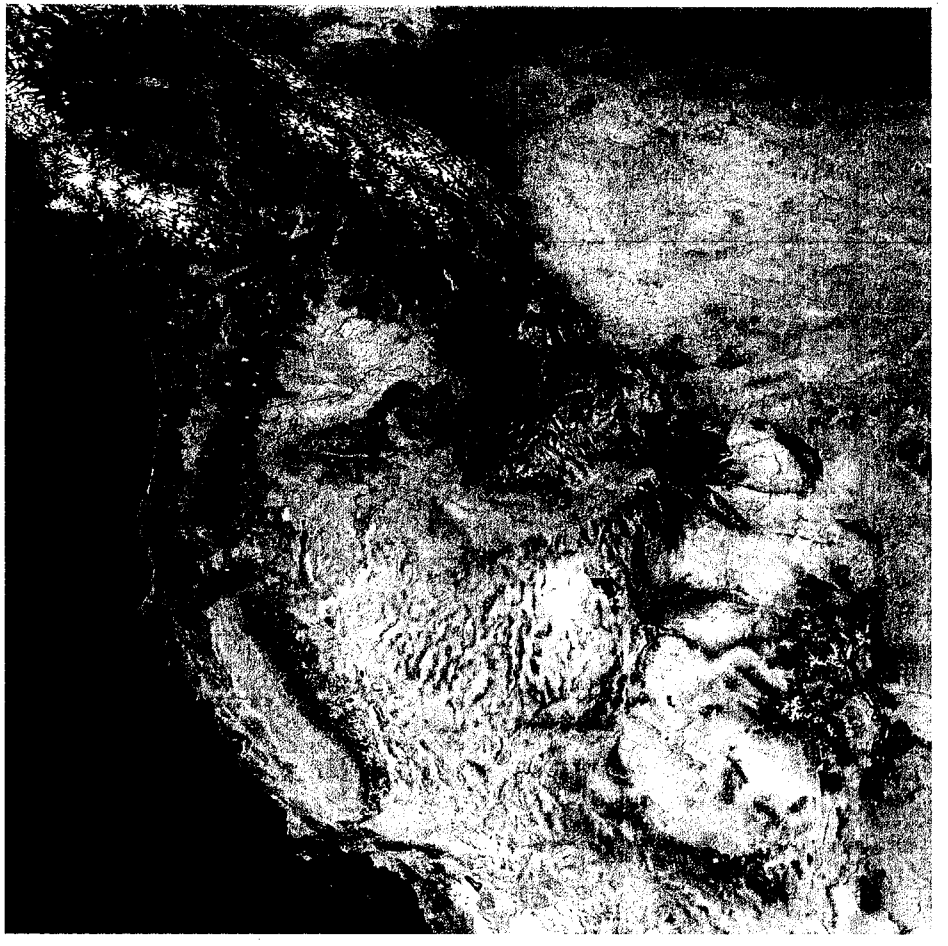
Reducing Size and Complexity of Very Large Satellite Data Sets

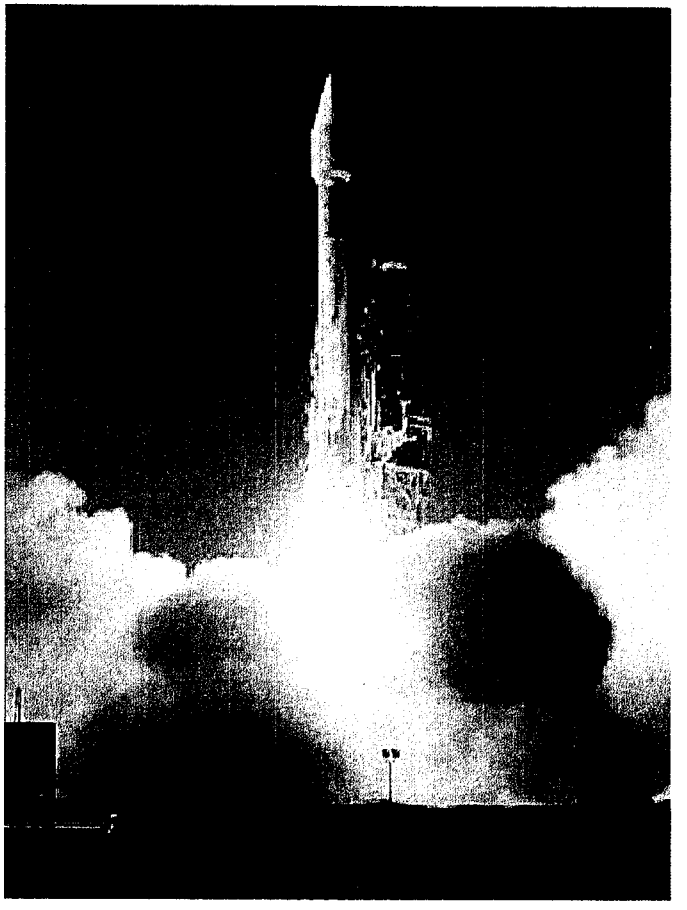


Amy Braverman
Jet Propulsion Laboratory
California Institute of Technology
Mail Stop 169-237
4800 Oak Grove Drive
Pasadena, CA 91109-8099
Amy.Braverman@jpl.nasa.gov

Larry DiGirolamo
Department of Atmospheric Science
University of Illinois
at Urbana-Champaign
105 S. Gregory St.
Urbana, IL 61801
larry@atmos.uiuc.edu

Outline

1. Motivation
2. Approach
3. Introductory example
4. Quantization
5. Algorithm
6. More examples
7. Summary / conclusions



Motivation

- Earth Observing System satellites return "massive" data volume.

