MET3220C & MET6480
Computational Meteorology

Exploratory Data Analysis
Empirical Distributions

Descriptive Statistics
Robustness of Statistics

Exploratory Data Analysis

- Exploratory data analysis employs statistics to summarize characteristics of the data.
  - Converts data to information.
  - Graphical representations are often used to examine the data, and to infer additional qualities of the data set.
- We will discuss many graphical statistical/graphical techniques for examining data.
- One key question is 'how robust are the statistics?'
- If the summary statistics change greatly depending on the subset of the data that is being examined (assuming the subset is sufficiently large), then we should place little value in the statistics!
  - We will discuss which techniques are more reliable than others.
  - Sometimes the consequences of the data not meeting our assumptions (e.g., a bell curve) are quite serious.
- Key Point: ALWAYS LOOK AT THE DATA!!!!

3M: Mean, Mode, and Median

- The 3 M's each give a measure of typical values.
  - The mean is the average. The notation for the mean of x is $\bar{x}$.
  - Mode is the most frequently occurring number.
  - If we order a data set, then the median is the value in the middle of the list.
- Consider the grades on a hypothetical homework assignment:
  - 21 values
  - 4, 5, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10
  - The mean is 8.1 (A–) $\left(\sum_{i=1}^{N} grade_i \right) / N$
  - The mode is 10 (A)
  - The median is 8 (A–) The 11th value, $grade_{(N+1)/2}$

Quartiles and Percentiles

- Quartiles and percentiles are used to describe the spread or distribution of a data set.
- Consider an ordered set of data $\{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$
- The median is defined as
  - $$q_{0.5} = \begin{cases} x_{(n+1)/2}, & \text{if n is odd} \\ (x_{n/2} + x_{n/2+1})/2, & \text{if n is even} \end{cases}$$
    - Half way through the series
- Quartiles are 25% through the series for the lower quartile $q_{0.25}$, and 75% through the series for the upper quartile $q_{0.75}$.
- Percentiles are the value that are greater than a percentage of the data set.
  - Example: the 50^{th} percentile is the median.
  - In data set with 100 values, the 99^{th} percentile is the greatest value.

FORTRAN: Arrays

- We will discuss a new type of variable designed to hold a series of data.

```
REAL fake_obs(125)    ! A variable that can hold 125 real values.
DO index = 1, 125
    fake_obs( index ) = ( REAL( index) + 0.5 ) ** 2
    ! Converts index to a real number, then adds 0.5, the squares to total
ENDDO
PRINT*, fake_obs(10:20)   ! Prints the 10th to 20th elements of the array
```

- Consider a sorted array $\{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$.
- If the change with index number is uniform, and n is large, then the mean can be approximated as $x_{(n+1)/2}$
  - This value is the median, even if the change with index number is non-uniform
- And the 30th percentile can be approximated as x(0.3 * n + 1)

More Robust Estimates of Central Location

- The mean is sensitive to outliers
- Consider the data set {11, 12, 13, 14, 15, 16, 17, 18, 19}
  - The mean and median are 15.
- However, if the final value (19) is replaced with 91, then the mean becomes 23.
  - The robustness can be examined by examining the differences between the mean of *the whole data set* and the *mean with a small fraction of the data removed*.
  - This test should be done for many samples.
  - This example is a form of *cross validation*.
- A trimean is a more robust measure of the central location.
  - Trimean = ($q_{0.25} + 2q_{0.5} + q_{0.75}$) / 4
- A mean could also be determined from a trimmed portion of the data set:

$$\bar{x}_{\alpha} = \frac{1}{n - 2\alpha n} \sum_{i=\alpha n+1}^{n-\alpha n} x_i$$

Example Cross Validation Code

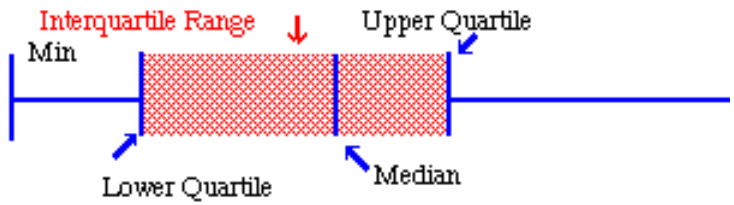
Useful for testing the robustness of a mean of a small number of independent obs.

```
REAL sum
REAL, dimension(365) :: daily_rain   !array of 365 daily temperatures
REAL test_means(365)   !An alternative version of similar declaration
INTEGER index, n, skip
N = 365   ! Assume that the values are read by the program
sum = 0.0
DO index = 1, n
    sum = sum + daily_rain(index)
ENDDO
DO index = 1, n
    test_means(index) = sum – daily_rain(index)
    test_means(index) = test_means(index) / REAL(n – 1)
ENDDO
sum = sum / n
```

Determines the mean for n subsets of the data. These can then be compared to the mean for the full data set.

Spread: Interquartile Range (IQR)

- One measure of spread is the interquartile range.
  - Spread in an indication of departure from the mean.
- IQR = $q_{0.75} - q_{0.25}$
- The IQR is a very robust measure of the spread of values near the mean, but does not give any information on outliers.



Graphic from http://ellerbruch.nmu.edu/classes/CS560w96/students/WELLERBR/boxplots/box.learn.html

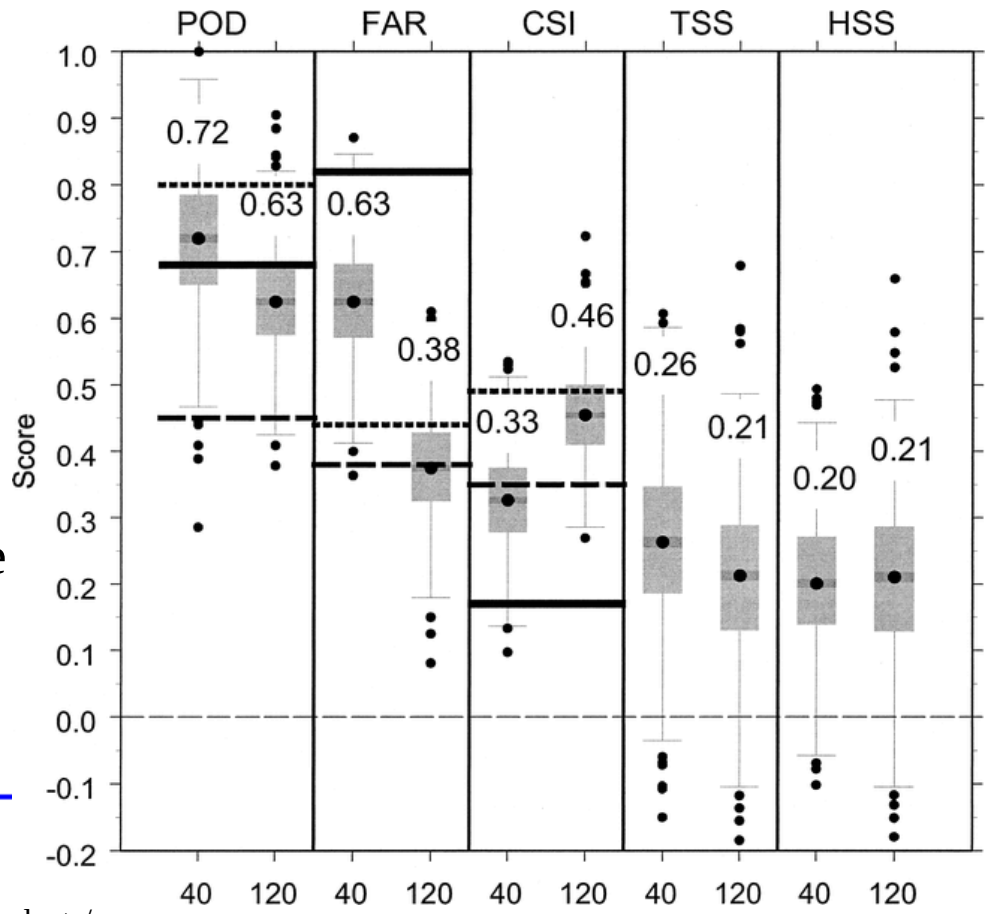


Image from http://ams.allenpress.com/perlserv/?request=display-figures&name=i1520-0434-18-5-953-f04

