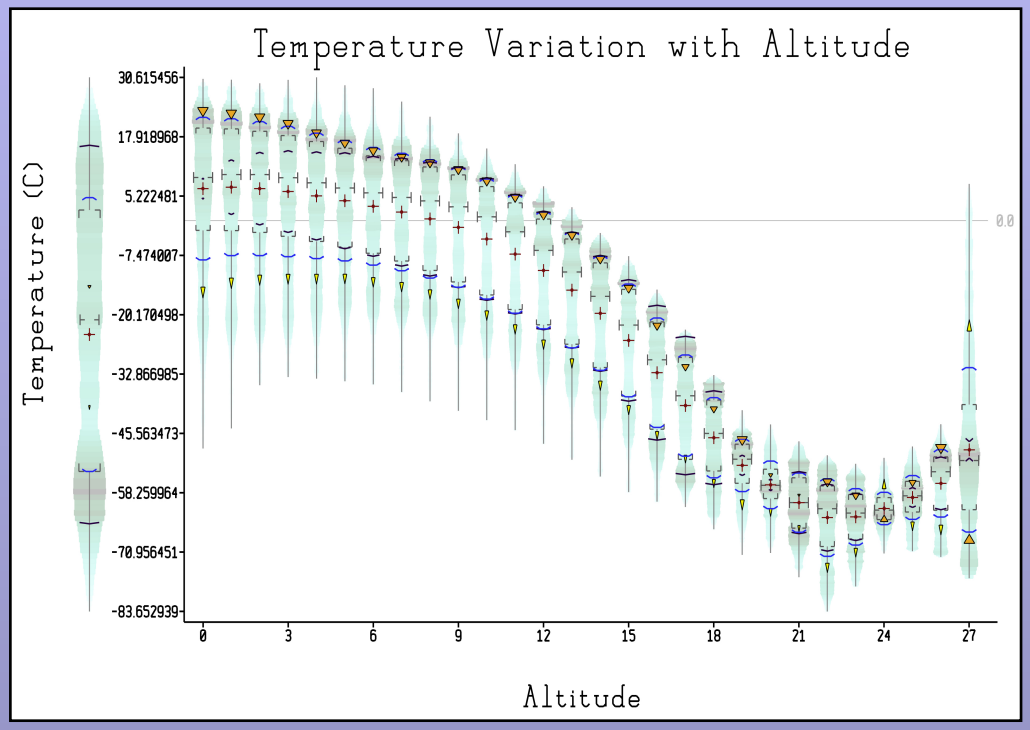
Visual Summary Statistics



Kristin Potter

June 14, 2006

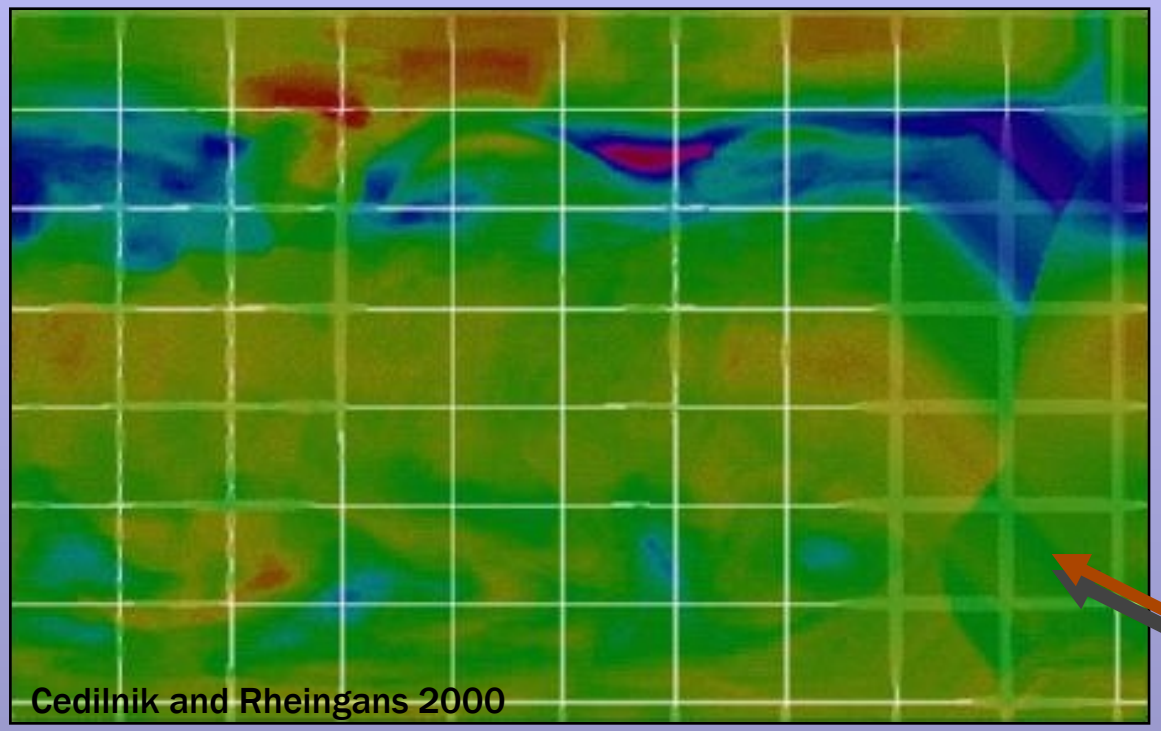
Visualization

- Quickly convey large amount of data/ information
- Create a visual representation of information that may not have a spatial locality
- Take advantage of the high bandwidth of the human visual system
- Give insights to relationships and interactions not possible when looking just at the "numbers"

Missing Qualitative Information

- Visualizations often lack uncertainty, error and confidence info
- Often presented alongside as tables, charts and graphs
- Should be incorporated into visualizations to maintain fidelity to data

Uncertainty Visualization



Area with high levels of uncertainty

- Add in qualitative information
- Annotate location and magnitude
- Typically modify a visual attribute of the data

Problems with Uncertainty Visualization

- Express syntax of uncertainty as "unknown" through blur, noise
- Can create distracting, unreadable visualizations
- Does not express the actual values of measures categorized as uncertainty

Goals of This Work

1. Rather than rendering "uncertainty" render specific, meaningful, known values
2. Understand the various measures that can describe uncertainty
3. Design visualization techniques to express these measures

What is uncertainty?

- Typically a measure of the standard deviation (σ) of the data

Given a data set $\{x_i\}_{i=0}^{N}$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i - \mu_1)^2}$$

where N is the number of samples, μ_1 is the mean

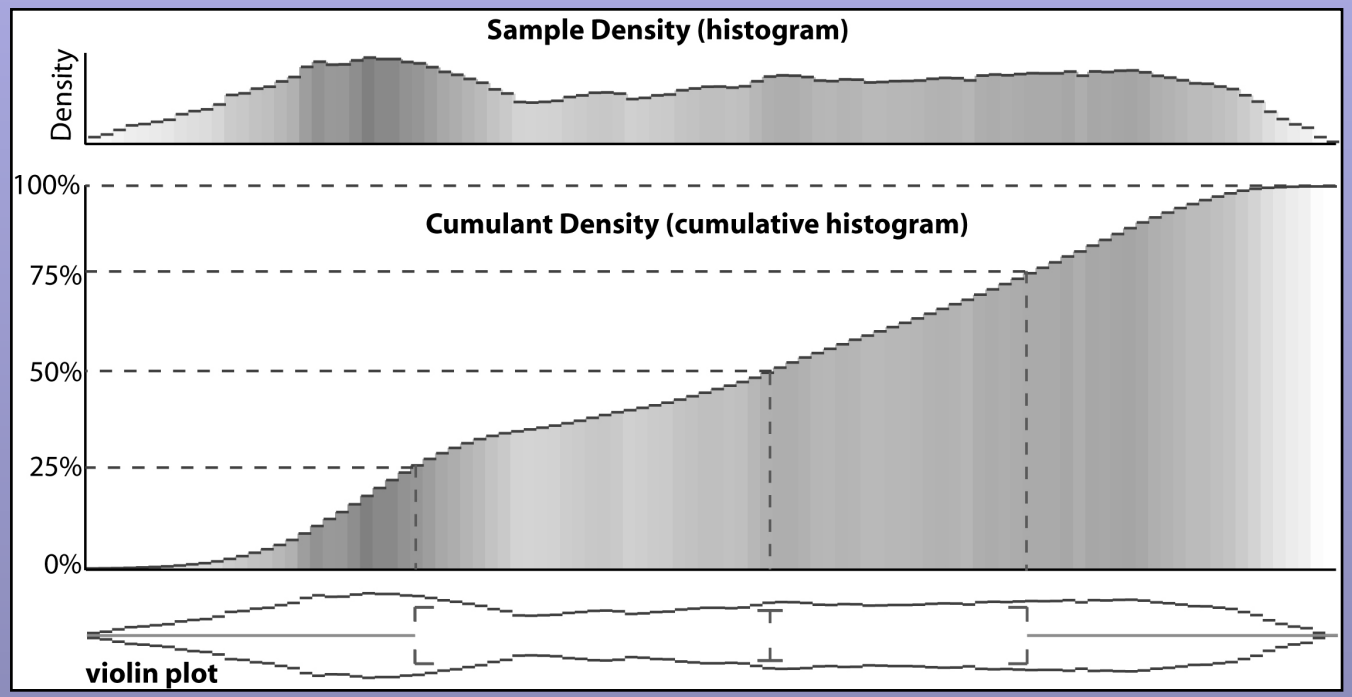
- A measure of the dispersion of the data, that is the amount of variation from the average/mean