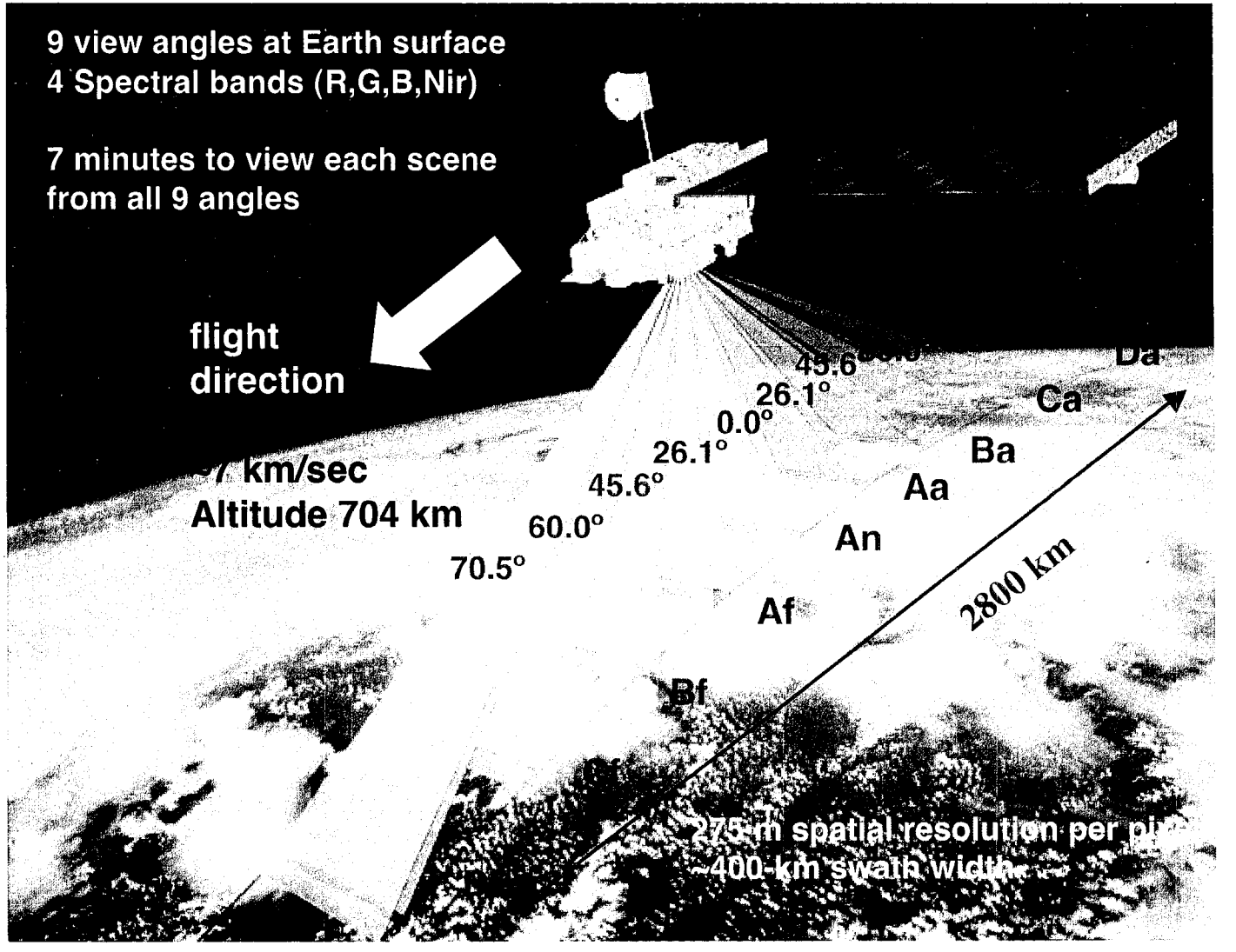
- Traditional approach to global data exploration: produce maps of one degree averages and standard deviations for each parameter of interest.

- Good news: this is easy, practical, and everybody understands it.

- Bad news: the method throws away almost all of the distributional information in the data including covariance and higher-order statistics.

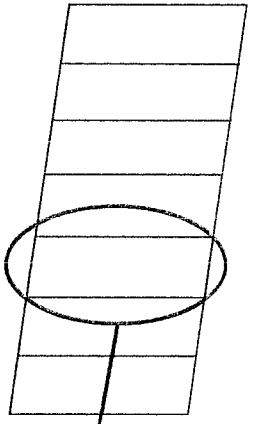
- **Need to produce lower volume proxy data sets suitable for exploratory data analysis; should be faithful to original data and parsimonious.**

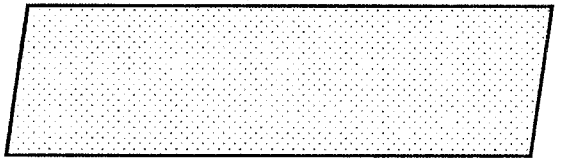
Motivation


9 view angles at Earth surface
4 Spectral bands (R,G,B,Nir)
7 minutes to view each scene
from all 9 angles
flight
direction
km/sec
Altitude 704 km
70.5°
60.0°
45.6°
26.1°
0.0°
26.1°
45.6°
Af
An
Aa
Ba
Ca
Da
Bf
2800 km
275 m spatial resolution per pi
400-km swath width

Motivation

180 blocks per swath

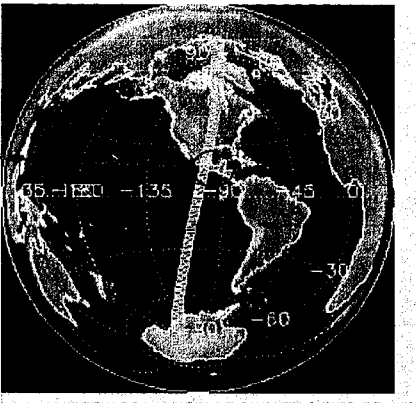




1 block= 128×512 1.1 km pixels
512×2048 275 m pixels

Level 1B2 product:

