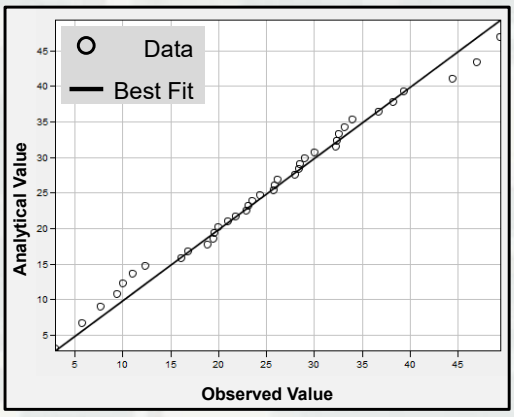


• Plotting the observed data against the model on a Cumulative Distribution Function, or CDF, plot and Probability Density Function, or PDF, plot can provide valuable insight to the goodness of fit provided by the model.
• In the plots shown here, the same model is fit to the same data within a CDF and PDF plot. Also, the same number of bins and bin sizes are used in both plots.
• The model appears to fit the data much better within the CDF plot than the PDF plot, but how much of that visual cue is based upon the chosen bin size in the PDF plot?
• Perhaps less bins would allow better visualization of the model fit within the PDF plot.
• You can use different numbers of bins and/or bin sizes when visualizing the data in this way.

Model Result Visualization



Probability-Probability Plot



Quantile-Quantile Plot

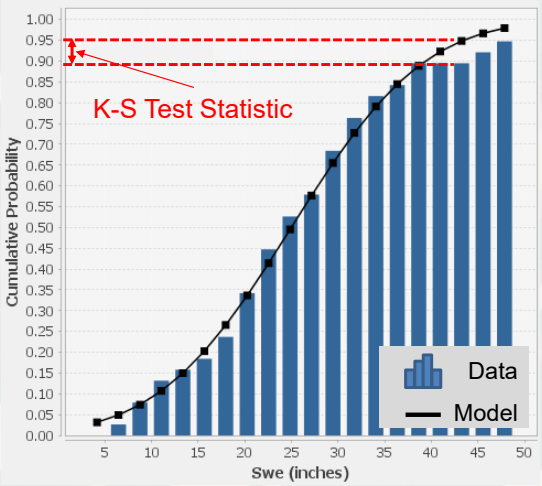
36

BUILDING STRONG®

- Probability-Probability, or PP, plots compare the probability of the data (using the chosen plotting position formula) against the inferred probability of the model.
- Quantile-Quantile, or QQ, plots compare the values of the actual data against the inferred value as predicted by the model.
- Both plots are valuable for comparing trends. However, the QQ plot is better at visualizing tail behavior while the PP plot is better at visualizing and comparing the “center” of the data.
- In both plots, solid black lines of perfect agreement are included for comparison.

Kolmogorov-Smirnov Test

- **Kolmogorov-Smirnov (K-S)**
  - **Non-parametric method**
  - Uses the **maximum difference** in **CDF** between the model and the data
  - If the difference is large based on the sample size, the null hypothesis that the data come from the proposed model would be rejected
  - Example:
    - *K-S Test Statistic = 0.05*

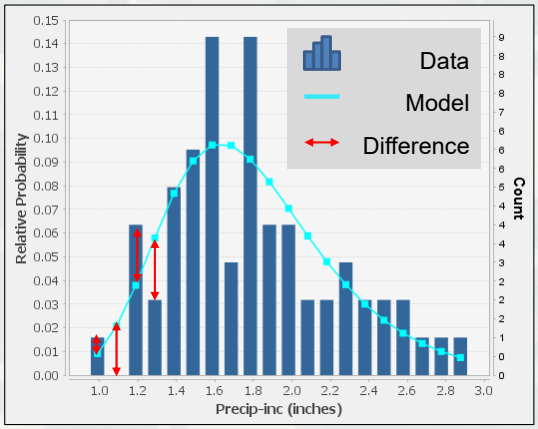


BUILDING STRONG®

• Now, let's talk about quantitative goodness of fit tests.
• First up, is the Kolmogorov-Smirnov, or K-S, test.
• This test is probably the most commonly-used goodness of fit test in water resources because it's simple and non-parametric in that it doesn't rely upon the assumption of an underlying distribution.
• This test measures the largest difference in the CDF between the model and the observed data.
• In the example on the right, the largest difference between the CDF of the model and observed data is approximately 0.05.
• Therefore, the K-S test statistic is 0.05.