Alternative Definitions of Uncertainty

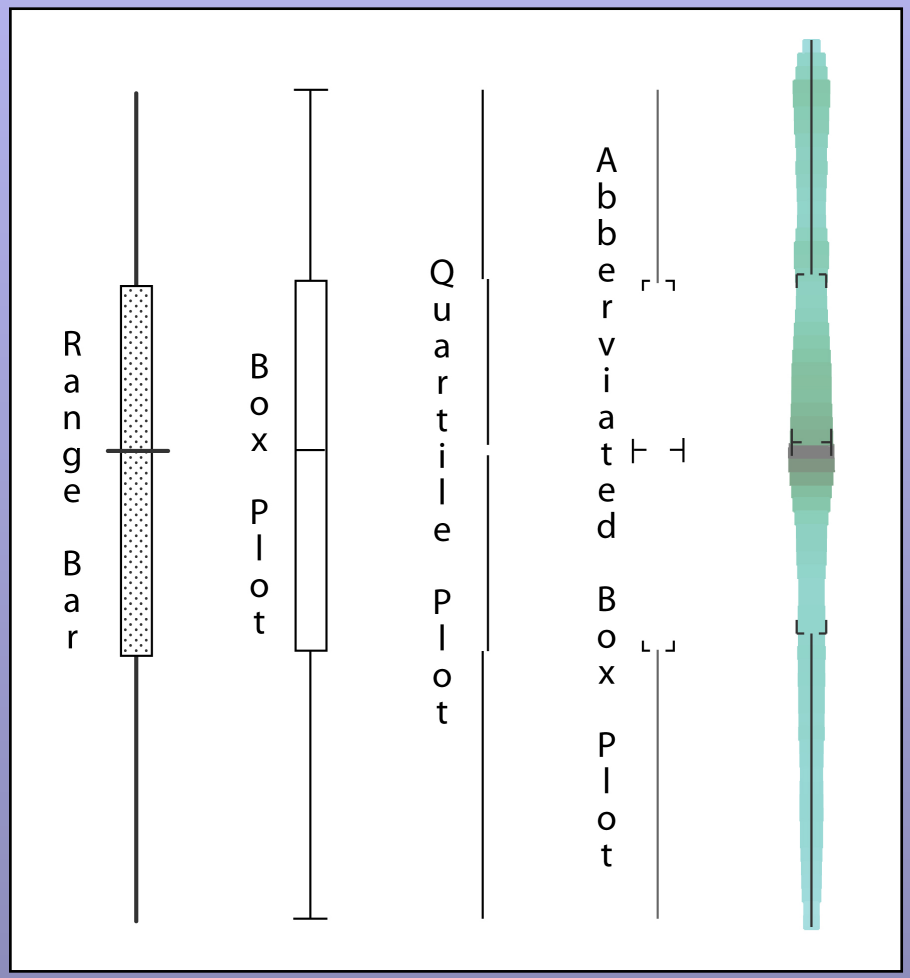
- Describe the distribution of the data, not only standard deviation
  - Quartiles
  - Moments
- Measures of Error
- Confidence Levels

Quartiles

- Partition the data into four equally sized subsets
  - Median, upper/lower quartiles, min/max



Box Plots

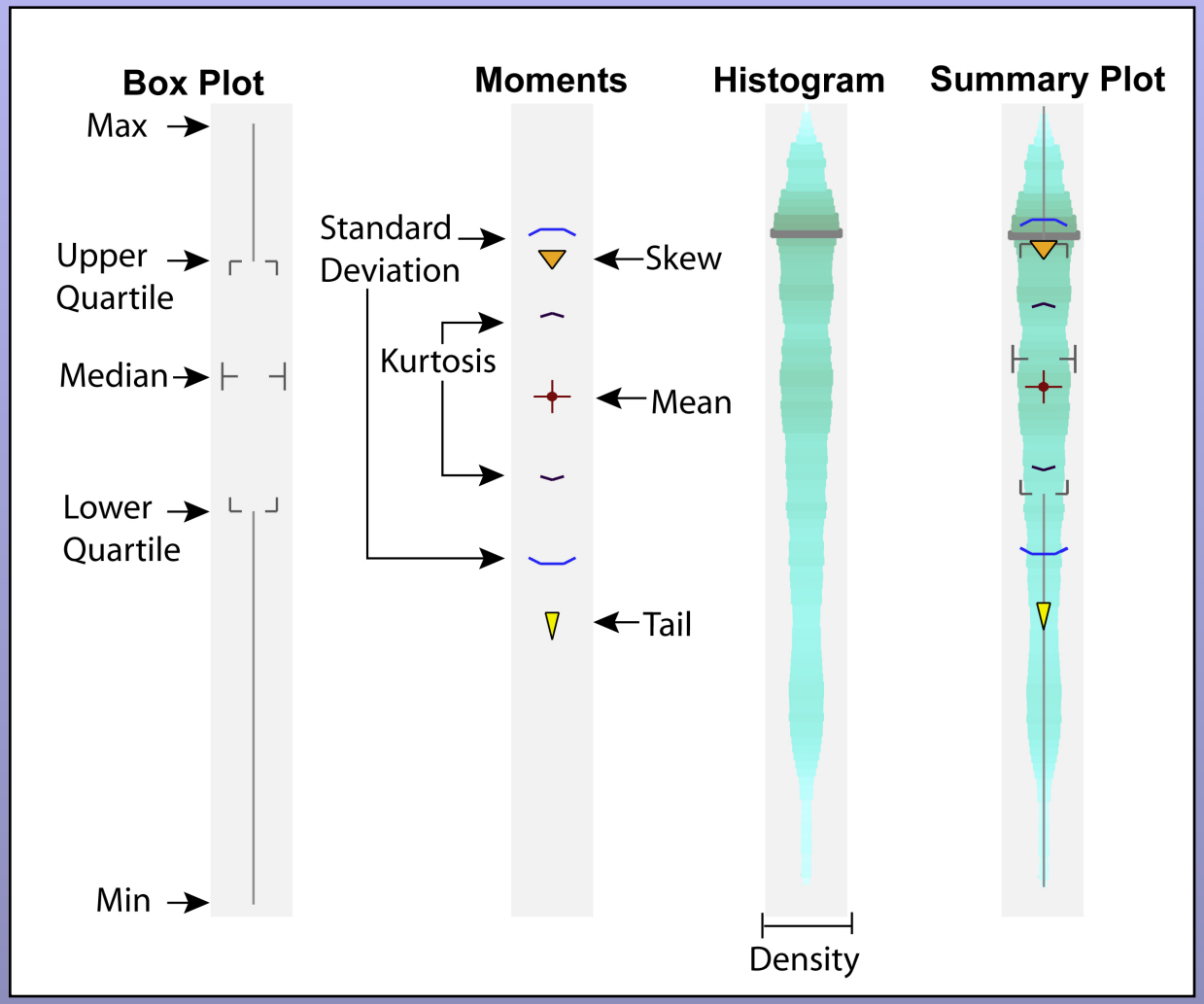


- Typical method to render quartiles
- Median, upper/lower quartiles, min/max
- Iteratively simplified over the years
- Simplicity can allow the addition of more information

Why Add More Information?