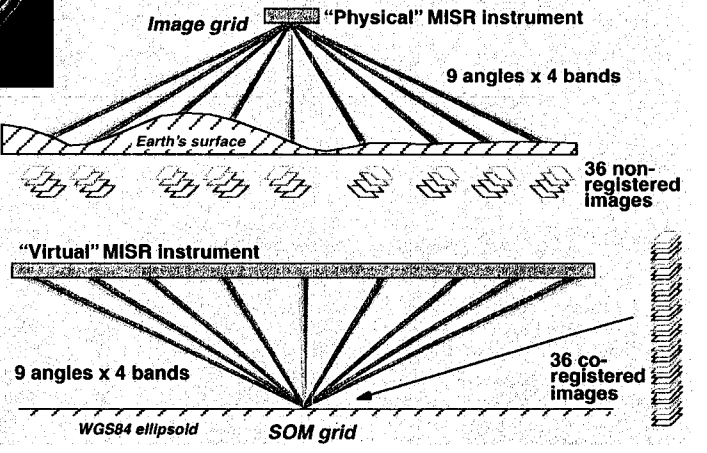
36 radiance measurements per pixel



SPACE OBLIQUE MERCATOR PROJECTION

Space Oblique Mercator projection

233 unique paths in 16-day repeat-cycle of Terra orbit



Level 2 derived geophysical products:

- Stereo product
- Albedo product
- Aerosol product
- Land surface product

Approach

- Old approach: bin data for a month into one degree grid cells, report means and standard deviations of parameters for each cell.

- New approach: bin data for a month into one degree grid cells, produce an estimate of the joint (empirical) probability distribution of parameters of interest for each grid cell.

How?

◆ Use a clustering algorithm to partition data for each grid cell into groups, represent each group by its centroid and (normalized) membership count. Collection of clusters and counts is a non-parametric estimate of the data distribution in that grid cell.

◆ Union of all 180 x 360 = 64,800 grid cell distribution estimates is a proxy for the original data. Approximates multivariate ("joint") statistical relationships in original data. Preserves spatial (temporal) relationships among grid cells.

◆ Similar to data squashing (DuMouchel, 2001).

Information example

- May 2003, three MISR cloud parameters observed simultaneously at 1.1 km resolution: red band albedo, stereo height, stereo-derived cloud mask.
- 54,600 populated, one degree grid cells.

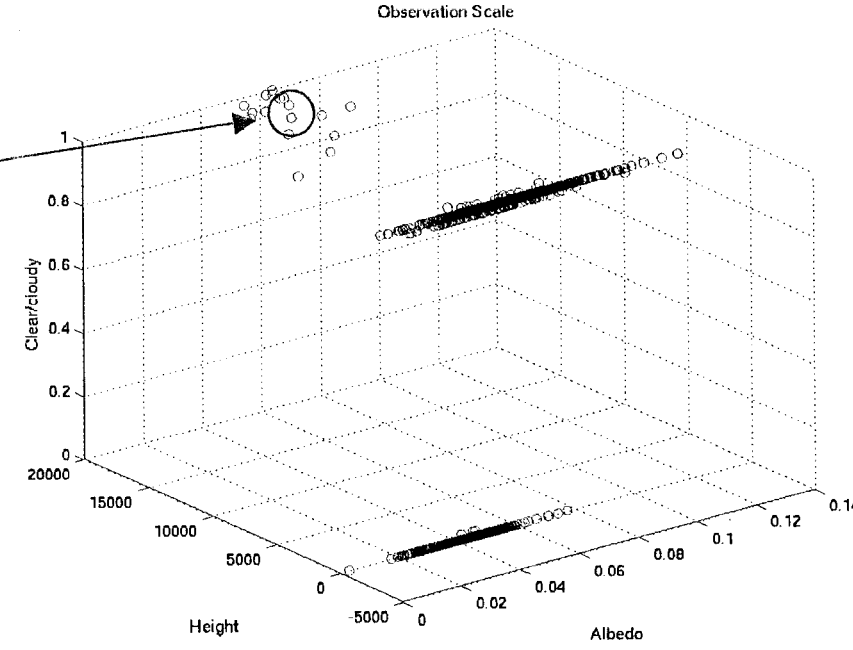
$$\mathbf{x}_n = (x_{n1}, x_{n2}, x_{n3})$$