Standard Deviation

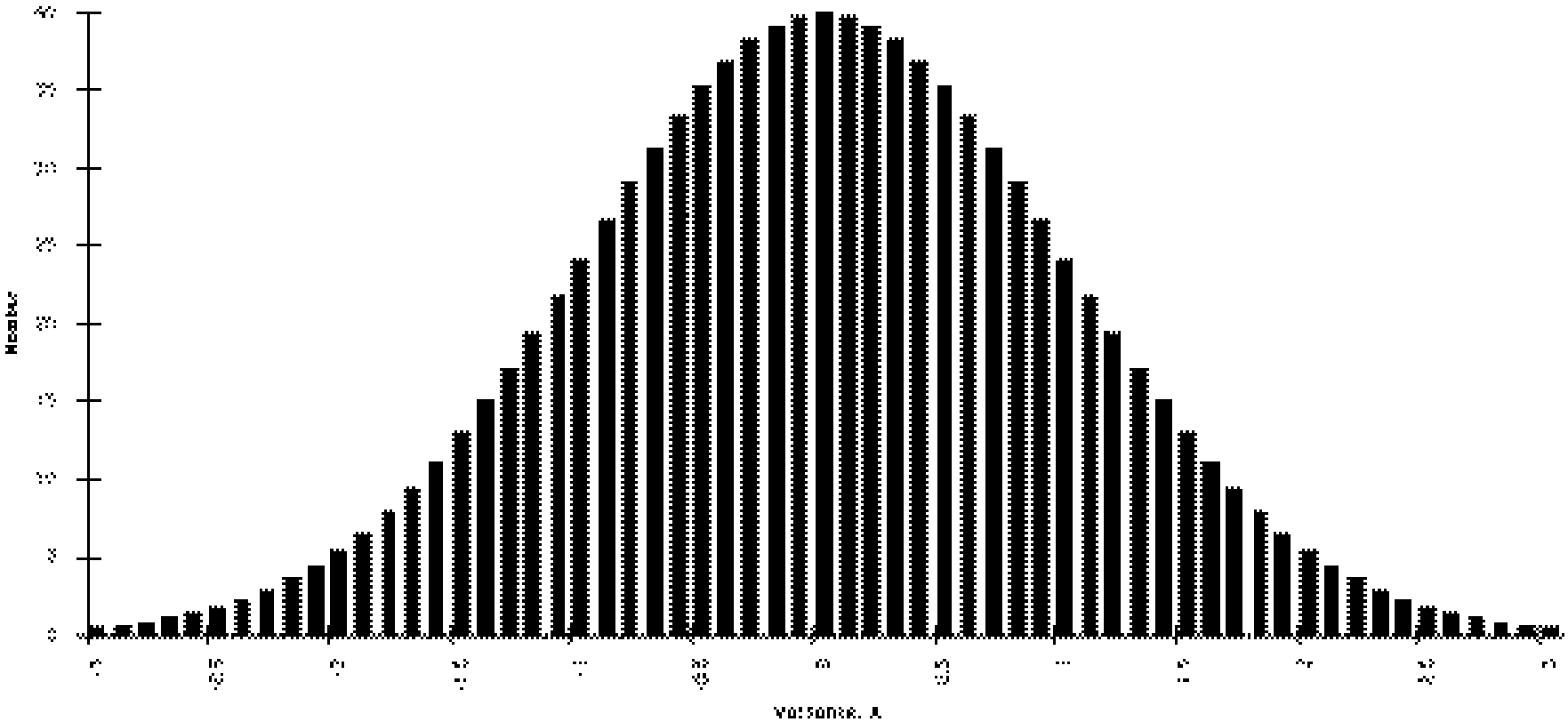
- The standard deviation is the most common measure of spread.
  - Unlike the IQR, is does consider outliers.
- The standard deviation is defined as

$$s = \sqrt{\frac{1}{n-1}\sum_{i}^{n}\left(x_i - \bar{x}\right)^2}$$

  - Where s is an estimate of the standard deviation
    - Estimate because it is assumed (?) to be based on an incomplete sample of the population.
    - The true standard deviation is usually notated as σ

- The standard deviation is highly sensitive to outliers. Why?
- Because of the square of the difference from the mean.
- If the data has a Gaussian distribution, then
  - 68% of the data are within 1 standard deviation from the mean
  - 99% of the data are within 3 standard deviations from the mean

Gaussian Distribution

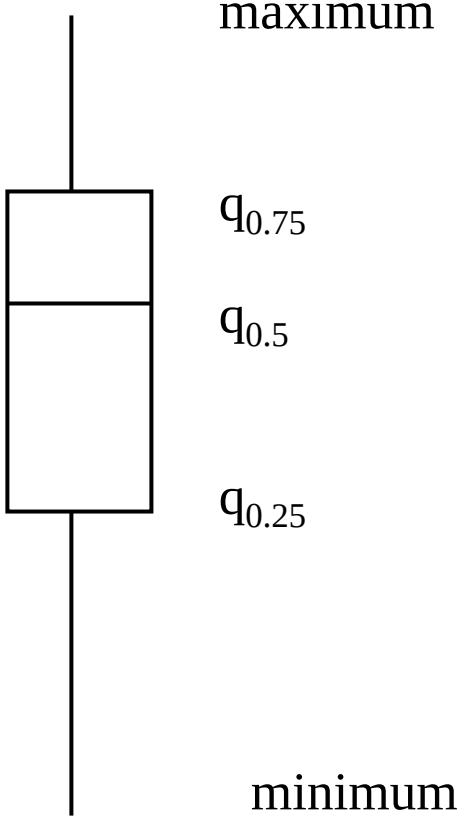


Graphic from www.cimms.ou.edu/~doswell/ Normals/normal.html

Median Absolute Deviation (MAD)

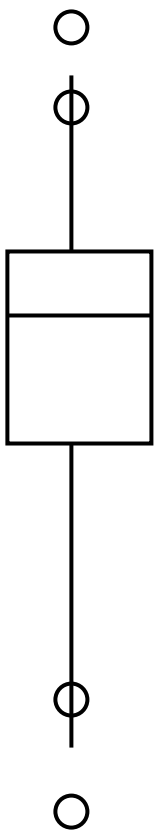
- The MAD does consider outliers, but unlike the standard deviation, the outlines have similar influence to the non-outliers.
- MAD = median($|x_i - q_{0.5}|$)
- Why is the influence of outliers reduced?
- Two reasons:
  - No square of the difference from the central location
  - The median (rather than the mean) is not influenced by outliers.

Box and Whiskers Plots



- The box plot is modified to show the extreme values of the data set.
- Alternatively, the whiskers can indicate the 5^{th} and 95^{th} percentiles
- Alternatively, the whiskers can indicate a multiples of the IQR.

Box and Whiskers Variant



maximum

$q_{0.75}$

$q_{0.5}$

$q_{0.25}$

Minimum within the whiskers

minimum

- The box plot is modified to show the an additional measure of spread, and extreme values of the data set.
- Also shown are the most extreme values within the measure of spread.