Chi-Square Test

- **(Pearson) Chi-Square**
  - **Parametric method**
  - Is there a significant difference between the expected frequencies and the observed frequencies?
  - Create a number of "bins" for the data and compare the observed proportion of the data in each bin compared to the expected proportion of the data according to the model
  - If the proportions are significantly different, then the null hypothesis that the data arose from the proposed model would be rejected

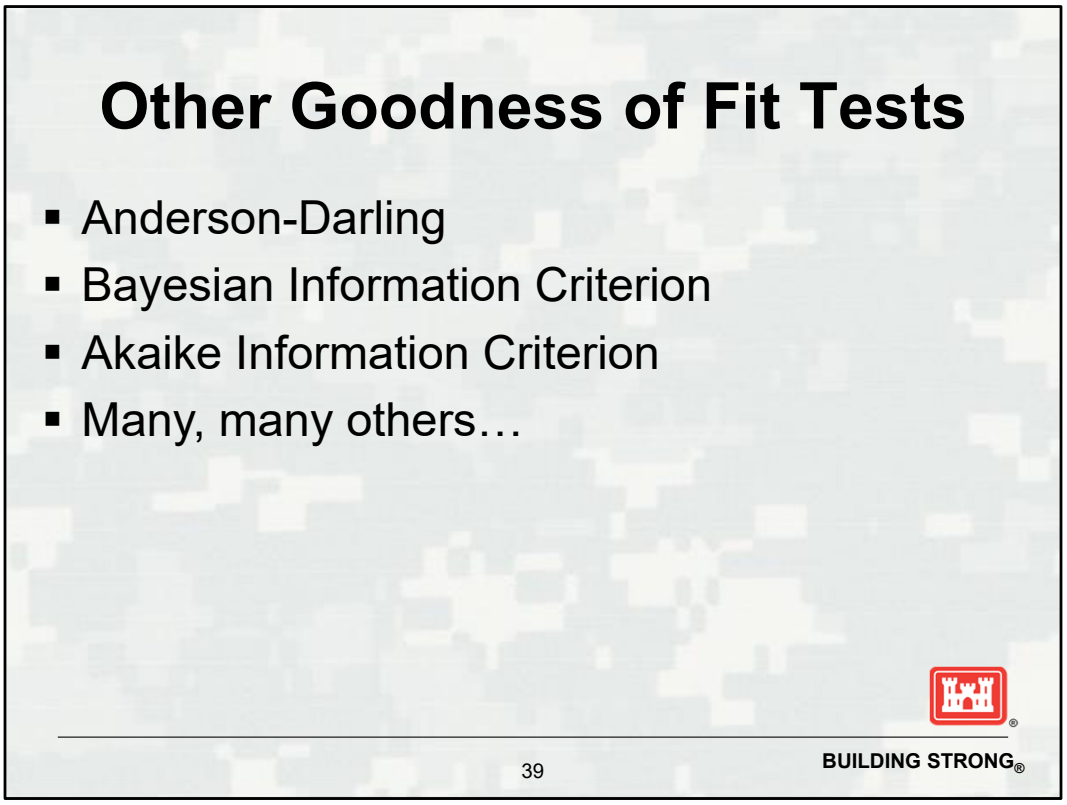


$$Test\ Statistic = \sum |Data - Model|$$

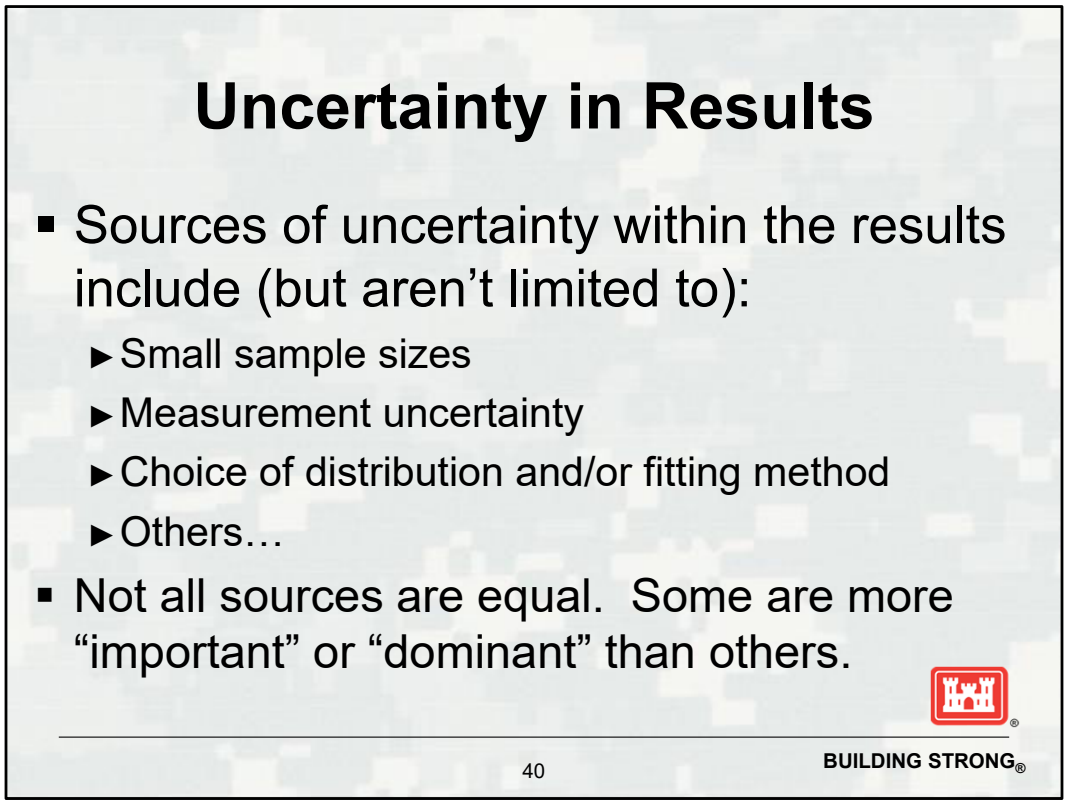
38

BUILDING STRONG®

- The Pearson Chi-Square test, which is often shortened to just Chi-Square, compares the entire model and distribution of data, not just the maximum in the CDF.
- It essentially asks, "Is there a significant difference between the model-predicted frequencies and the frequencies within the observed data?"
- This test first ranks the data then places the values within bins. In HEC-SSP, exactly five values are placed in each and every bin. This can result in the use of bins that are not uniform in size. Values can not be shared within bins; every possible value is in exactly one and only one bin. Finally, the first bin edge has a CDF value = 0 and last bin has a CDF value = 1.
- Once the bins are created, the test statistic is computed using the sum of squared residuals between the model and data across the entire probability range. In the figure above, only a few bins are compared with red arrows, but all bins are compared.
- The name "Chi-Square" test arises from the fact that when the number of sampled elements in each bin is equivalent to the expected value, it is predicted by the chi-square distribution with n-1 degrees of freedom.



- There are many, many goodness of fit tests that have been developed for different types of data, distributions, and fitting methods. I'll briefly describe three additional tests that are available within HEC-SSP. This list is by no means exhaustive as everybody and their brother has a goodness of fit test.
- The Anderson-Darling test assesses whether a sample comes from a selected distribution.
    - This test works by first assuming that the data does arise from the chosen distribution and then tests the data for uniformity with a simple distance test.
- Bayesian Information Criterion, or BIC, and Akaike Information Criterion, or AIC, are similar tests
    - Both the BIC and AIC tests make use of likelihood functions, which were briefly introduced in the last lecture for use with the maximum likelihood fitting method.
    - This means that these tests aren't appropriate for use with fitting methods like method of moments, which limits their applicability.
    - However, these tests are really cool in that they penalize complex models in an attempt to weed out overfitting.
    - In this case, model complexity refers to the number of parameters. So, if you use a 3-parameter model like Log Pearson Type III, it'll be penalized more than a 2-parameter model, like Log Normal.