- Multiple datasets can have the same quartile description, but may not be the same distribution
  - i.e. multi-modal distributions
- Quartiles describe the density distribution of the data, but not more subtle characteristics

The Summary Plot



Moments

Central Moments:

Given a data set $\{x_i\}_{i=0}^{N}$

$$\mu_k = \frac{1}{N}\sum_{i=0}^{N}(x_i - \mu_1)^k$$

where N is the number of samples, μ_1 is the mean

k refers to the desired moment

Mean and Standard Deviation

- k = 1 is the mean, the "expected" value

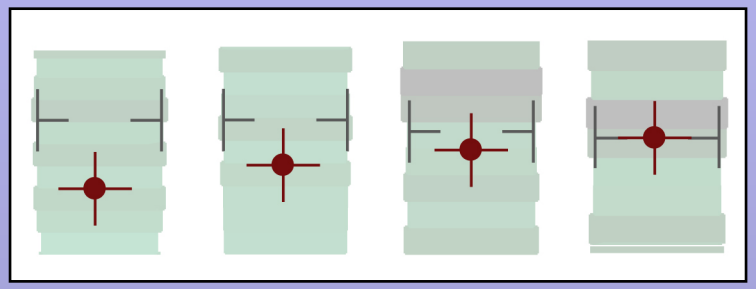
$$\mu_1 = \frac{1}{N} \sum_{i=0}^{N} x_i$$

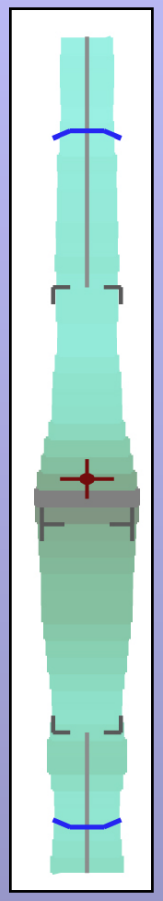
- k = 2 is the variance, σ is standard deviation

$$\sigma = \sqrt{\mu_2}$$

- Measures how spread out the data values are

Mean and Standard Deviation Visualization



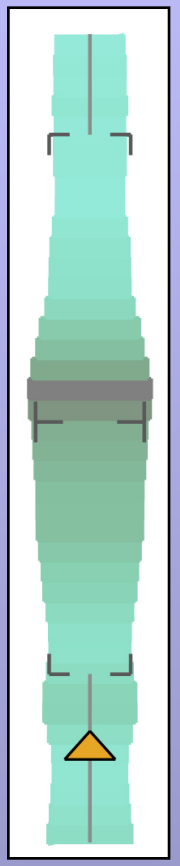


- Mean rendered as a red bull's eye
  - Lines up with median when equal
- Standard Deviation rendered as blue parenthesis
  - equal #'d moment, so 2 glyphs

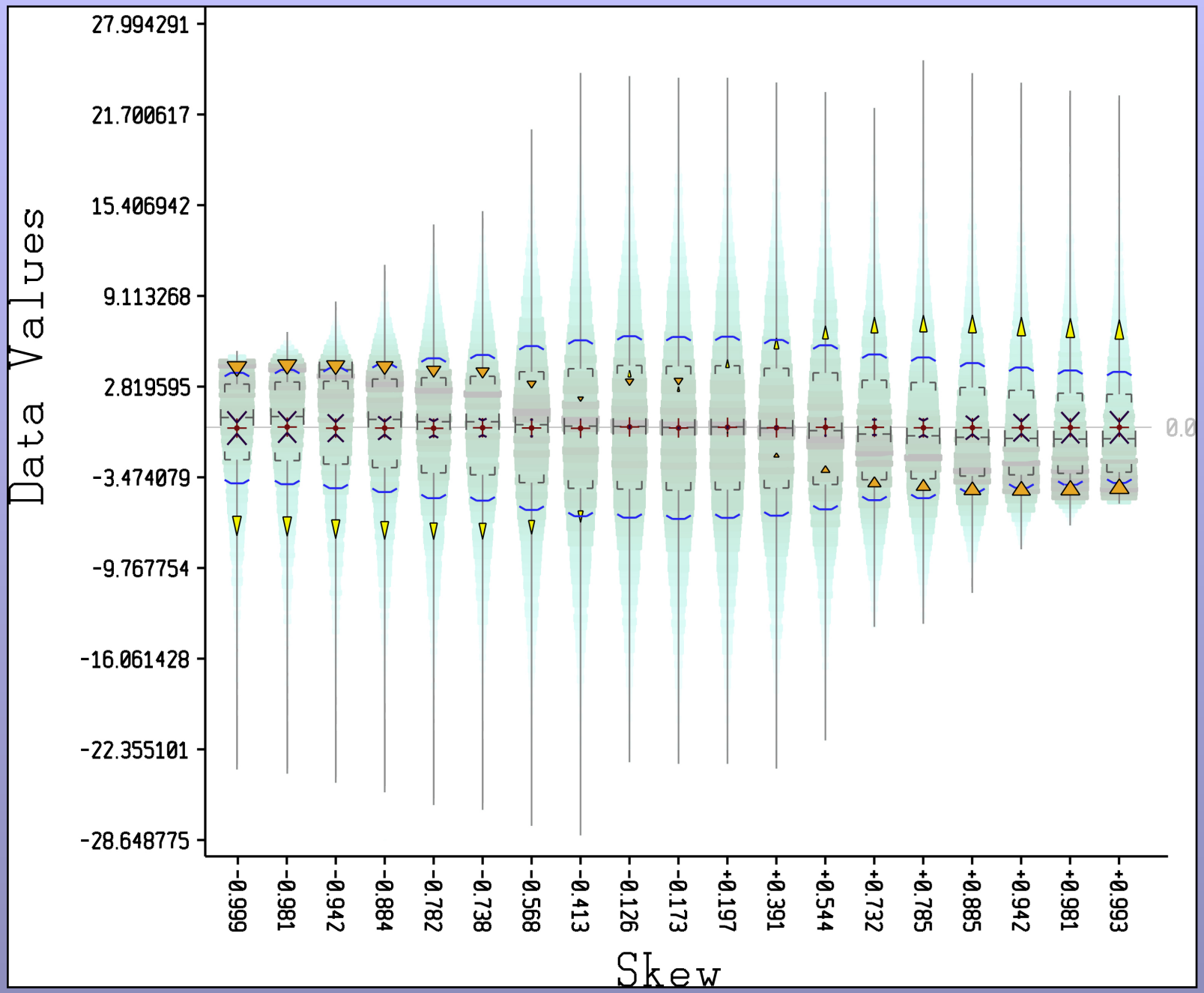
Skew

- Measures how asymmetrical the distribution is

$$\gamma = \frac{\mu_3}{\sigma^3}$$



- Normalized by standard deviation
- Visualized as a scaled, orange triangle resting on the end of the distribution with the most weight, pointing towards the long tail


Data Values
27.994291
21.700617
15.406942
9.113268
2.819595
-3.474079
-9.767754
-16.061428
-22.355101
-28.648775
0.0
-0.999
-0.981
-0.942
-0.884
-0.782
-0.738
-0.568
-0.413
-0.126
-0.173
+0.197
+0.391
+0.544
+0.732
+0.785
+0.885
+0.942
+0.981
+0.993
Skew

Kurtosis

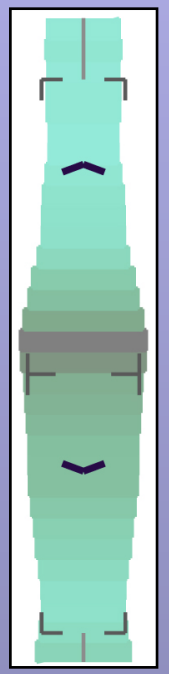
- Measures the "peakiness" of the distribution