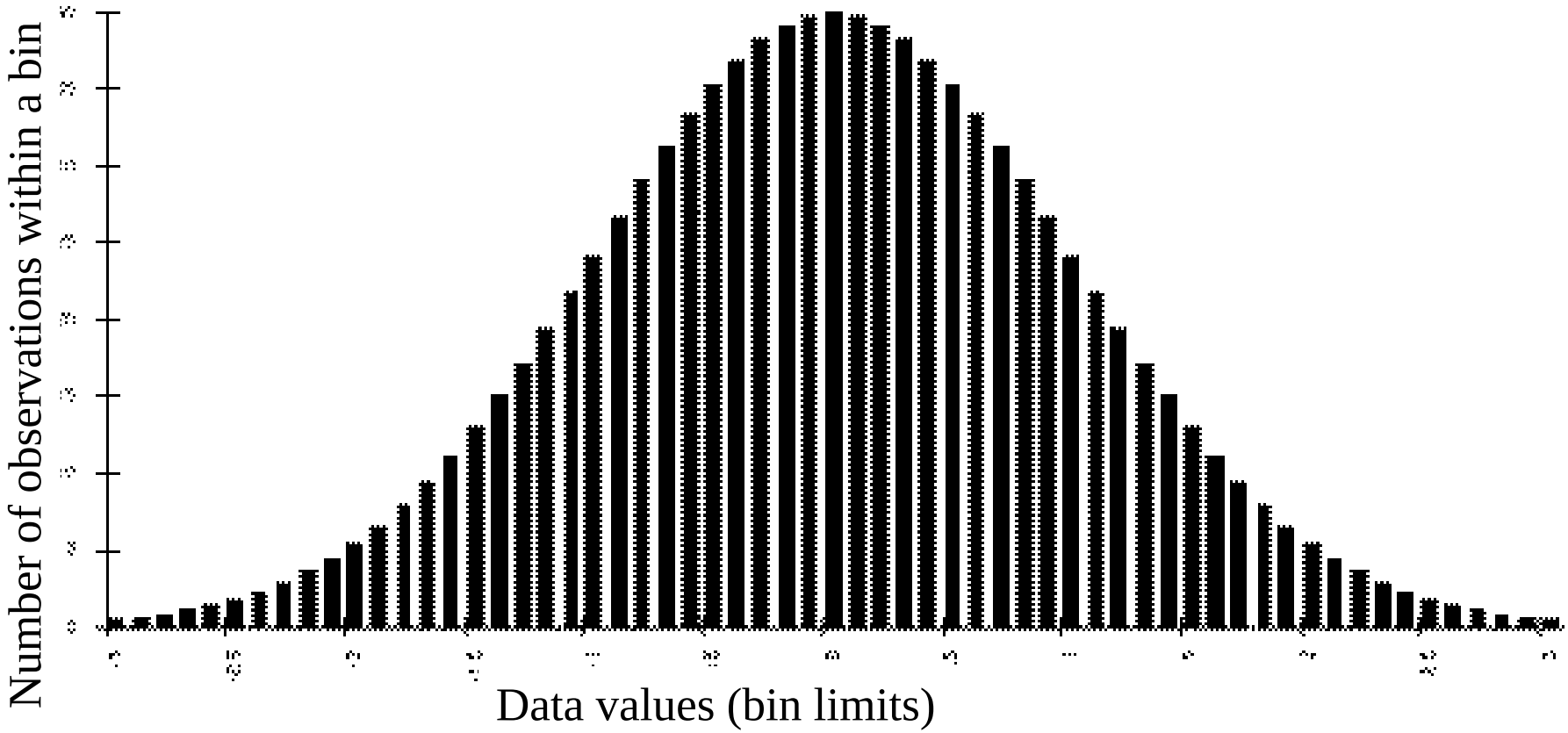
Symmetry – or a lack thereof

- 'Is the data symmetrical?' is a key question for many assumptions
  - Example: are the ENSO impacts for El Niño equal and opposite those for La Niña? If so, forecasting is a lot easier!
- Skewness is one measure of asymmetry

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3}$$

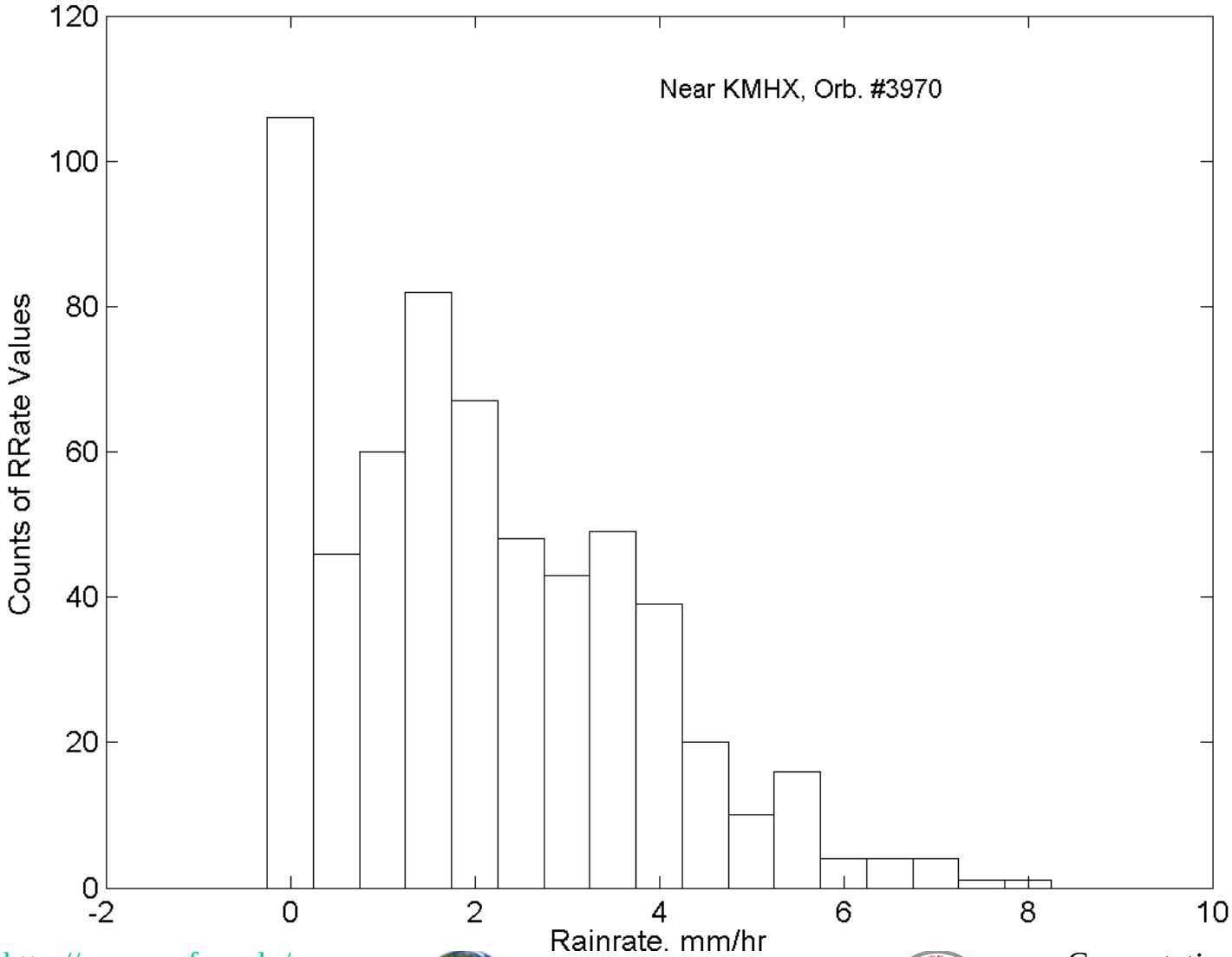
- Skewness is far from a robust statistic
  - There are several alternatives.
  - However, these are not commonly used in statistical analyses