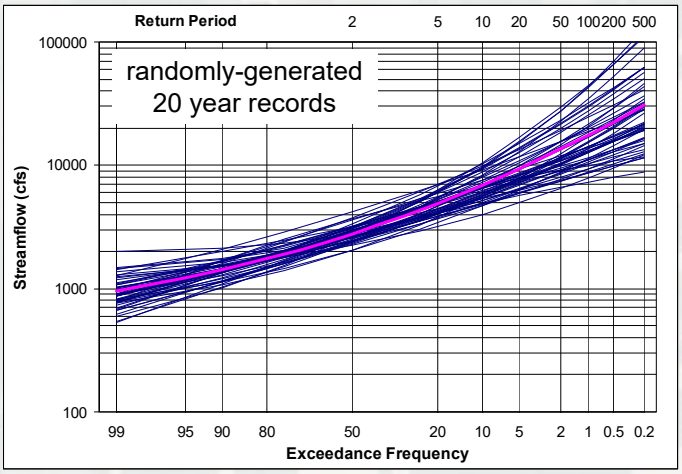


- There are multiple ways in which uncertainty enters into Parametric Modeling and 99.9% of the time, uncertainty should be shown and discussed along with the results.
- Uncertainty in your results can arise from things like small sample sizes, uncertainty in measurements in inferences of measurements, and modeling choices. This is by no means an exhaustive list, but only meant to provoke some discussion. We'll discuss a few of these sources in more detail in the next few slides.
- I gotta mention that not all sources of uncertainty are equal. Some sources provide greater uncertainty in different "portions" of the results than others. For instance, source "X" may provide greater uncertainty in the more frequent range of the resultant probability distribution than source "Y". But, source "Y" provides

much greater uncertainty in the extremely rare range of probabilities.

Confidence Intervals

- **Model results** (i.e. median curve) account for **natural variability**
- **Confidence limits** account for **knowledge uncertainty**
- **Uncertainty** in the model results can be **visualized** through the use of **confidence intervals**
  - Uncertainty is actually a distribution. More on that later…
  - B17 procedures provides these intervals
  - But, they can also be determined using other approaches

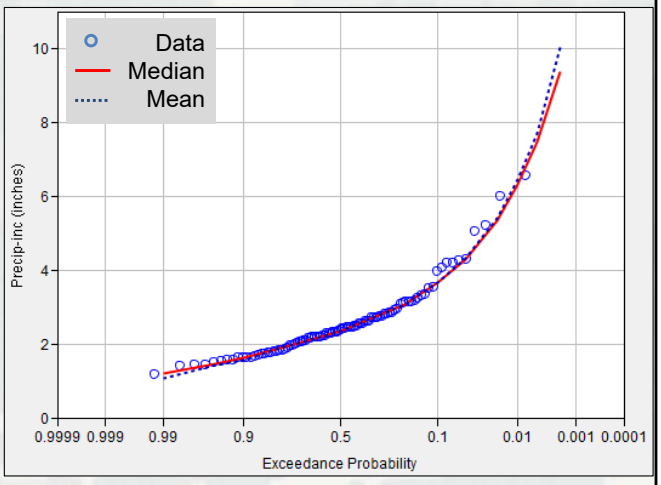


BUILDING STRONG®

• Uncertainty is most commonly visualized using confidence intervals, which are meant to encompass many sources of uncertainty (but usually not all). Confidence intervals typically represent knowledge uncertainty, which can be reduced with additional study (conversely, epistemic uncertainty or natural variability cannot be reduced with additional study).
• Confidence intervals simply show two values with each side of a "confidence interval" implying a "confidence limit". However, in actuality, the true uncertainty is best represented using a distribution of uncertainty.
• Confidence intervals can be estimated using monte carlo approaches where multiple samples of the model are made using an effective record length many, many times. This approach is visualized in the figure within this slide. Each blue line is a separate sample.
• In specific cases, confidence intervals can be estimated using closed-form equations. For instance, both Bulletin 17B and Bulletin 17C procedures will produce confidence interval estimates for the Log Pearson Type III distribution. However, closed-form equations aren't available for all distributions.
• Both Beth and Greg will go into much greater detail regarding this topic within later lectures.