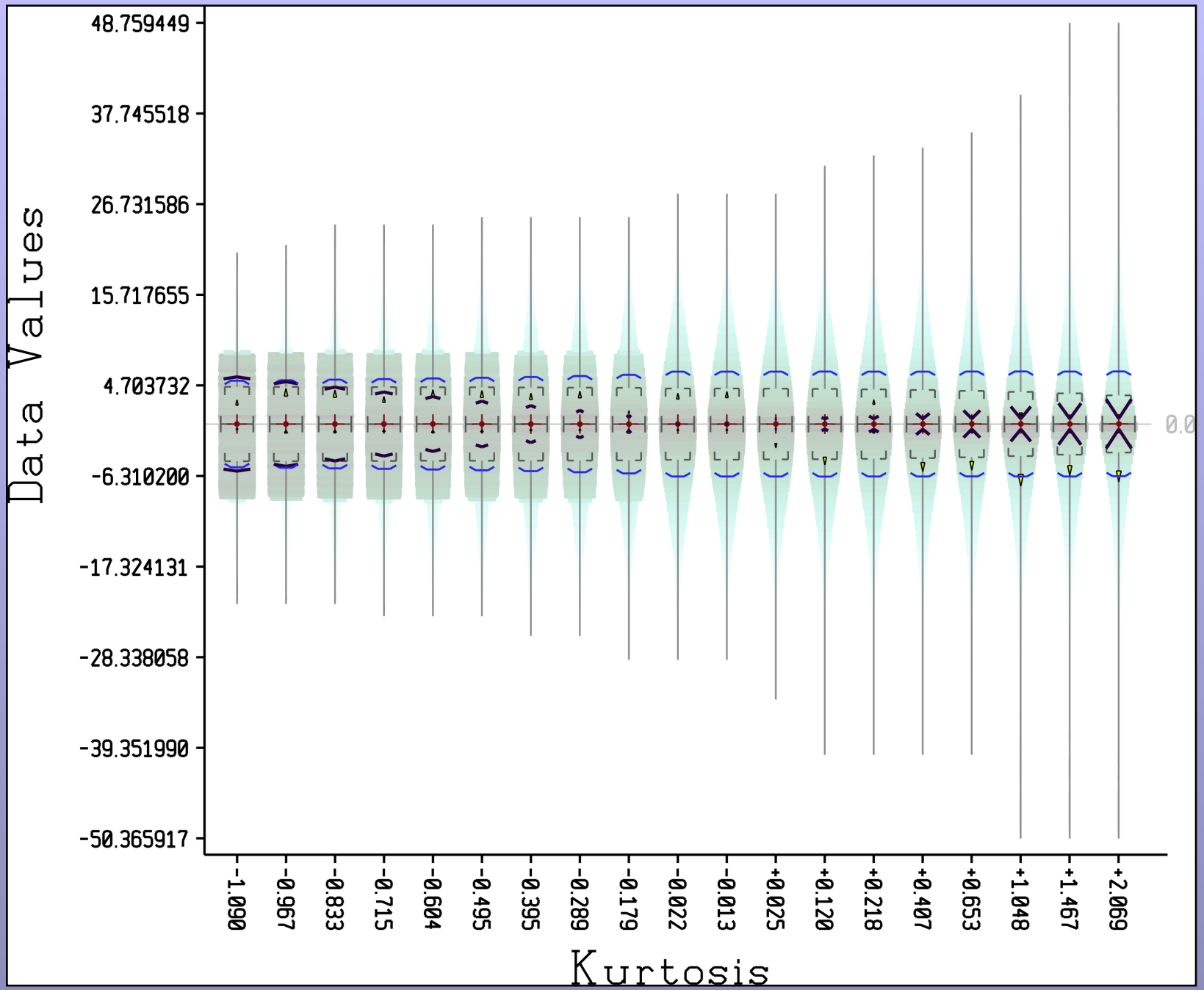
$$\kappa = \frac{\mu_4}{\sigma^4}$$

- Excess Kurtosis (normalize by the Gaussian distribution)

$$\kappa_e = \frac{\mu_4}{\sigma^4} - 3$$

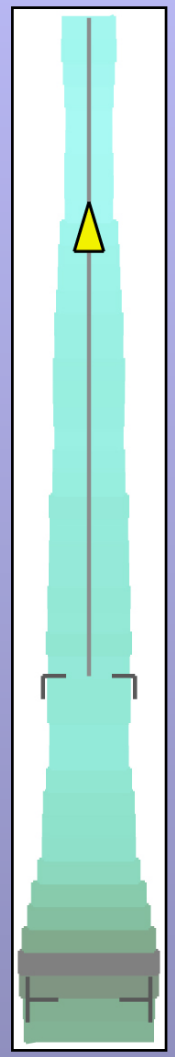
- Rendered as 2 glyphs that are scaled by the value, similar to the parens of standard deviation, but "pointiness" scaled by sign of kurtosis

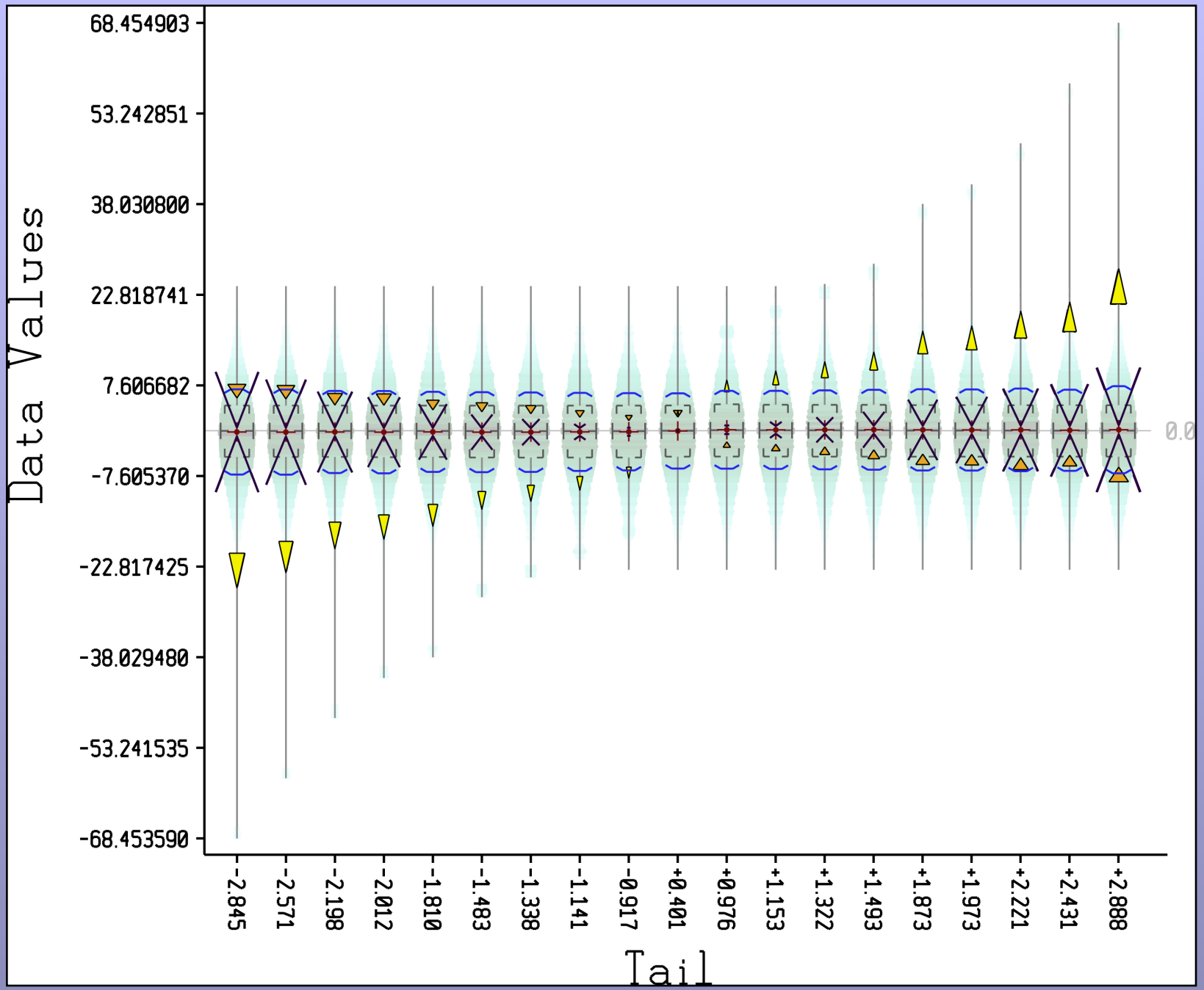



Data Values
48.759449
37.745518
26.731586
15.717655
4.703732
-6.310200
-17.324131
-28.338058
-39.351990
-50.365917
0.0
-1.090
-0.967
-0.833
-0.715
-0.604
-0.495
-0.395
-0.289
-0.179
-0.022
-0.013
+0.025
+0.120
+0.218
+0.407
+0.653
+1.048
+1.467
+2.069
Kurtosis