distortion

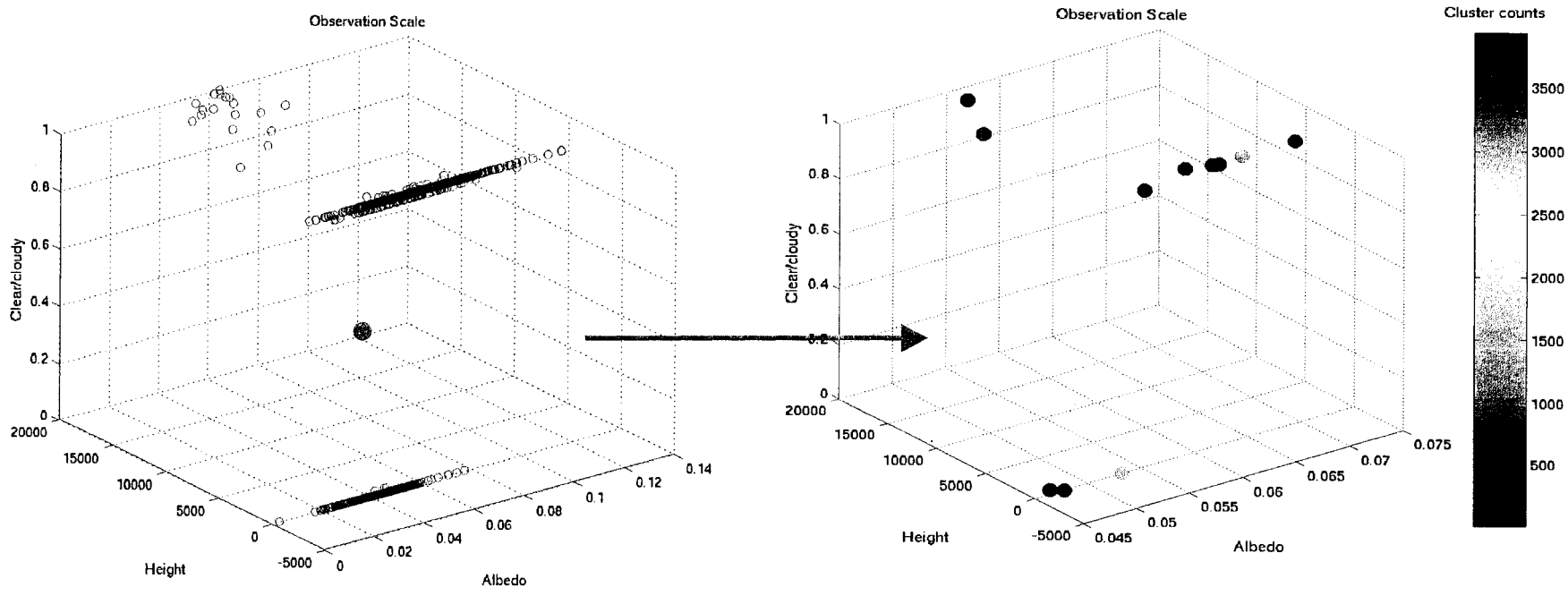
count

1.0049	-134.9029	0.047706	554	0	1	0
1.0042	-134.8931	0.047706	554	0	1	0
1.9106	-134.8741	0.056573	1110	1	1	0
1.9091	-134.8543	0.062676	1110	1	1	0
1.9083	-134.8445	0.062676	1110	1	1	0
1.9076	-134.8346	0.059605	1110	1	1	0
1.9068	-134.8248	0.059605	1110	1	1	0
1.9060	-134.8149	0.087852	1110	1	1	0

lat lon albedo stereo height cloud indicator

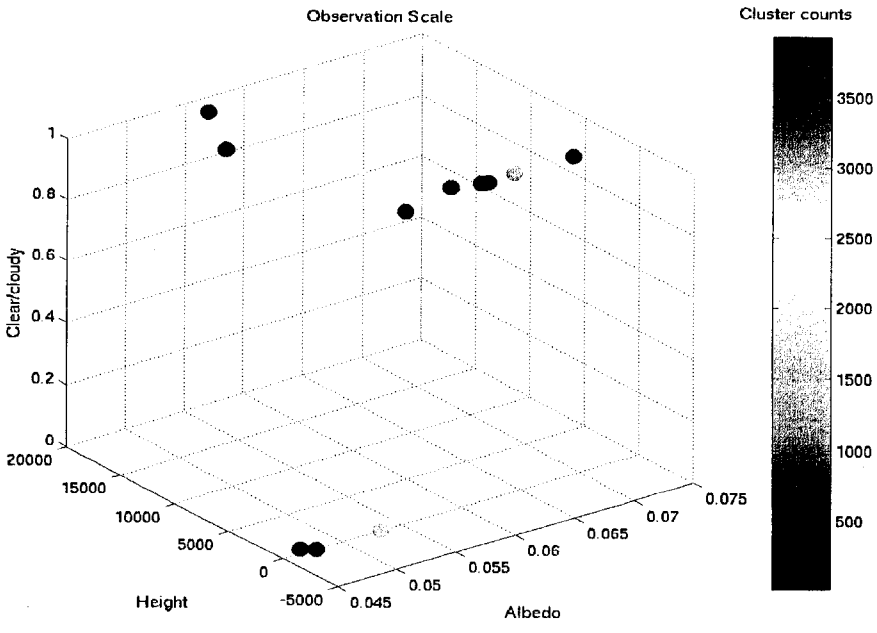


data for each spatio-temporal grid cell



- Union of all clusters for all grid cells is a compressed version of the original data. Global data set is about 40 MB.

- MISR Level 3 Quantization product:



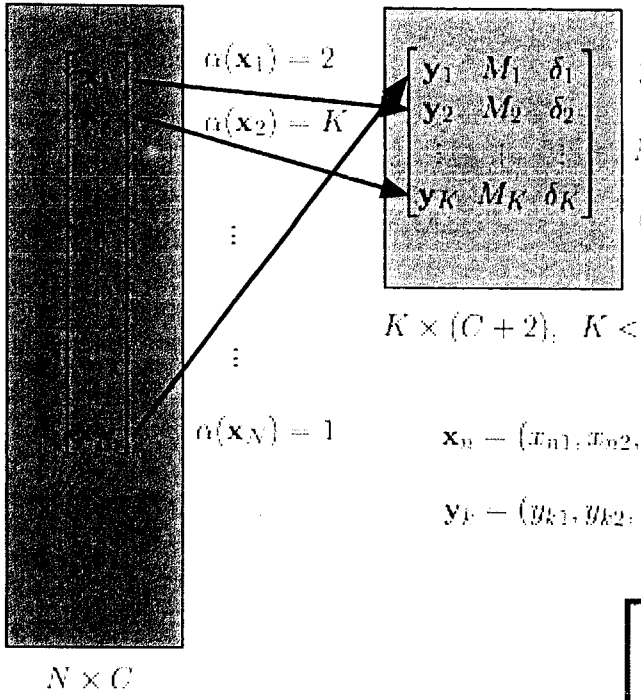
1	-135	0	0.0469255	554.000	0	4	0.000
1	-135	1	0.0477967	-7.810	0	3939	1.3788e-08
1	-135	2	0.0626818	1209.600	1	5	9.6774e-05
1	-135	3	0.0556817	528.713	1	80	9.6981e-05
1	-135	4	0.0532033	- 9.026	0	1525	2.2987e-09
1	-135	5	0.0525103	13480.000	1	6	0.4503107
1	-135	6	0.0631486	1025.386	1	101	4.6231e-06
1	-135	7	0.0628238	1044.123	1	212	2.9235e-07
1	-135	8	0.0703639	1050.800	1	133	1.8671e-07
1	-135	9	0.0654584	1109.720	1	1194	8.6598e-08
1	-135	10	0.0554375	18333.250	1	12	0.1011733
1	-135	11	0.0605958	1674.615	1	13	0.0029561

lat, lon, cluster id, albedo, stereo height, cloud indicator, count, distortion (standardized)

- Compute as if using original data, but answers are estimates of quantities computed from original data.
- How good? Depends on *both* distortion and the computation:

$$E\|g(X)-g(Y)\|^2 \approx E\left[(X-Y)\dot{g}(Y)'\dot{g}(Y)(X-Y)'\right]$$

Quantization



$$\mathbf{y}_k = \frac{1}{M_k} \sum_{n=1}^{N} \mathbf{x}_n 1[\alpha(\mathbf{x}_n) = k]$$

$$M_k = \sum_{n=1}^{N} 1[\alpha(\mathbf{x}) = k]$$

$$\delta_k = \frac{1}{M_k} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{y}_k\|^2 1[\alpha(\mathbf{x}_n) = k]$$