Uncertainty Due to Sample Size

- **Expected probability** of exceedance represents the **combined uncertainty** due to both natural and knowledge uncertainty
  - Exp. Prob. curve = mean
  - At ½ AEP, median = mean
  - B17B procedures included an approximation (B17C does not)
  - Adjustment can be determined using other approaches



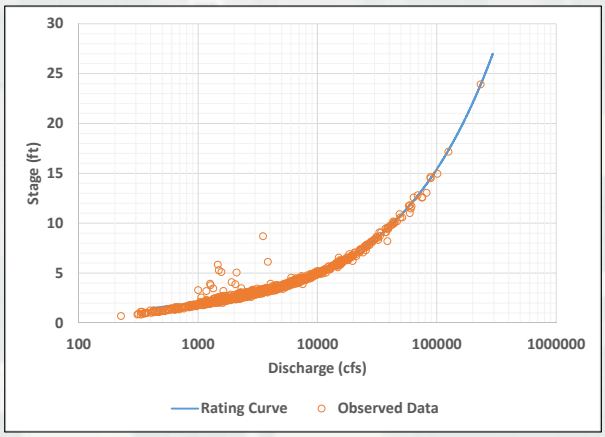
42

BUILDING STRONG®

- Parametric modeling uncertainty can also arise from small sample sizes. This is a very common problem that is encountered in nearly all water resources applications. From a frequentist standpoint (which I found to be much, much easier to understand early in my career), as the sample size gets larger and larger, the uncertainty should get smaller and smaller. Essentially, a model that is fit to 100 years of data should have tighter confidence intervals than a model that is fit to 50 years of data, all other things being equal.
- To correct for small sample sizes, an expected probability adjustment can be made. Remember that the result of a parametric modeling exercise is a parameterized distribution which provides a median quantile estimates. Also, remember that the median represents the point in which 50% of the possible values are above that value and 50% of the possible values are below. The best representation of the true likelihood is the mean.
- At the ½ annual exceedance probability, or AEP, the mean and median values are the same.
- For AEP less than ½, the mean value is greater than the median.
- For AEP greater than ½, the mean value is smaller than the median.
- However, it's not necessarily straightforward to estimate the mean, or an expected probability curve, at the various quantiles of interest. Bulletin 17B included procedures to estimate the expected probability curve but Bulletin 17C does not.
- However, we've recently added tools to HEC-SSP to estimate the mean and produce an expected probability curve when using Bulletin 17C procedures.
- The expected probability curve is most often determined using monte carlo approaches.

Uncertainty Due to Measurement Error

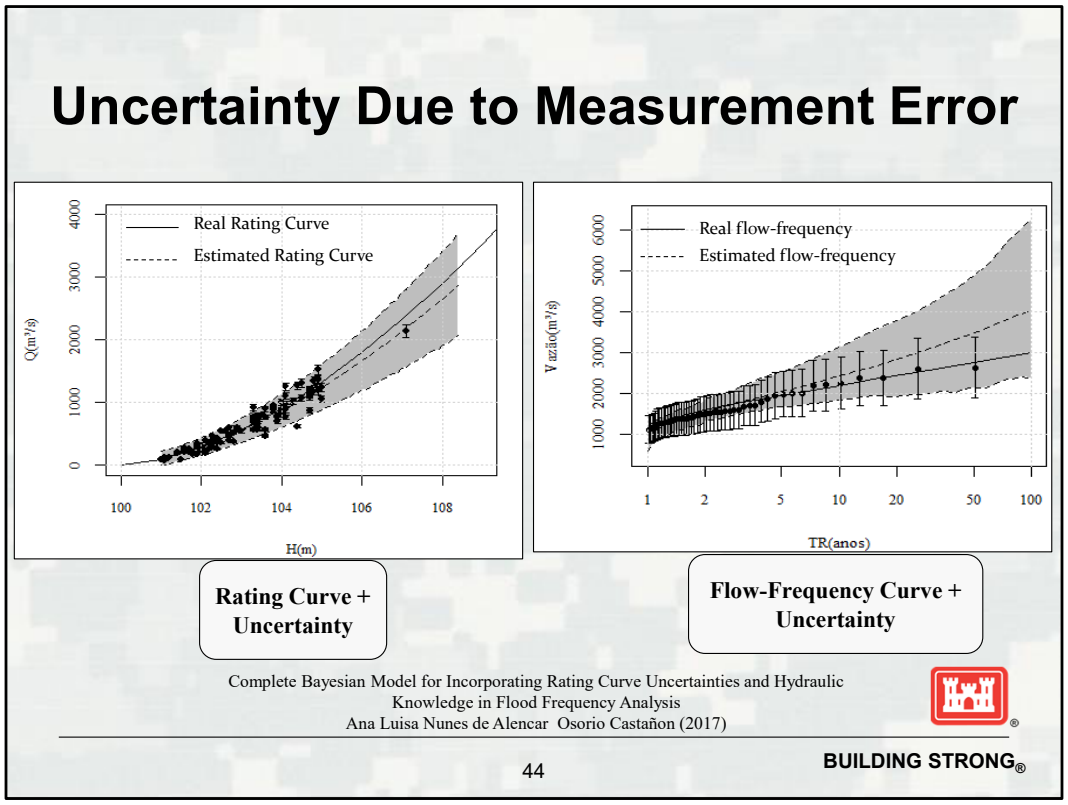
- Uncertainty can also arise due to errors in the measurements of data
  - Prevalent within stage-flow, radar reflectivity-precipitation rate, etc
- Errors can get larger with magnitude
  - Heteroscedasticity