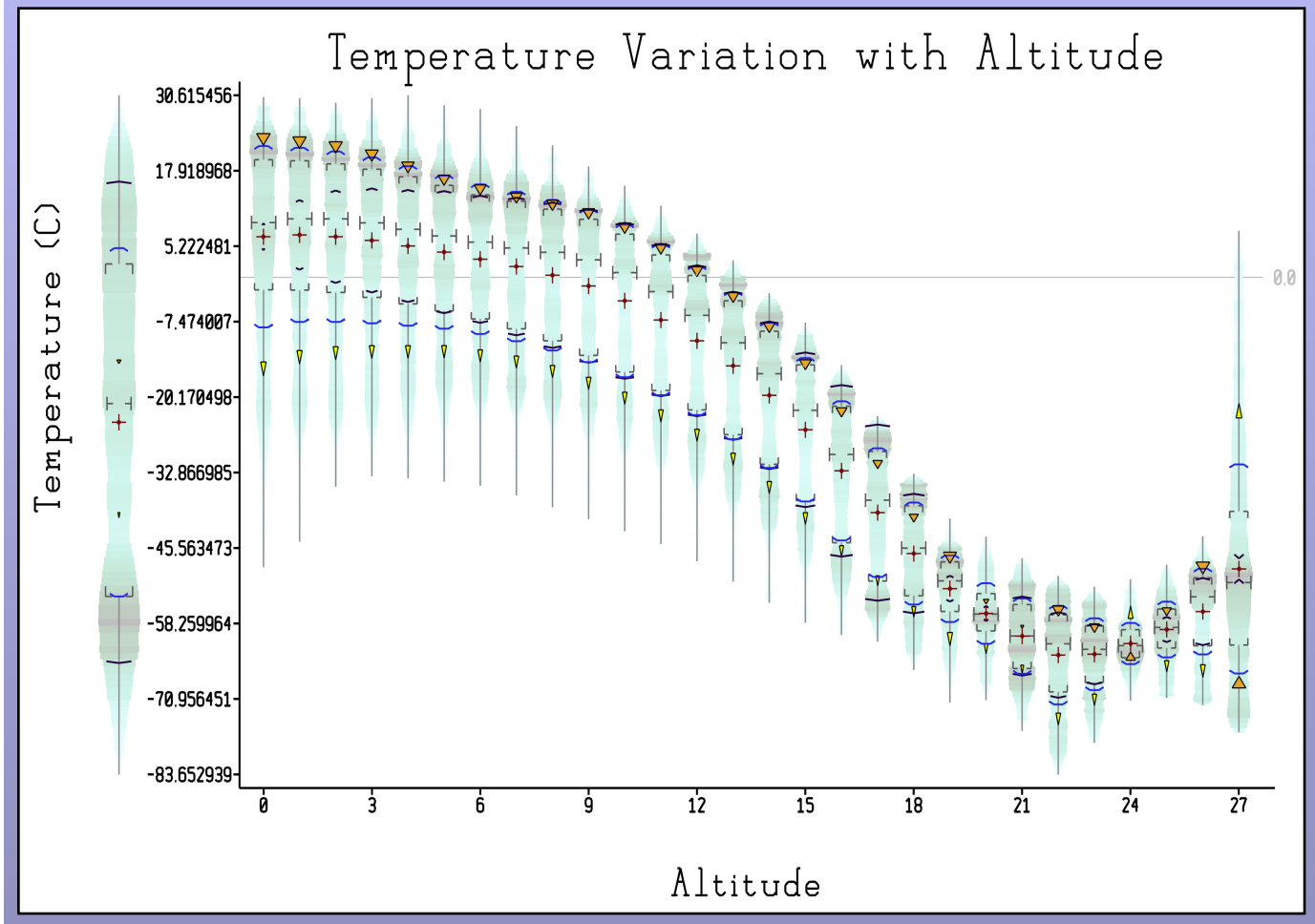
Fifth Central Moment

- Measures distribution asymmetry farther away from the mean when compared to the skew
- Rendered as a smaller, yellow triangle, size and sharpness scaled by value, placed pointing towards asymmetry

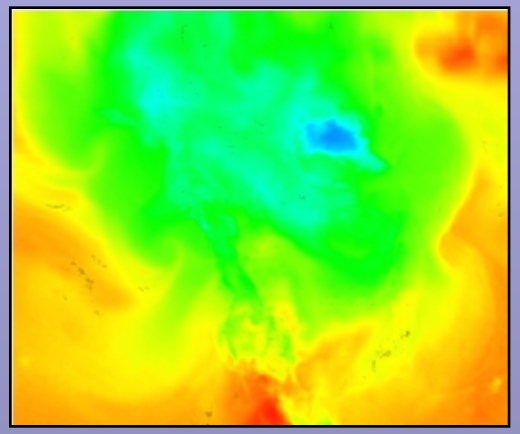



Data Values
68.454903
53.242851
38.030800
22.818741
7.606682
-7.605370
-22.817425
-38.029480
-53.241535
-68.453590
0.0
-2.845
-2.571
-2.198
-2.012
-1.810
-1.483
-1.338
-1.141
-0.917
+0.401
+0.976
+1.153
+1.322
+1.493
+1.873
+1.973
+2.221
+2.431
+2.888
Tail

Temperature Example

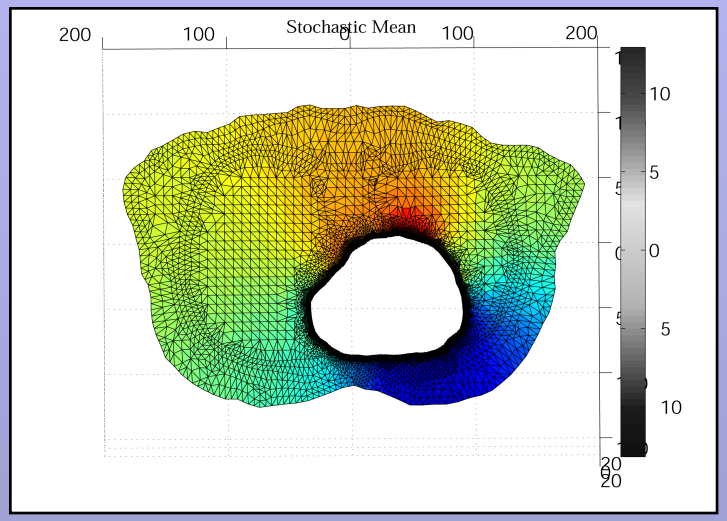


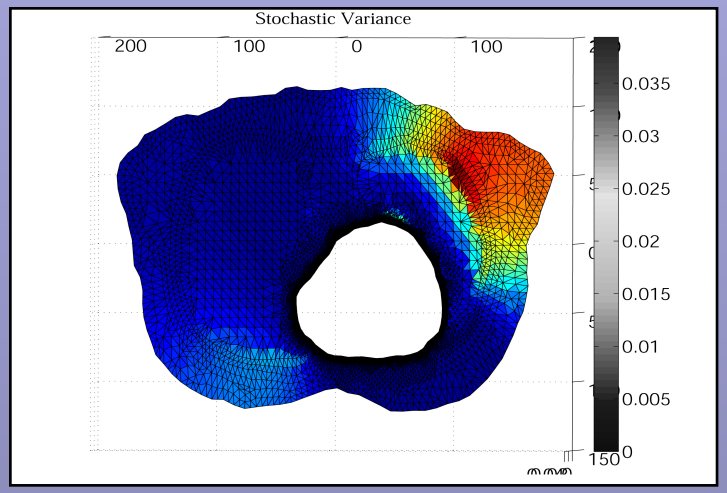
Map

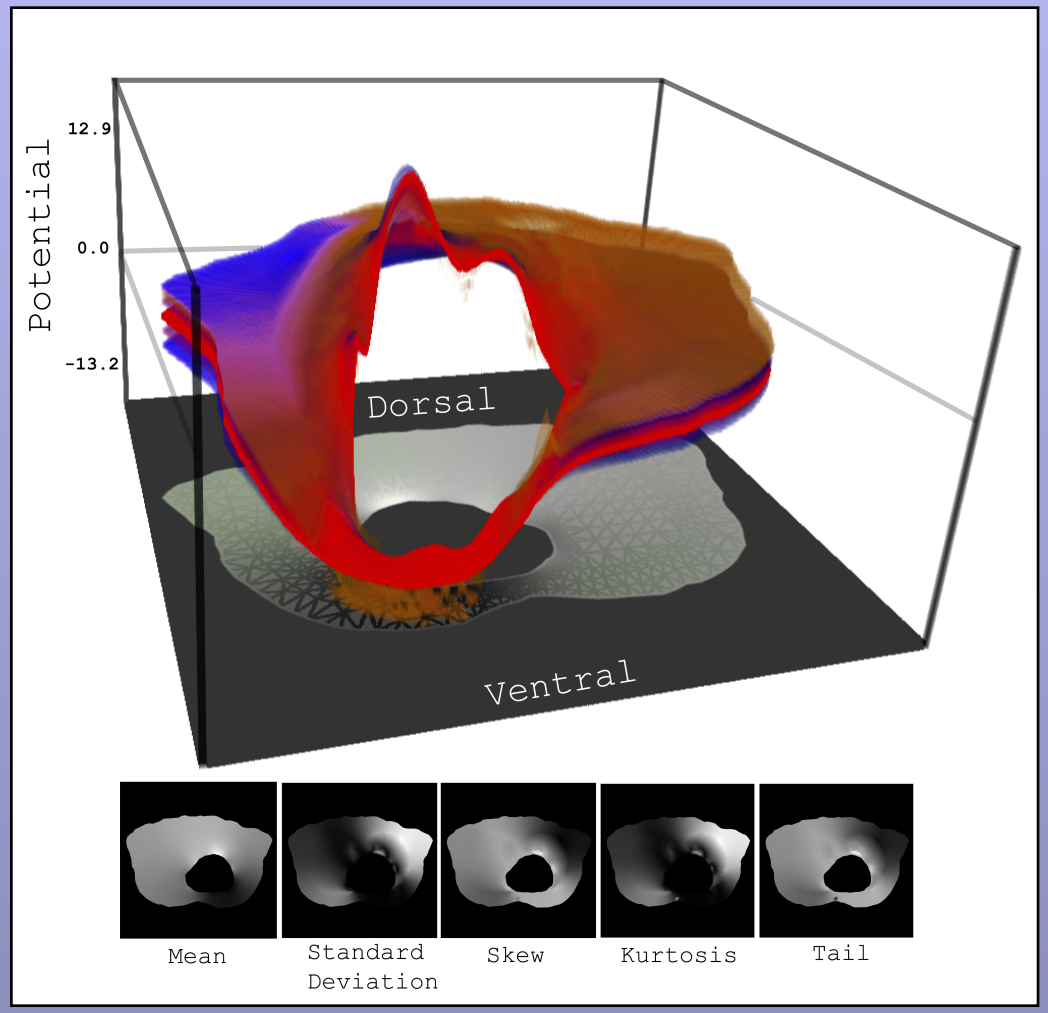


Temperature

Electric Potentials Example







Correlated Data

- Visualize relationships between multiple sets of correlated data
- Joint histogram
- Covariance