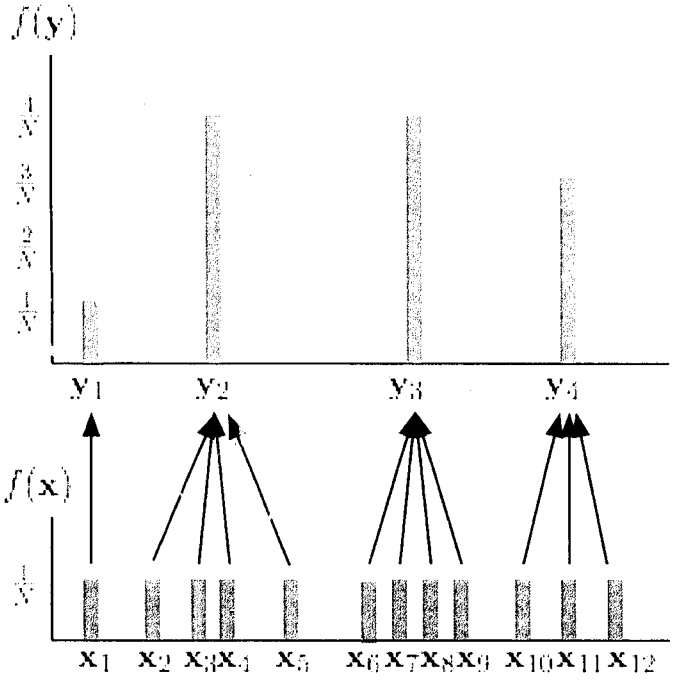
$$\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nC})'$$

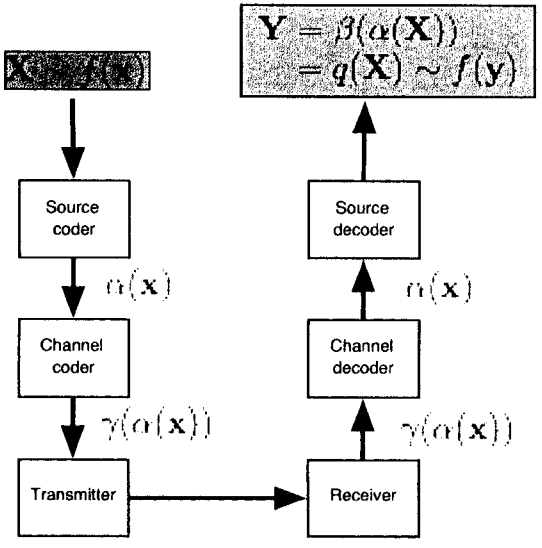
$$\mathbf{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kC})'$$

$$\mathbf{Y} = q(\mathbf{X}) = E(\mathbf{X} \mid \mathbf{Y})$$

$$\delta(\mathbf{X}, \mathbf{Y}) = E\|\mathbf{X} - \mathbf{Y}\|^2$$

$$h(\mathbf{Y}) = -E(\log f(\mathbf{Y}))$$





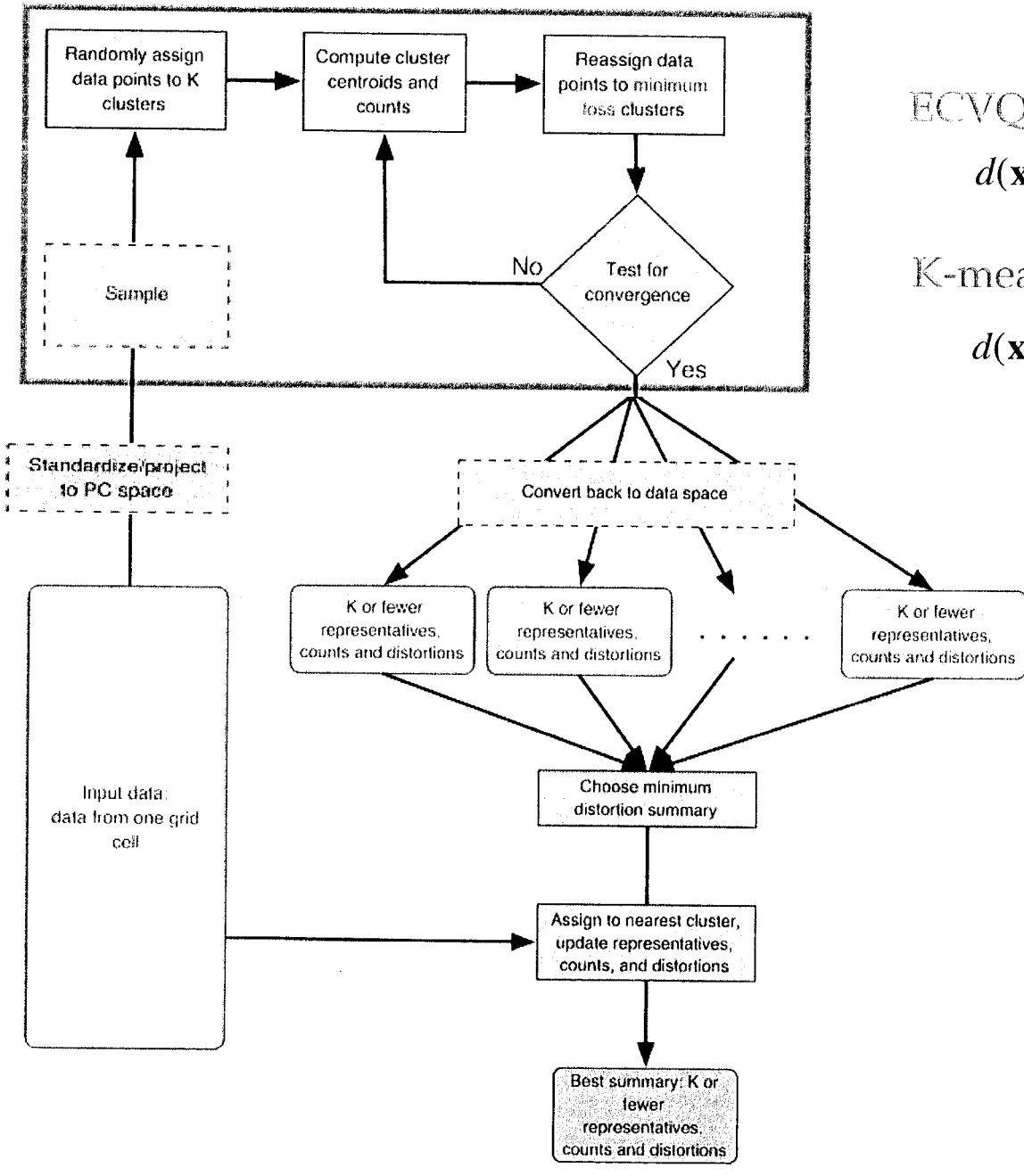
$\alpha(\mathbf{x})$ is the source coder:

$$\alpha(\mathbf{x}) = argmin_k \{d(\mathbf{x}, k)\}$$

$\beta(k)$ is the decoder: $\beta(k) = \mathbf{y}_k$

$\gamma(k)$ is the channel coder: $\gamma(k)| = -\log \frac{M_k}{N}$

Algorithm

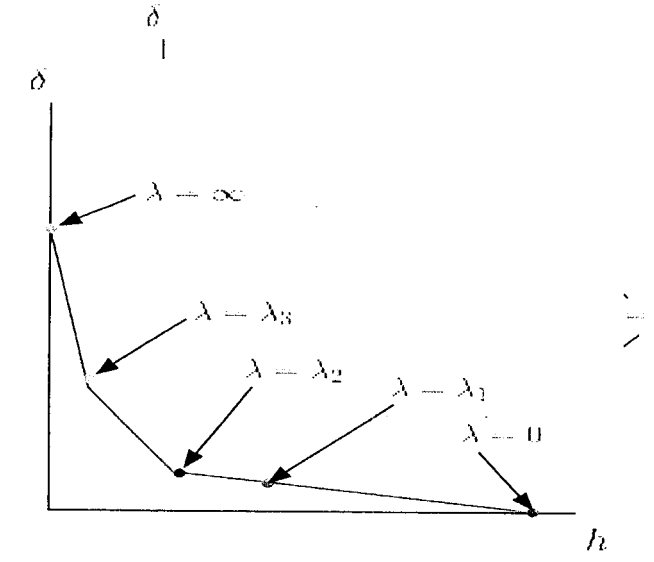


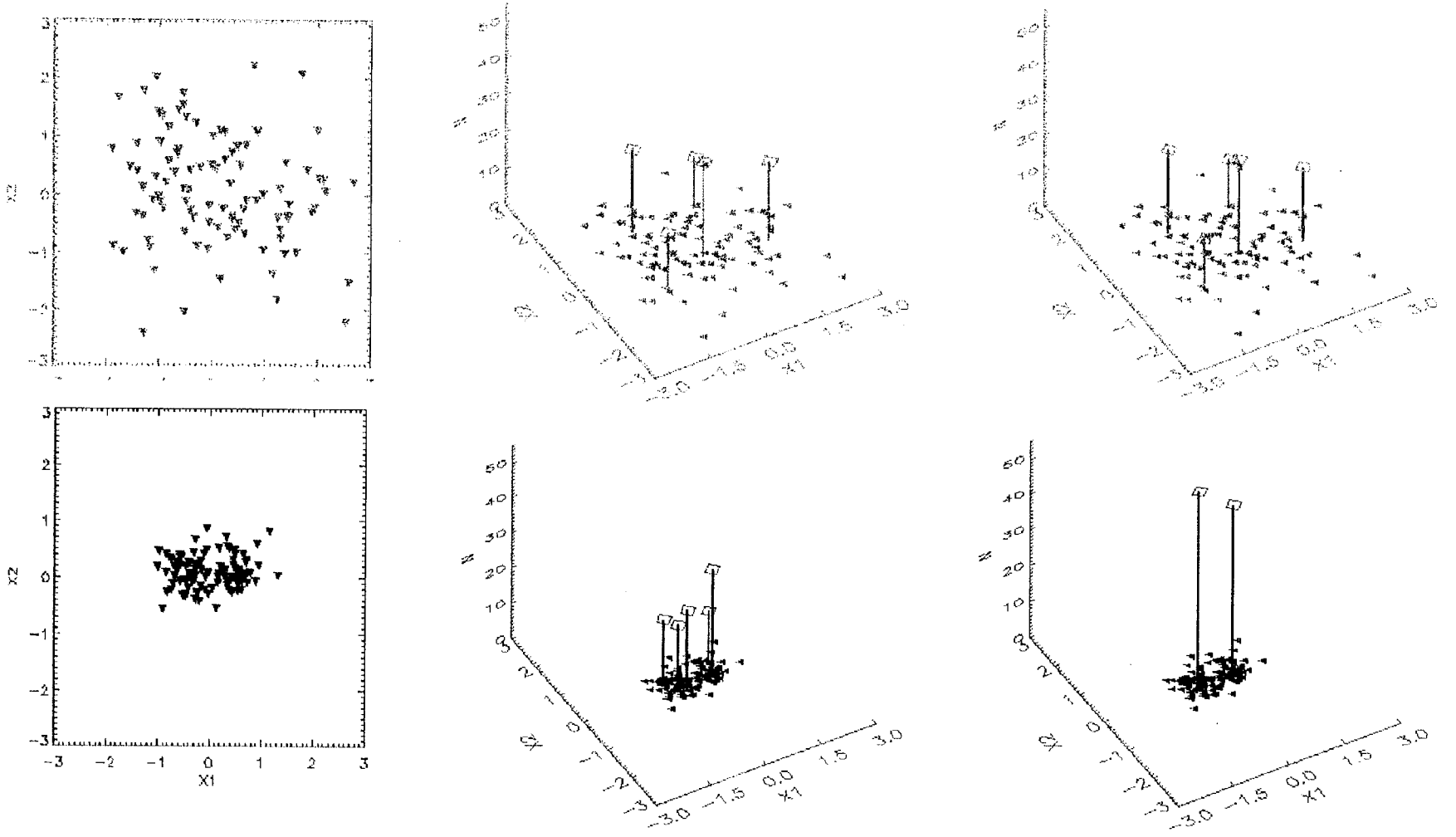
ECVQ loss:

$$d(\mathbf{x},k) = \|\mathbf{x} - \mathbf{y}_k\|^2 + \lambda\left(-\log \frac{N_k}{N}\right)$$

K-means loss:

$$d(\mathbf{x},k) = \|\mathbf{x} - \mathbf{y}_k\|^2$$





Choose λ to equalize δ .