BUILDING STRONG®

- Uncertainty can also arise from errors in the measurements of the data itself.
- For instance, the United States Geological Survey operates a network of thousands of stream gages throughout the United States.
- At these gages, stages are typically measured, not flow.
- Estimates of flow are then based upon the measured stage, an assumed or measured cross sectional shape, and an assumed or measured velocity distribution.
- The stage is commonly measured in regular intervals, like 5-, 15-, 60-minutes, while the cross sectional shape and velocity distribution are measured from time to time and updated when large changes or floods occur.
- The measurements are combined to create a rating curve which transforms the measured stage to a volumetric flow rate.
- This conversion introduces error which carries forward to our analyses since we commonly fit models to flow and not stage (since stage doesn't "behave" as well as flow).
- The errors aren't linear either due to heteroscedasticity, which I mentioned in the "Developing a Representative Data Set" video.
- Remember that common cross section shapes exhibit this

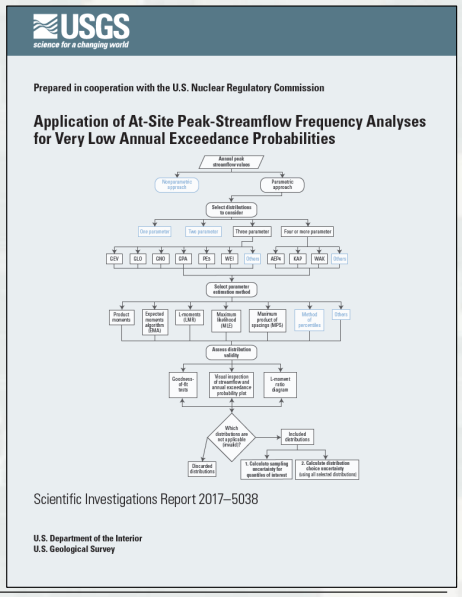
behavior due to different portions of the channel and floodplain being active at different stages/flow rates.



- A super duper awesome master's thesis delving into this topic was performed by a woman from Brazil named Ana Luisa who came to HEC to work with us for a short time a few years ago.
- Within this study, she incorporated rating curve uncertainty into estimated flow-frequency curves.
- As expected, there is a lot of uncertainty and the uncertainty becomes larger as the stage and/or flow rate increases.

Uncertainty Due to Model Choice

- What if all the models fit the sample reasonably well and you must extrapolate to rare exceedance probabilities?
- How much uncertainty is due to the choice of model?
- The problem gets worse with smaller sample sizes…
- This problem is encountered within many types of studies (not just dam safety)


USGS
science for a changing world

Prepared in cooperation with the U.S. Nuclear Regulatory Commission

Application of At-Site Peak-Streamflow Frequency Analyses for Very Low Annual Exceedance Probabilities