Histograms, or Probability Distribution Functions

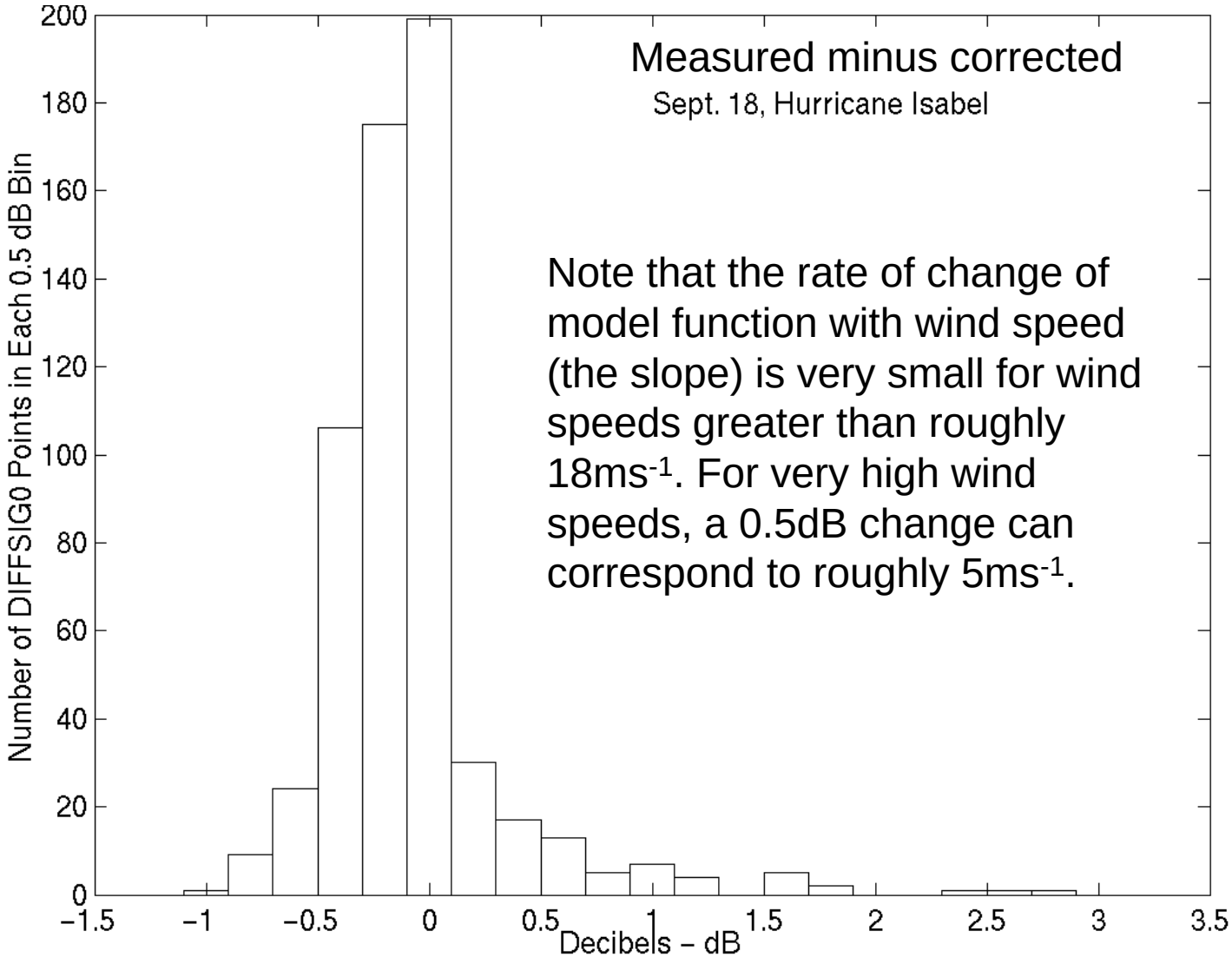


Graphic from www.cimms.ou.edu/~doswell/ Normals/normal.html

Hurr. Isabel Rainrate Value Distribution for Overlapping QSCAT & NEXRAD

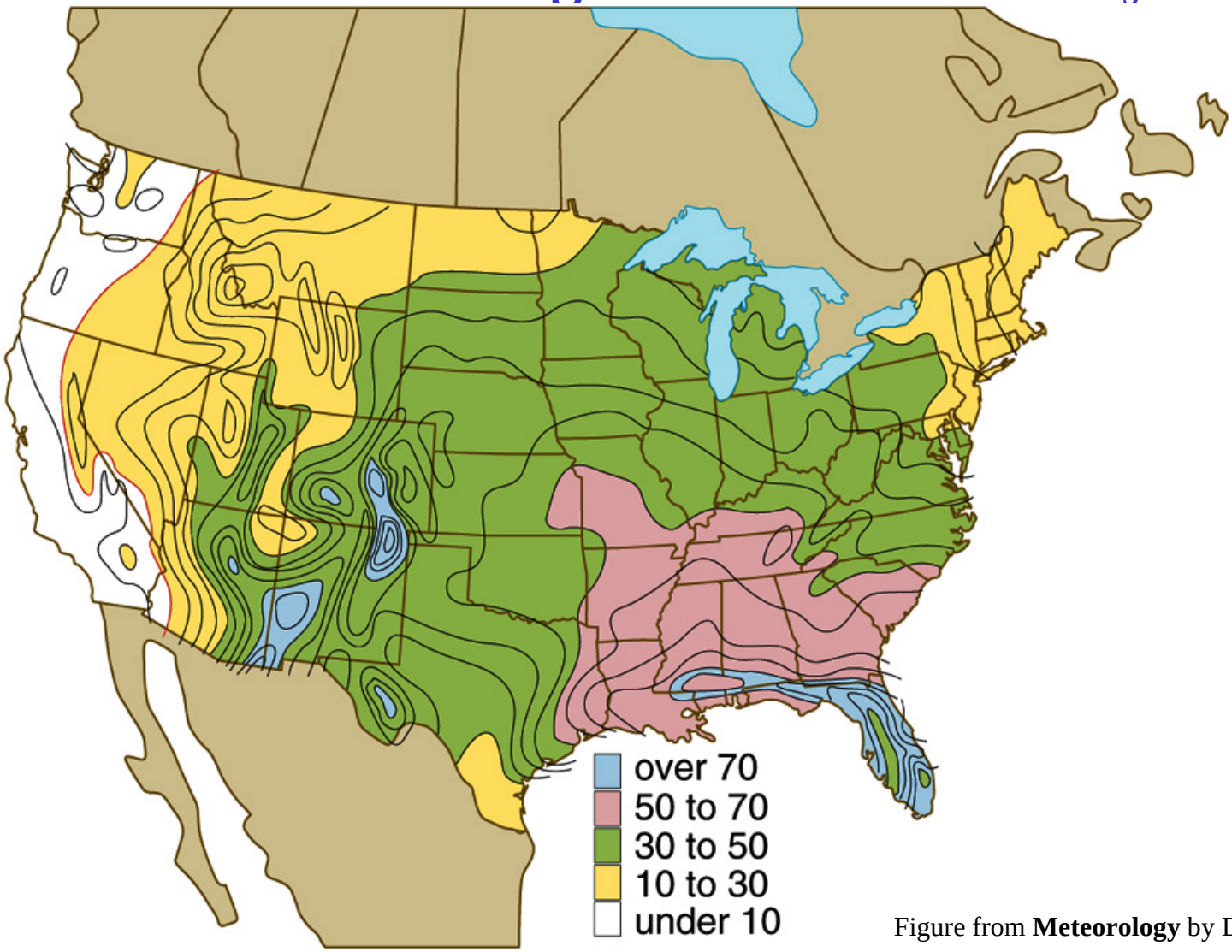


Physically Based Correction to QSCAT Radar Return



Note that the rate of change of model function with wind speed (the slope) is very small for wind speeds greater than roughly 18ms^{-1}. For very high wind speeds, a 0.5dB change can correspond to roughly 5ms^{-1}.

Examine of Histogram Bivariate Estimator
Average Thunderstorm Days



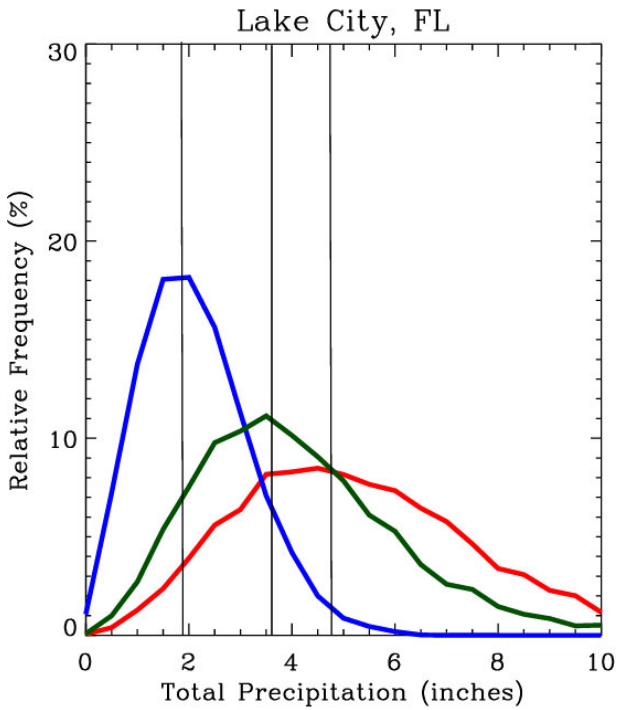
- The average number of days with thunderstorms.
- Contour interval is five per day.
- The two variables used as estimators are latitude and longitude

Figure from **Meteorology** by Danielson, Levin and Abrams

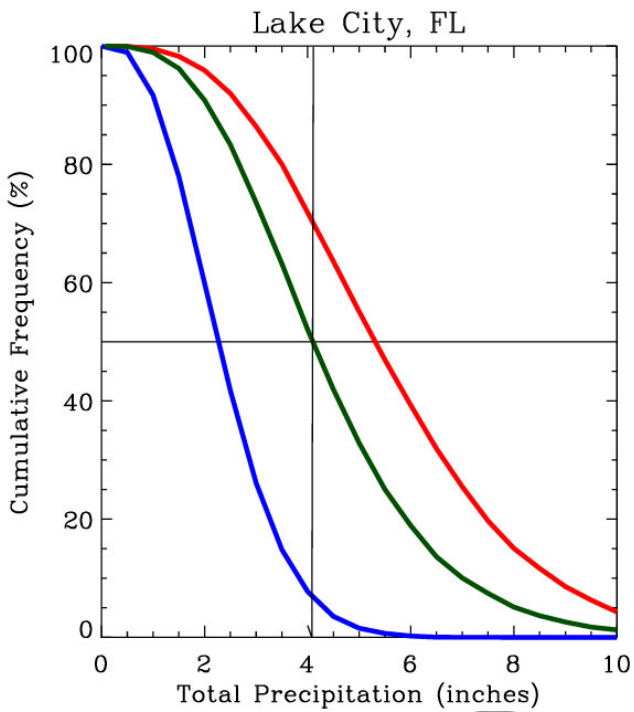
Cumulative Probability Distributions

- Cumulative probability distributions plot
  - X-axis: magnitude of events
  - Y-axis: cumulative probability of all events to the left of the point on the x-axis. P(x ≤ X)
    - Probability of exceedence is 1 – cumulative probability. P (x ≥ X)