Given two data sets, $\{x_i\}_{i=0}^{N}$, $\{x_j\}_{j=0}^{N}$,

$$V_{ij} = \frac{1}{N-1}\sum_{k=0}^{N}(x_{i_k} - \mu_i)(x_{j_k} - \mu_j)$$

where N is the number of samples

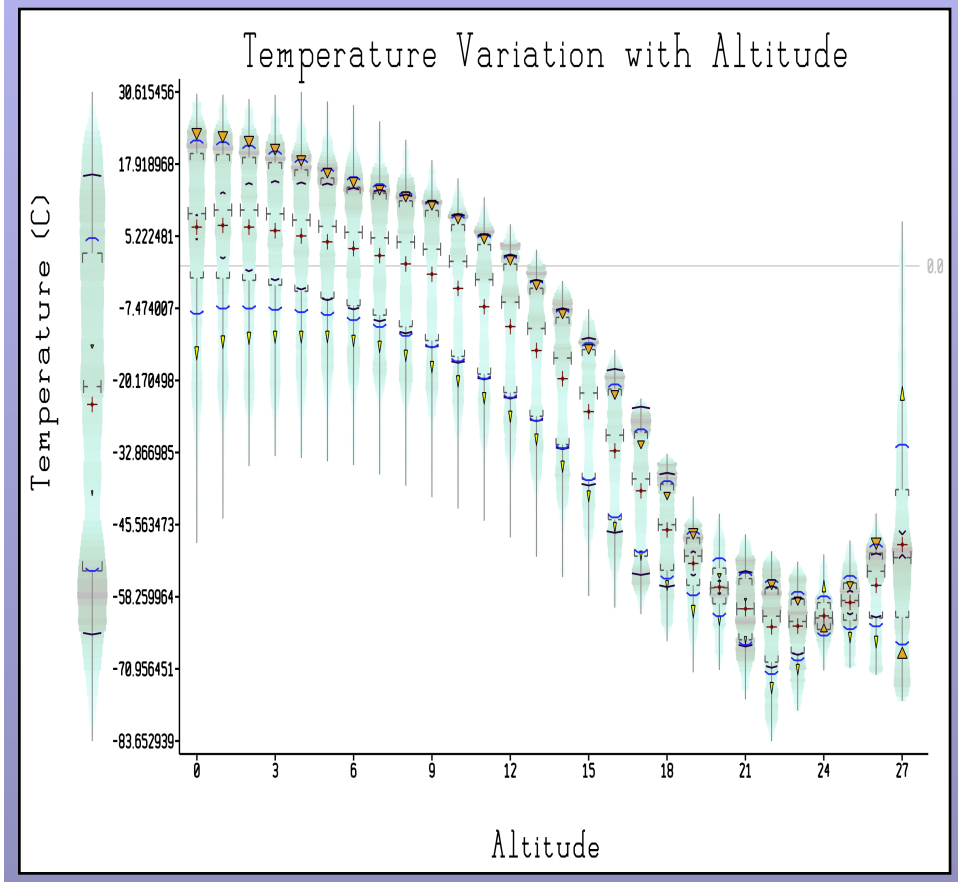
Skew variance

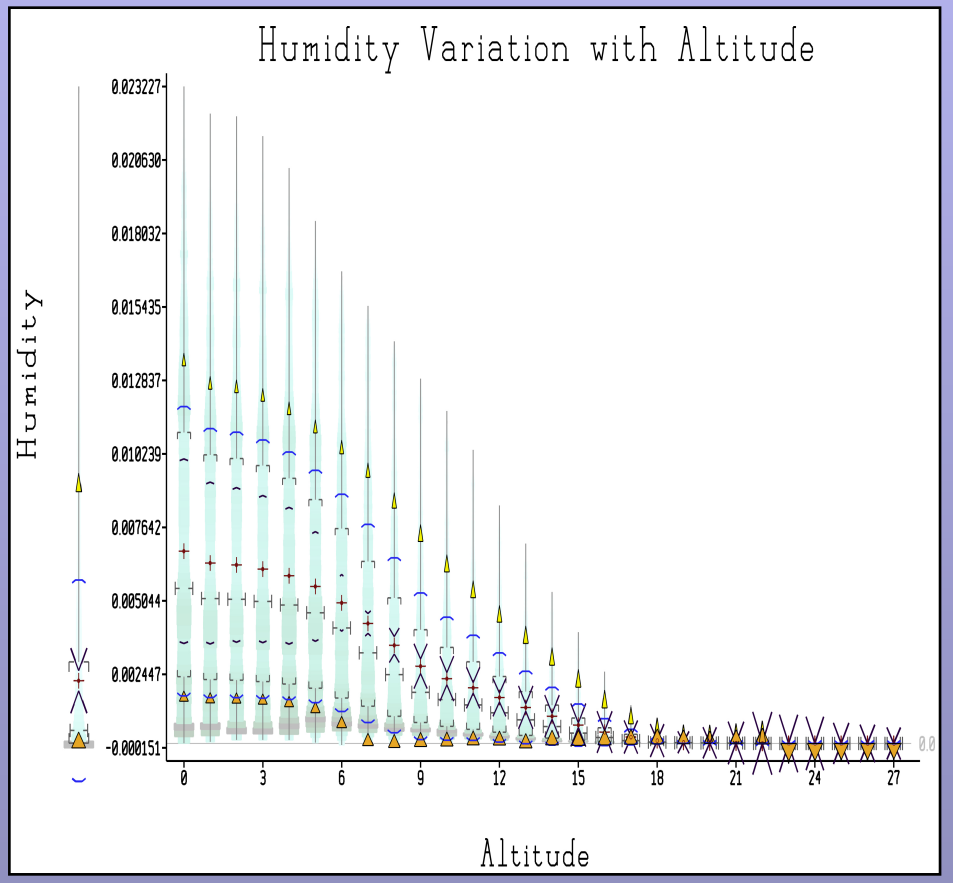
Given two data sets, $\{x_i\}_{i=0}^{N}$, $\{x_j\}_{j=0}^{N}$,

$$V_{i^m j^n} = \frac{1}{N-1}\sum_{k=0}^{N}(x_{i_k} - \mu_i)^m(x_{j_k} - \mu_j)^n$$