Scientific Investigations Report 2017–5038

U.S. Department of the Interior
U.S. Geological Survey

45

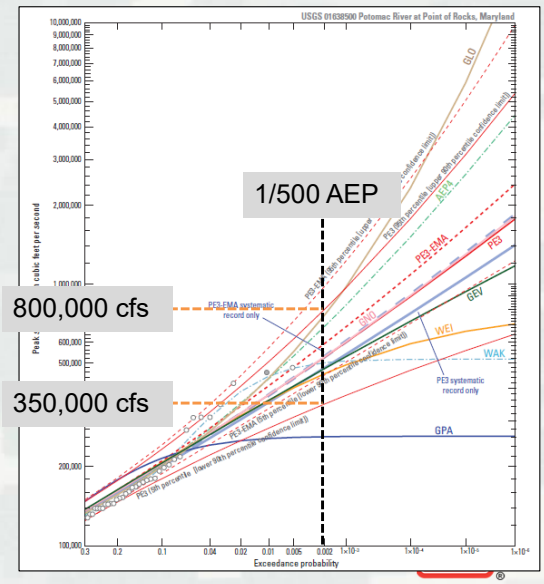
BUILDING STRONG®

- Uncertainty in your results are also due to the modeling choices that were made.
- Here's a commonly encountered problem within water resources applications: "If all the models fit the sample reasonably well, how do you have confidence in your extrapolation?"
  - How accurately are you predicting?:
    - The extents of the floodplain at the 1/500 AEP?
    - The likelihood of SWE exceeding 20 inches given a sample of 15 years?
- This source of uncertainty gets worse and worse with smaller sample sizes.

- The USGS published a really cool paper that quantified this uncertainty for several locations in the U.S. and I show the cover here.

Uncertainty Due to Model Choice

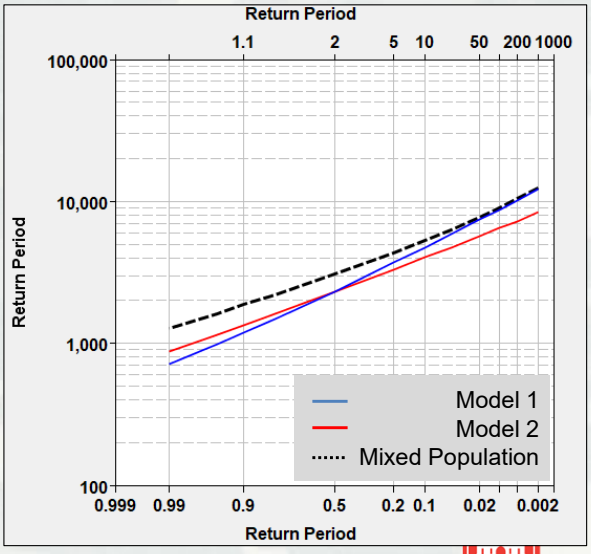
- Depending upon the model choice, huge amounts of uncertainty can result!
- What if all the models reasonably fit the sample?
- Must fall back upon the underlying assumptions of the model…



• Here's a screen capture from that paper. In this image, multiple models have been fit to the same data. These models are visualized as the various solid and dashed lines.
• At the 1/500 AEP, the different models (excluding the Generalized Pareto model) predict a range of 450,000 cfs in the instantaneous peak discharge.
• As the AEP decreases or becomes rarer, these differences increase.
• The moral of the story is that, depending upon the model choice, a huge amount of uncertainty can exist within your quantile predictions, especially at rare exceedance probabilities.
• If several models fit the sample reasonably well, you must fall back upon the underlying assumptions of the model when making a determination of which one is best.
• For instance, this data set is comprised of annual maximum flows. Therefore, the Generalized Pareto model isn't an appropriate distribution choice because that's only meant for use with partial duration series.