Histogram



Probability of Exceedence



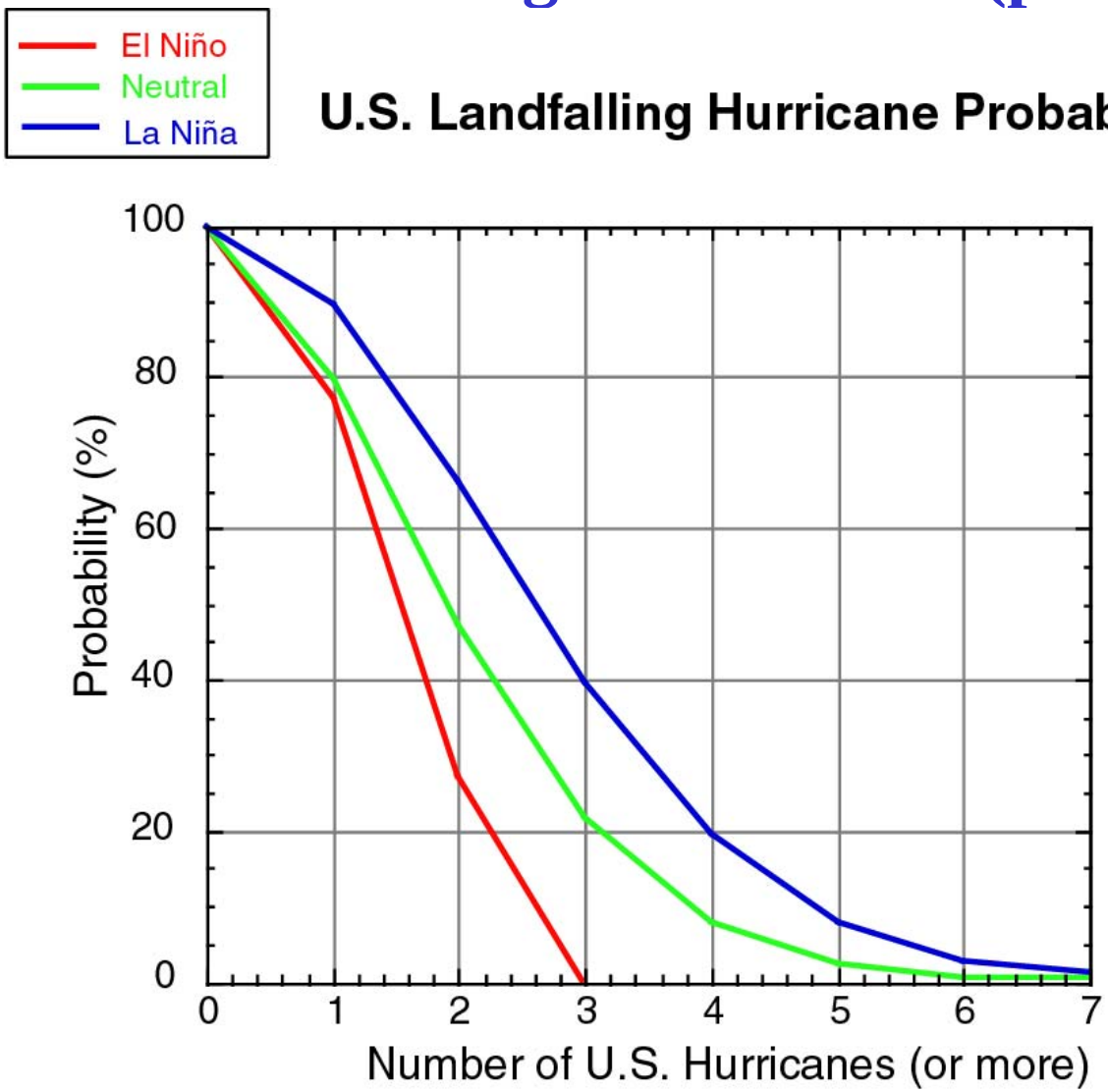
Colors:

El Nino

Neutral

La Nina

Probability of Exceedence Example: Landfalling Hurricances (pre 2005)



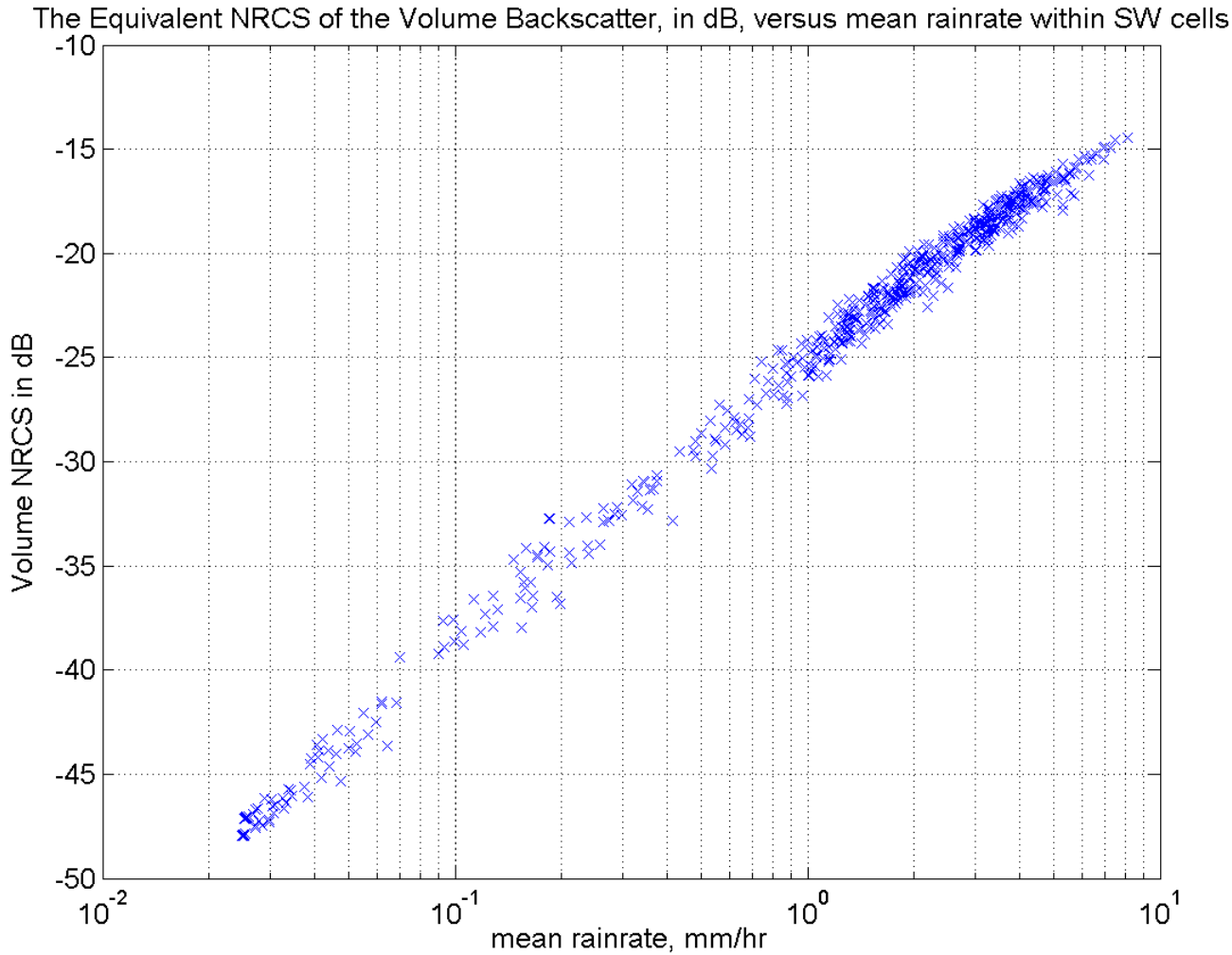
Extreme Value Distributions

- There are several types of extreme value distributions that can be used to describe the likelihood of an event (or an event of lesser magnitude).
- These methods are usually used on ordered (also called ranked) data, where *i* represents the rank (from lesser values to greater values).
- Most of these follow form given in the text book.
- They are used to estimate the likelihood of events of certain magnitude, often for engineering or insurance purposes.
- Alternatively, they can be used to estimate that average time (with large margins of error) between events of given magnitudes.
  - Example: the average time between floods of a certain level.