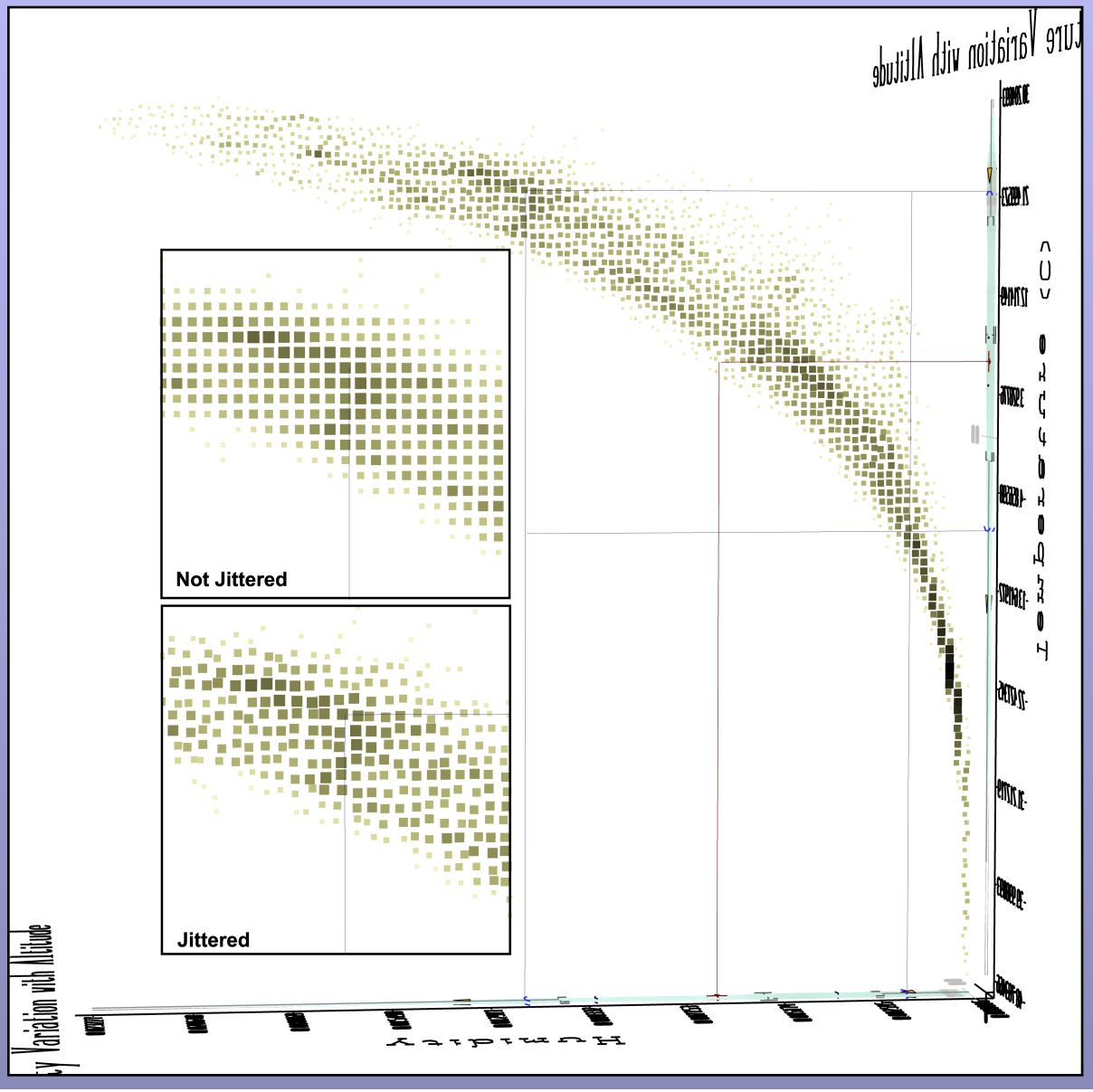
where N is the number of samples

Temperature/Humidity Example

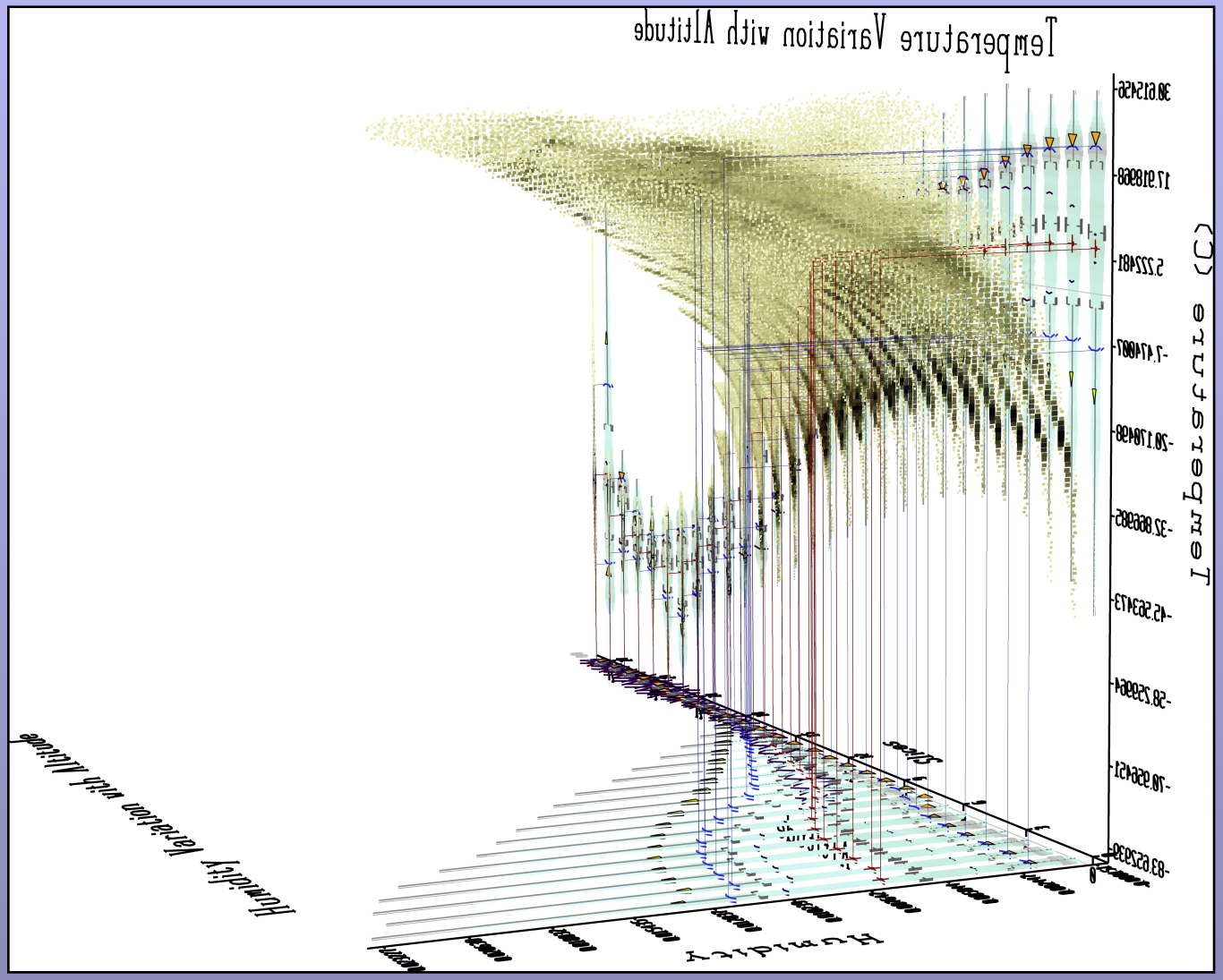


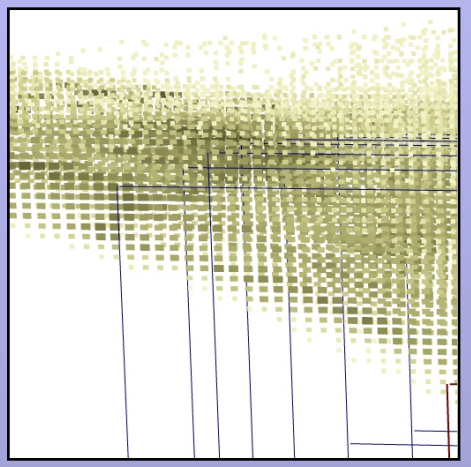


Joint Histogram – 1 category

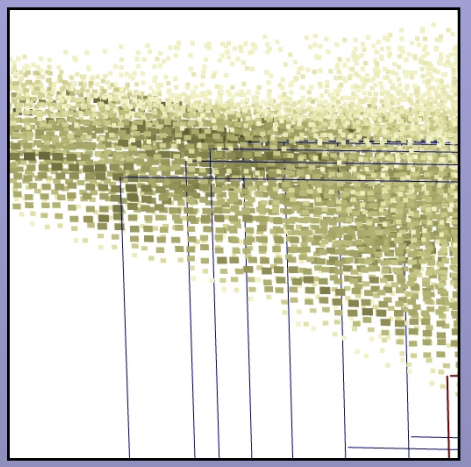


Joint Histogram – all categories



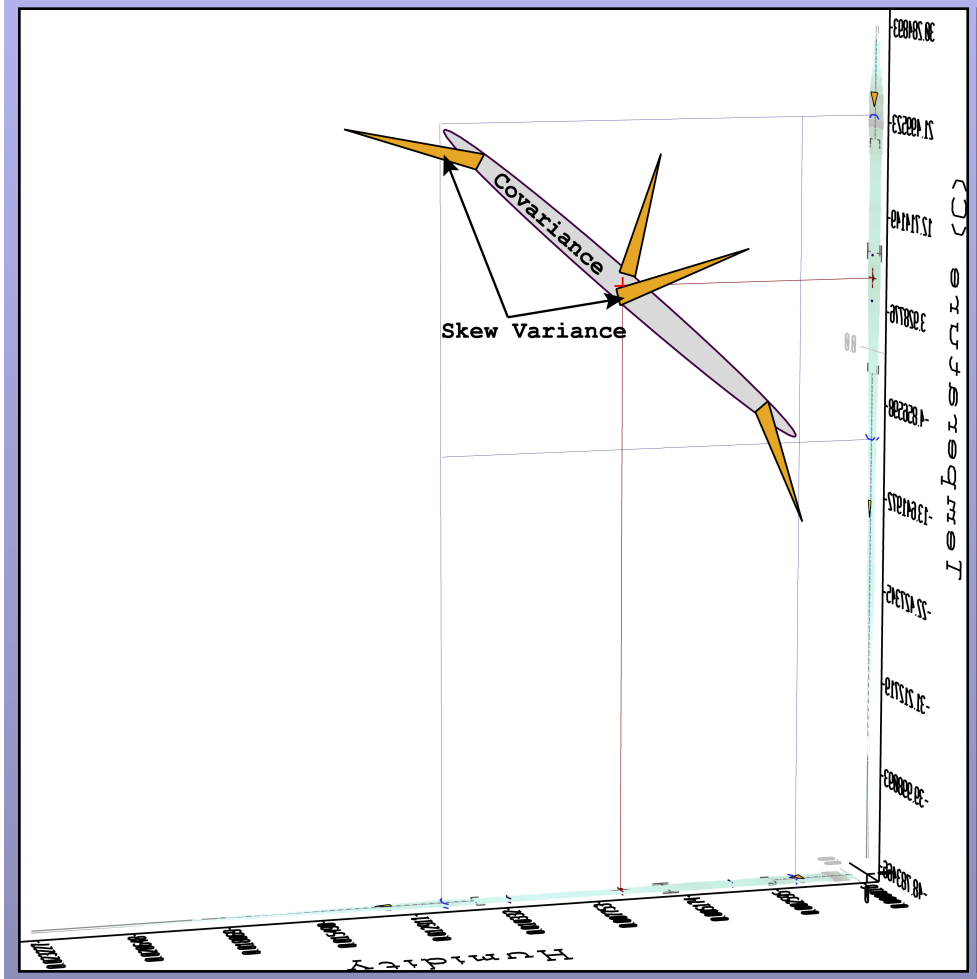


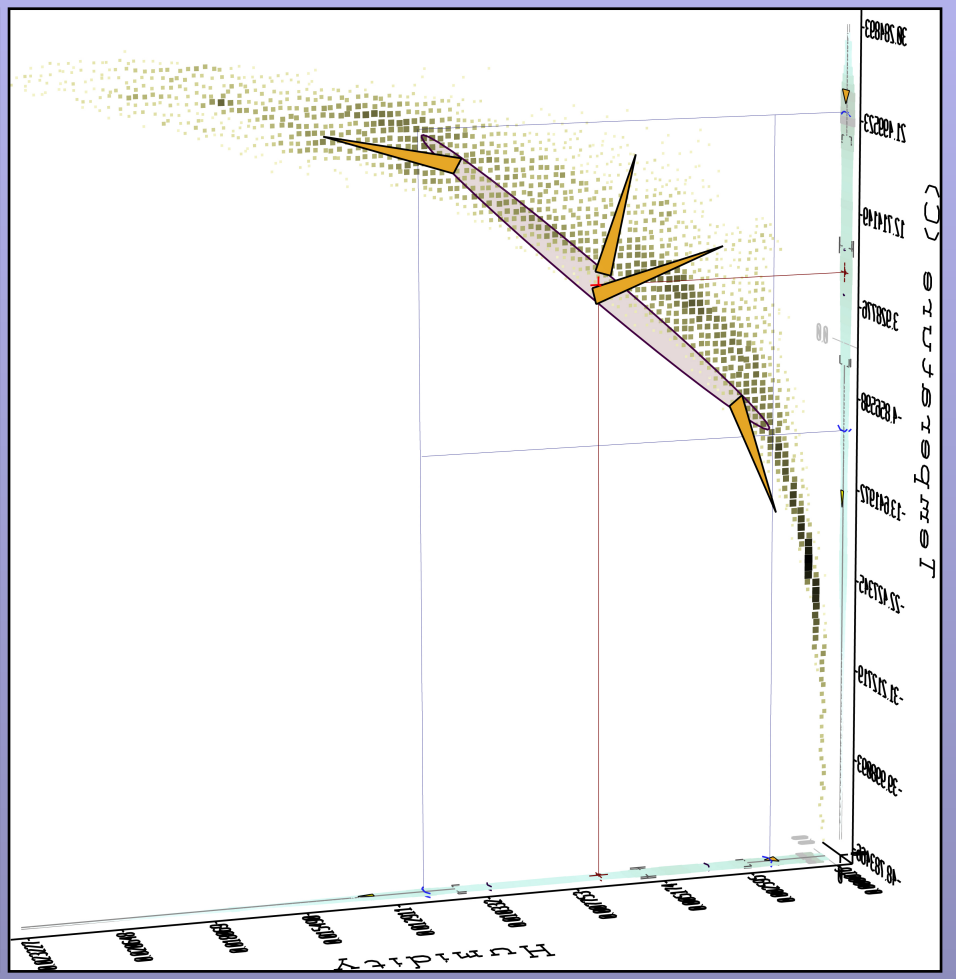
No Jitter



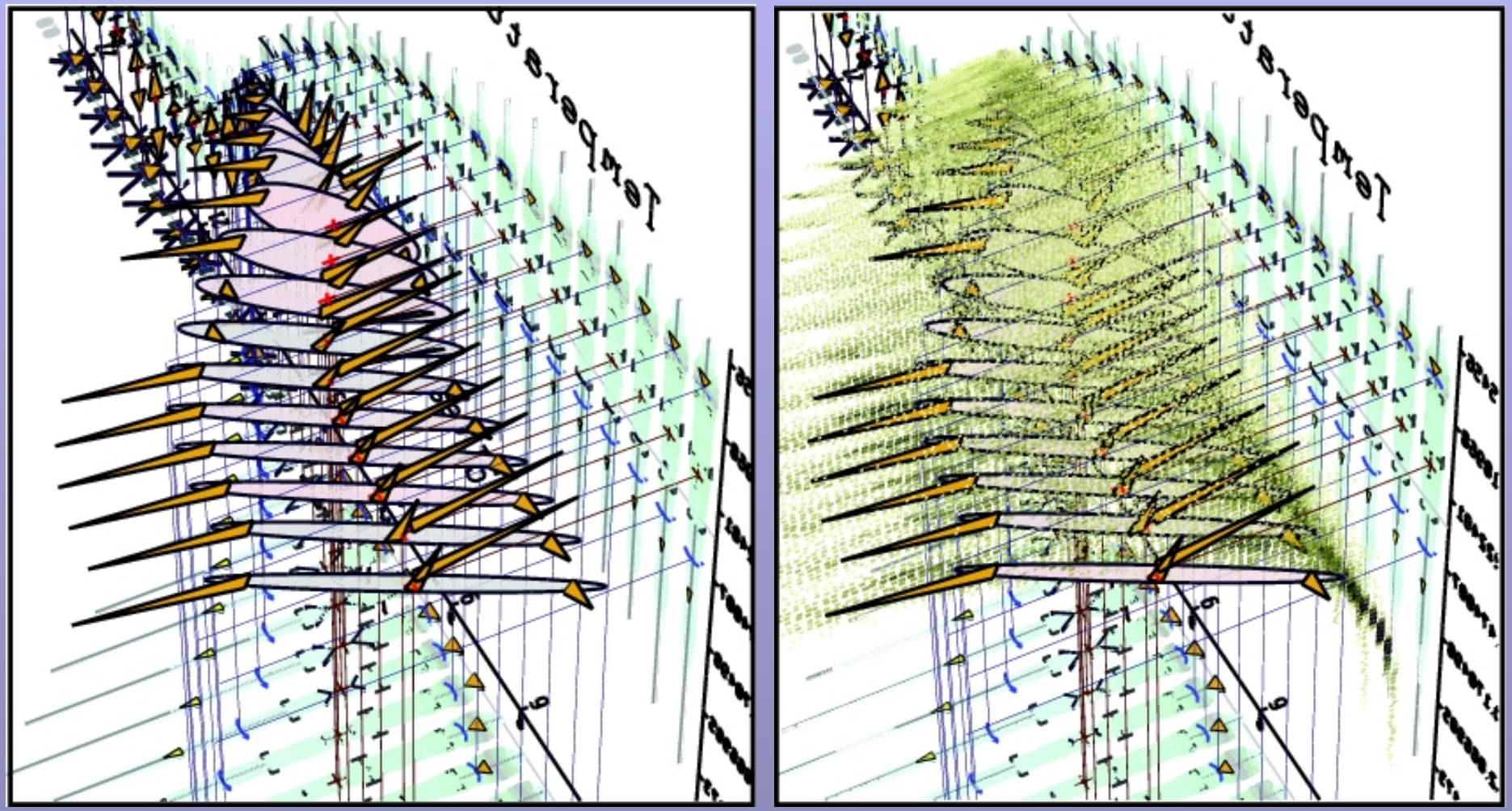
Jitter

Covariance and Skew Variance - 1 category





Covariance and Skew Variance - all categories



Discussion

- Issues arise with the sensitivity of the higher order moments.
  - Need more robust methods for computing these measures
- Too much visual clutter
  - Add a UI
- Add other measures of uncertainty
  - Error and confidence

Conclusion

- Visualizations should have associated qualitative measures included in images
- Current uncertainty visualization techniques could be improved by looking at other measures of uncertainty
- Adding information into visualizations is hard and requires good design

Thanks!

Acknowledgements:

Joe Kniss, Richard Riesenfeld, Chris Johnson and Chuck Hansen for contributions to this work.
The SCI Institute for lab space and equipment.
Mike Kirby and Sarah Geneser for the electric potential data, generated under ARO grant number W911NF-05-1-0395.
Weather simulation data courtesy of the Canadian Meteorological Center.

This work was funded under NSF grant # 03-12479.