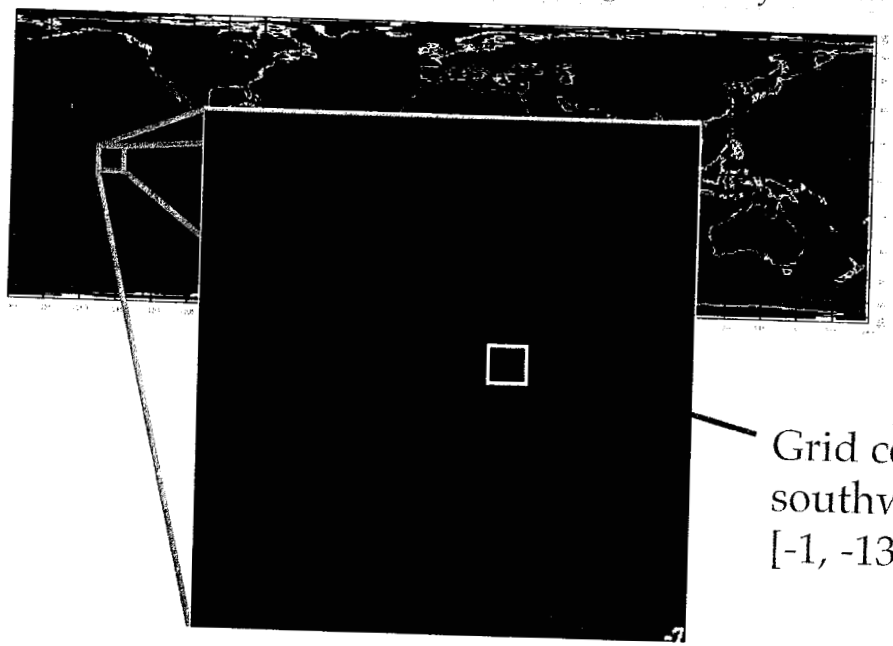
What's this good for?

- Compare climate model output to observational data:
  - Summarize large quantities of model output data;
  - Compare (estimates of) arbitrary non-linear statistics;
  - Compare whole distributions;
  - Construct formal hypothesis tests for consistency of model output and data.

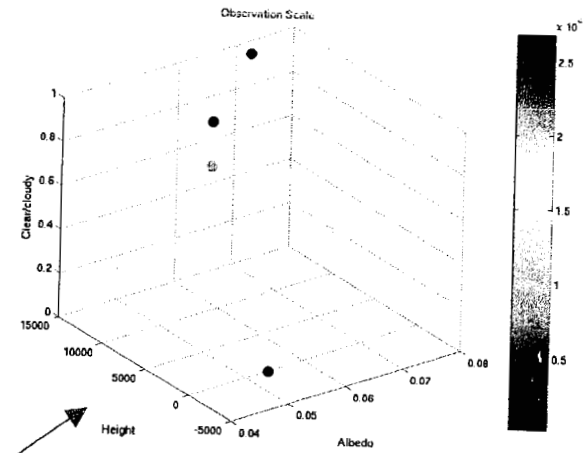
- Data mining/exploratory data analysis of observational and model data sets:
  - Understand/explore how multivariate distributions evolve in time and space;
  - Outlier/extreme event detection;
  - Find all occurrences of a given multivariate signature;
  - Characterize regions or time periods by probabilistic regime;
  - Explore multi-resolution (spatial, temporal, quantization) relationships.

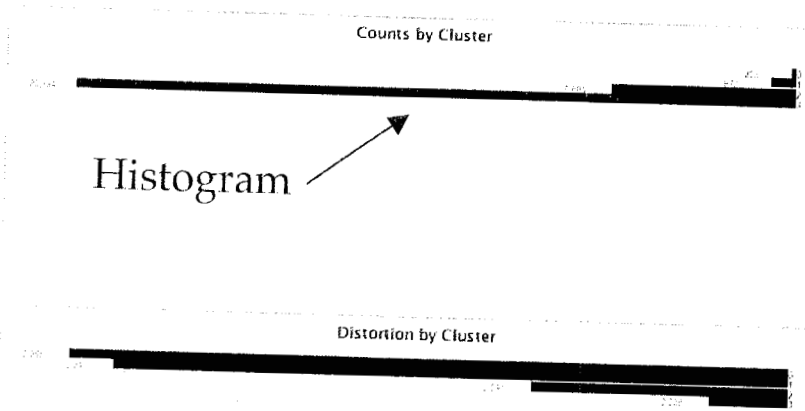
Example Data Visualization

Displayed: MISR May 2003 one-degree grid cloud fraction.