Standardized Anomalies

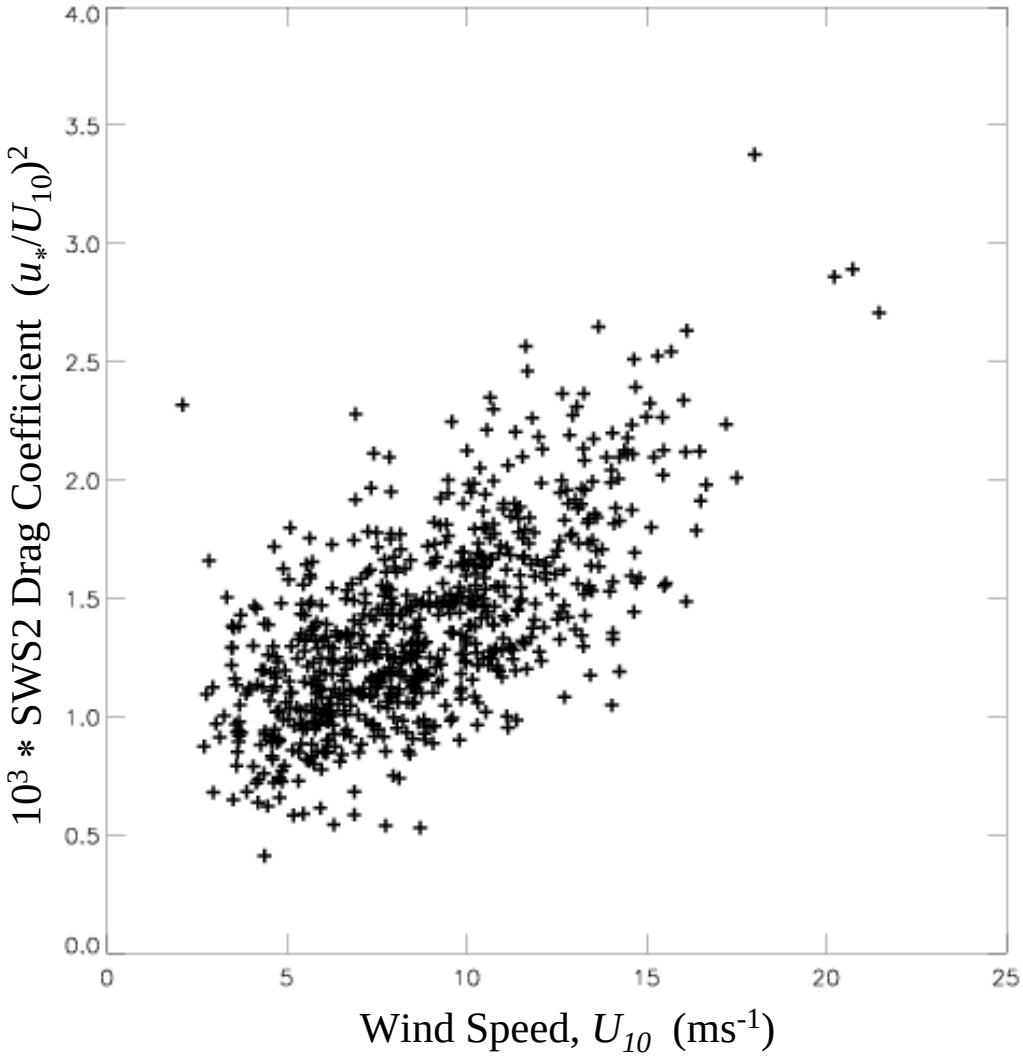
- We are often interested in departures from a mean value.
- If x_i is a value in a series, and <x> is the mean value, the the departure associated with x_i is usually written as x'
  - $x_i = \langle x \rangle + x_i'$
- The standardized anomaly (z) is defined as
  - $z_i = (x_i - \langle x \rangle) / s_x$
  - Where s_x is the sample standard deviation

Exploration of Paired Data
Scatterplots



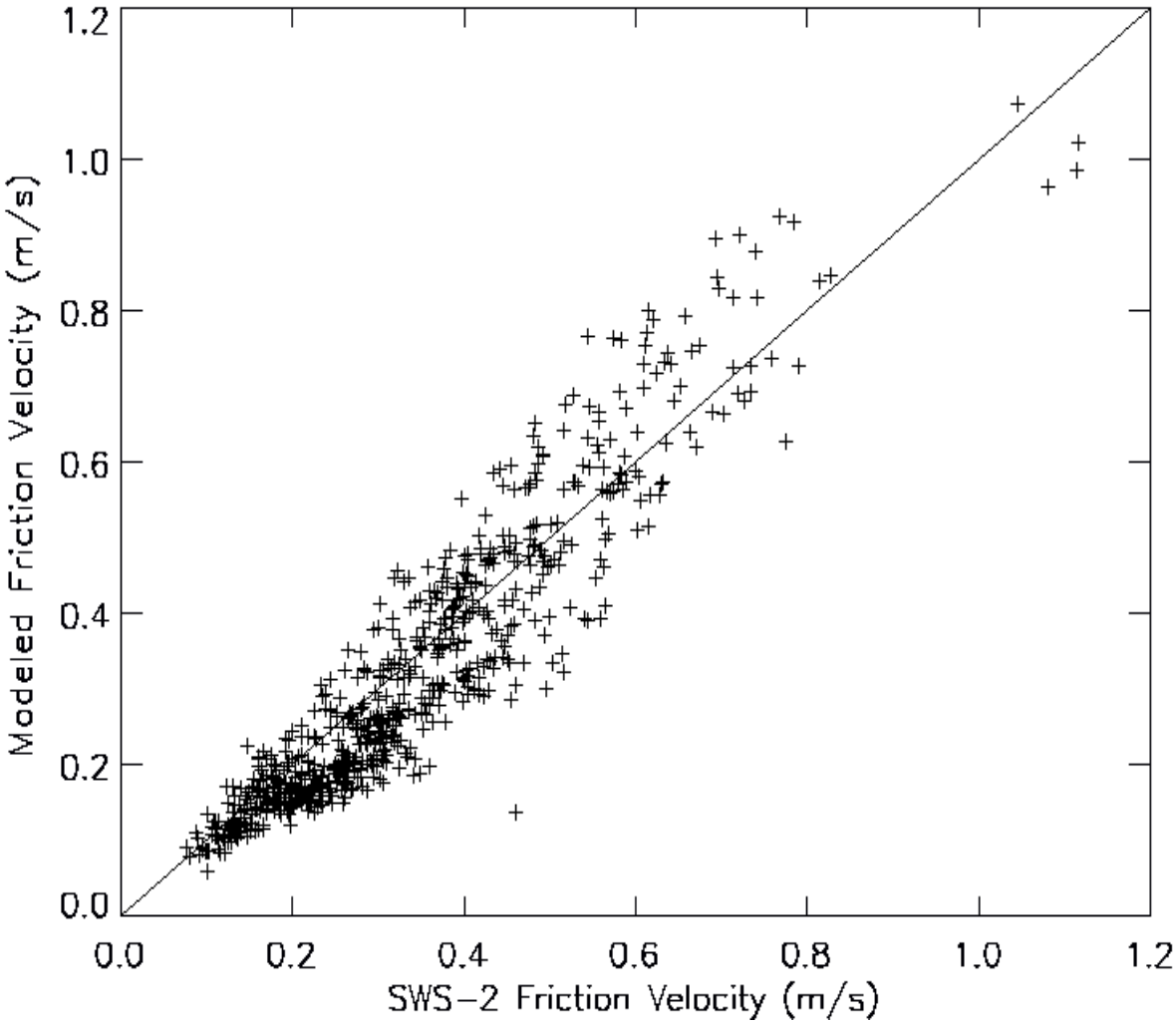
- Satellite radar backscatter vs. NEXRAD rainrates