Combining Multiple Models

- **Mixed Population**
- Curve Combination
  - Bulletin 17C recommendations
- Bayesian hierarchical modeling

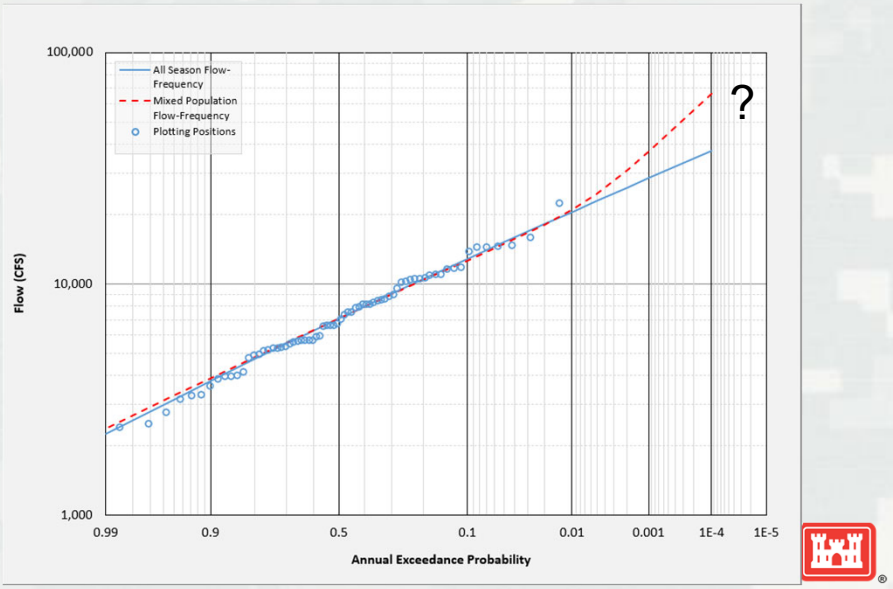


47

BUILDING STRONG®

• There are ways to combine multiple models including mixed population analyses, curve combination, and Bayesian hierarchical modeling. Mixed population analysis uses the probability of union theory while curve combination analyses typically use user-defined weights to combine the multiple inputs into a single resultant probability distribution.
• Bayesian hierarchical modeling is a much more complicated topic and would take far too much time completely explain than is allowed within this lecture. In short, this type of analysis combines multiple models using Bayes theorem. Our colleagues from the Risk Management Center have developed a piece of software that can perform this type of analysis, which is especially useful for dam safety applications. If you're interested in giving it a try, reach out to us and we'll point you in the right direction.

Combining Multiple Models



48

BUILDING STRONG®

- Mixed populations are prevalent all over the country and as a profession, we've kind of been lax on treating them appropriately.
- They really come into play when you start to extrapolate beyond observed data, which is where we're commonly interested when regulating floodplains or managing infrastructure.
- As an example, the solid blue line in this figure is a flood-frequency curve that was realized by fitting a single Log Pearson Type III distribution to an annual maximum series that contains more than one type of flood
    - **<u>This data set is not IID</u>**
- The dashed red line is a flood-frequency curve that correctly accounts for the possibility of more than one type of flood mechanism occurring in any given year
    - This was computed by "combining" separate flow-frequency distributions using probability of union
    - Data sets and resultant distributions adhere to IID assumption
- The differences between the two distributions get larger and larger as AEP decreases.
- Extrapolation is key since we don't have 100s of years of observed data but we still need to estimate flow-frequency for AEPs less than 1/100 for many applications

Summary

- Described parametric modeling, its advantages, disadvantages, and steps.
- Defined data requirements and how to develop a data set for analysis.
- Detailed commonly used probability distributions.
- Outlined fitting methods and parameter estimation.
- Explained multiple ways of validating goodness of fit.
- Briefly described uncertainty.

- I introduced parametric modeling along with some of the theory, advantages, and disadvantages when using this method.
- We talked at length about data that goes into a parametric modeling analysis.
- It's super important to investigate your sample to ensure it's comprised of independent and identically distributed values that are representative of the parent population.
- We talked about the selection of both an analytical distribution and a fitting method. When put together, these two create a model from which we can draw conclusions.
- We discussed commonly utilized distributions within water resources as well as several examples of fitting methods.
- we talked about the visualization of model results against observed data.
- Several qualitative visualization tools were presented in addition to several quantitative goodness of fit tests.
- Remember that when you ascertain whether your chosen model is appropriate for use, your judgement shouldn't be based upon a single visualization or quantitative test.
- Instead, you should use all of the tools that are available.
- Remember that uncertainty arises from multiple sources like small sample sizes, measurement error, and model choices.
- Not all sources of uncertainty are equal.
- Also, we briefly talked about combining different models together to form a single model and what tools are available to perform those combinations.