Example Drag Coefficients From Severe Wind Storms 2 Experiment

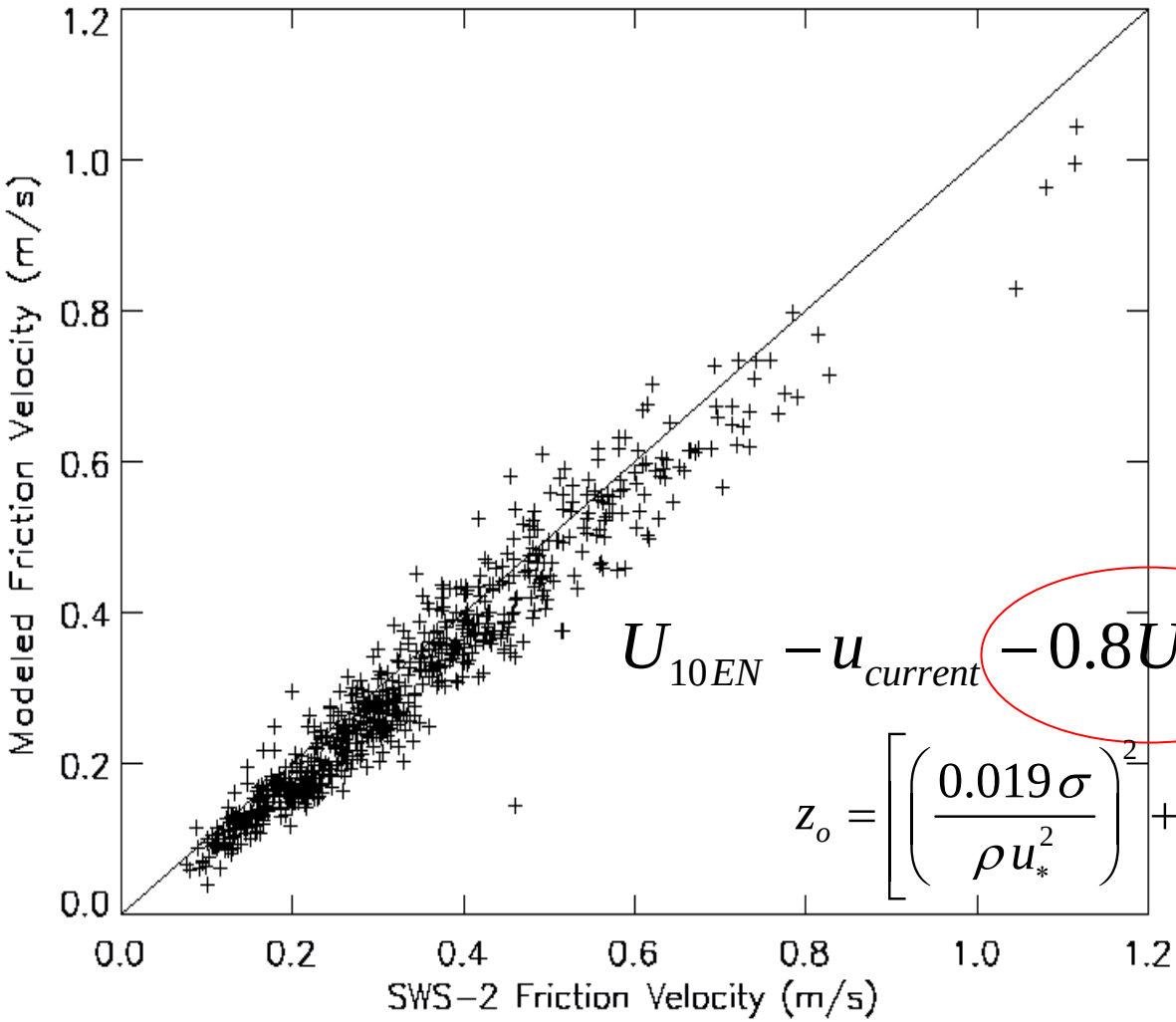


- Preliminary version of the data set provided by Peter K. Taylor.
- These drag coefficients are based on high quality observations.
- Observations that are mostly from rough seas.

Results of Taylor and Yelland's Parameterization on SWS2 data



Bourassa (2004) Comparison to Observations



- This Bourassa guy claims that using better physical assumptions leads to a better model.
  - *Bourassa* (2004, *ASR*)

$$U_{10EN} - u_{current} - 0.8U_{orbital} = \frac{u_*}{k_v}\log\left(\frac{z}{z_o}\right)$$

$$z_o = \left[\left(\frac{0.019\sigma}{\rho u_*^2}\right)^2 + \left(0.068\frac{u_*^2}{g}\right)^2\right]^{0.5}$$

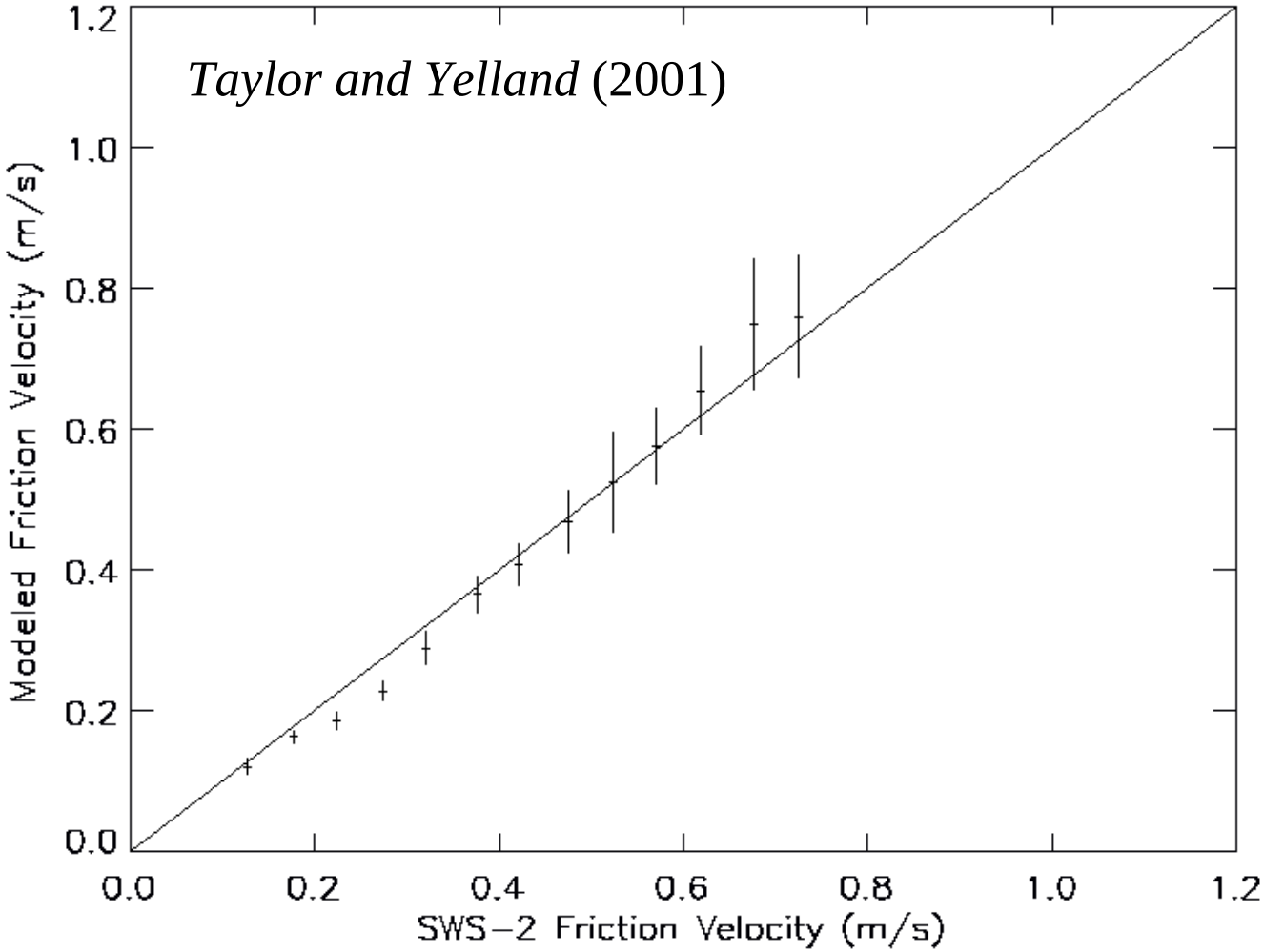
Uncertainty in a Mean

- A problem with using overall measurements of error is that many are very sensitive to outliers.
  - Comparison statistics are largely dependent on a small fraction of the data set.
  - The results are similar even for very different models.
- It is more useful to examine the statistics for small sub-samples of the data set.
  - This can also be misleading for some cases, which will be addressed in later lectures.
  - One useful diagnostic statistic is the uncertainty in the mean.
    - If the errors have a Gaussian distribution, then the uncertainty in the mean is

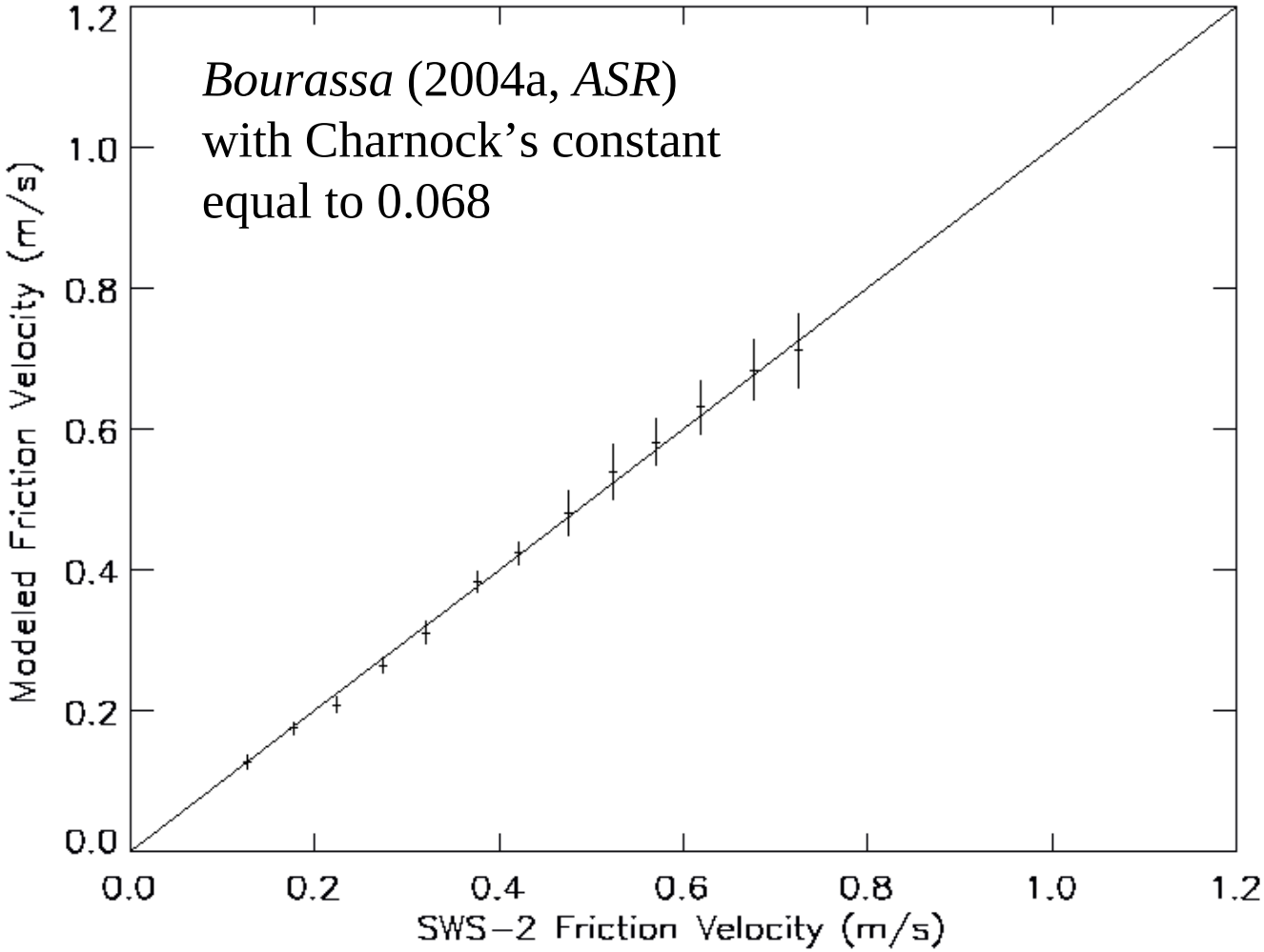
      $$s_{\bar{x}} = s_x / \sqrt{n}$$

    - Where n refers to the number of independent points in the sample (or sub-sample)

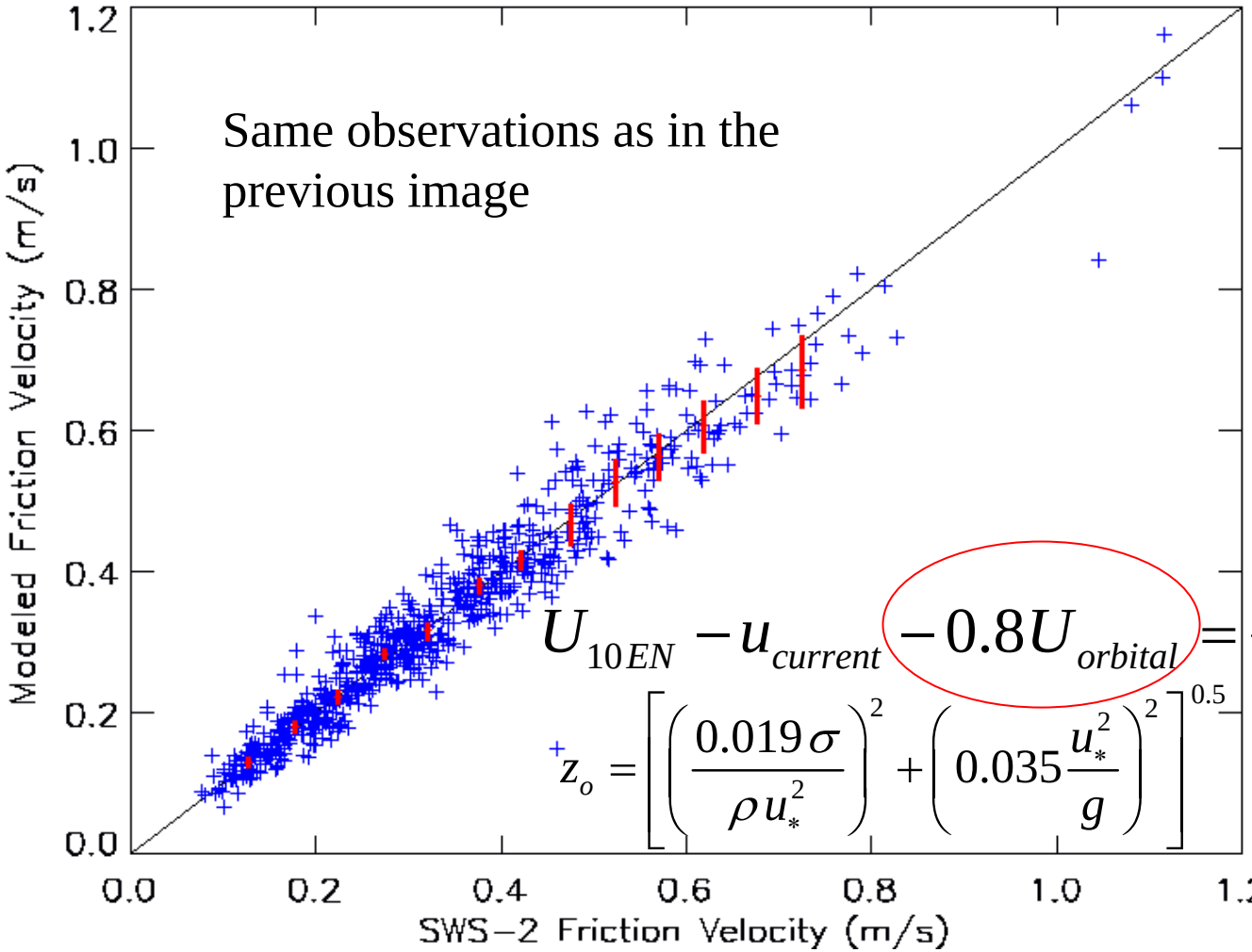
More Results:
Means & Three Standard Deviations



More Results:
Means & Three Standard Deviations



Results of Bourassa (2005) Compared to SWS2 Observations



$$U_{10EN} - u_{current} - 0.8U_{orbital} = \frac{u_*}{k_v}\log\left(\frac{z - 0.8H_s}{z_o}\right)$$

$$z_o = \left[\left(\frac{0.019\sigma}{\rho u_*^2}\right)^2 + \left(0.035\frac{u_*^2}{g}\right)^2\right]^{0.5}$$

- A purely descriptive technique can be combined with a quantitative technique to get a great deal of information.