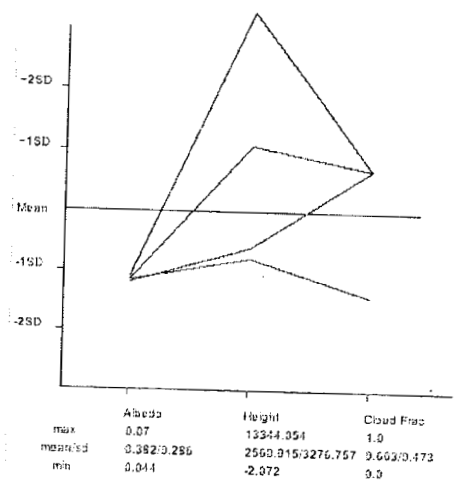
Grid cell with southwest corner [-1, -137]





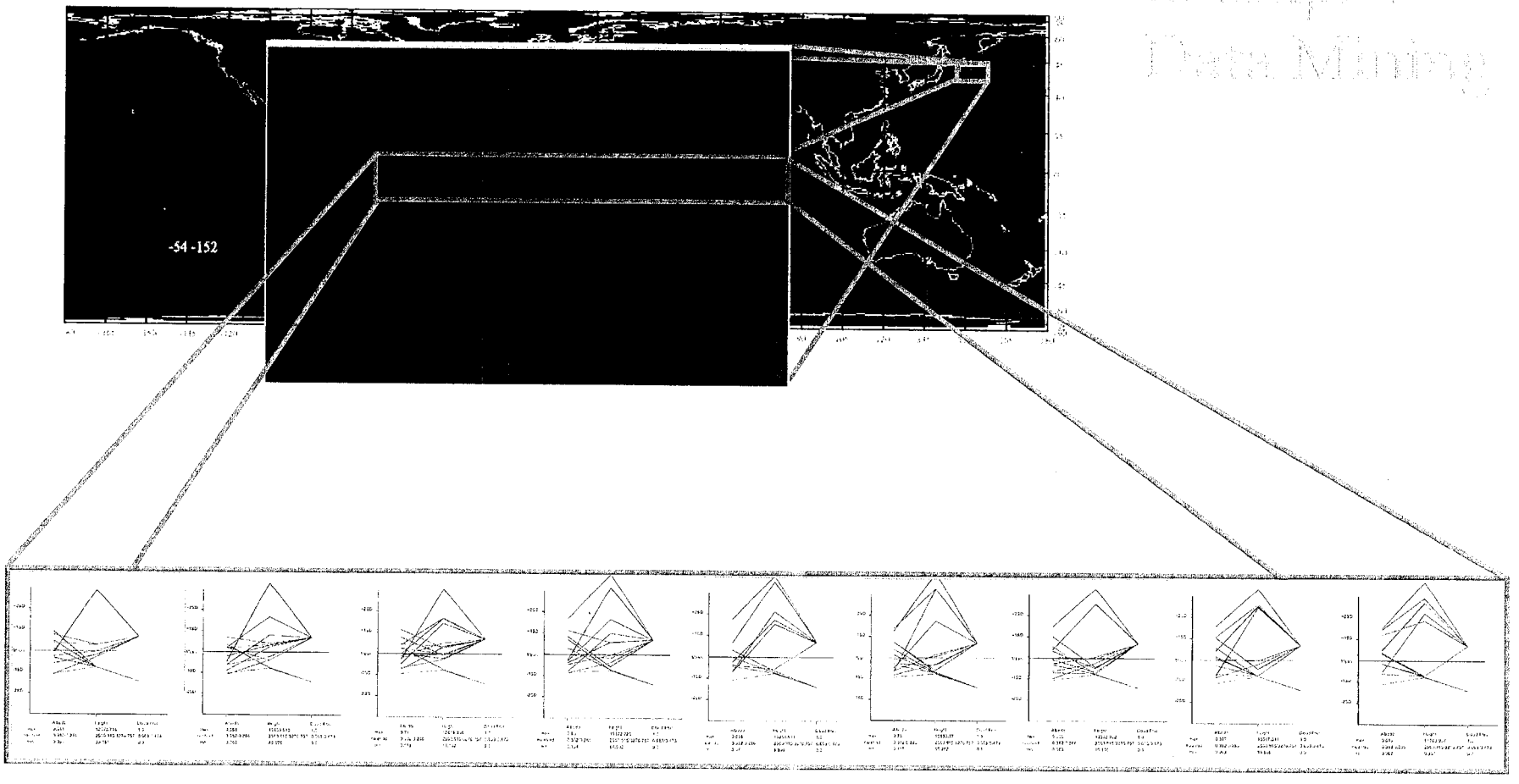
Histogram

Bar plot of within-cluster variation (distortion)



Parallel coordinate plot

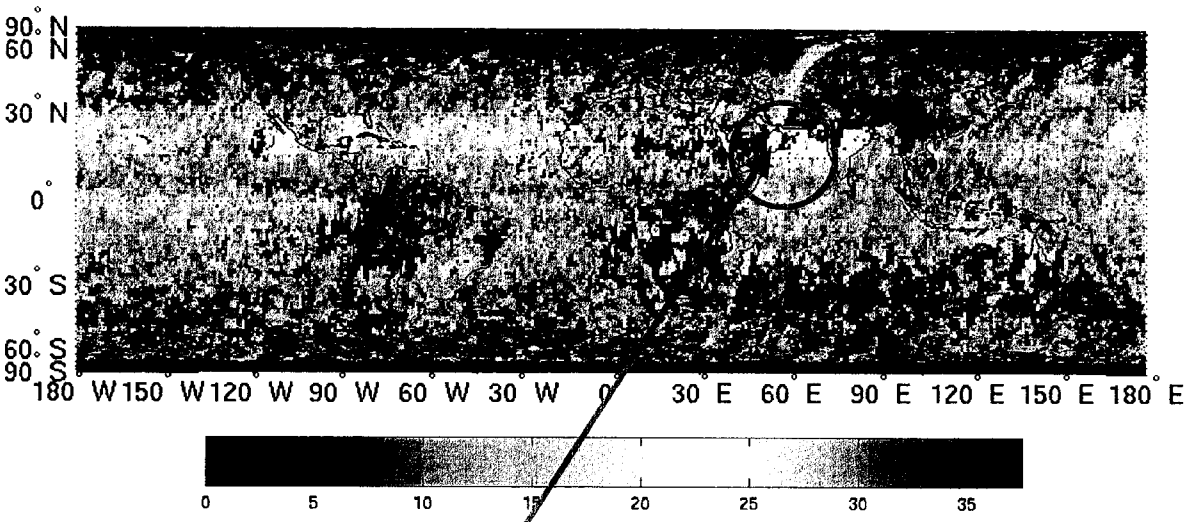
Example: Data Mining



Grid cells with southwest corner latitudes 40 degrees, longitudes 149 to 157.

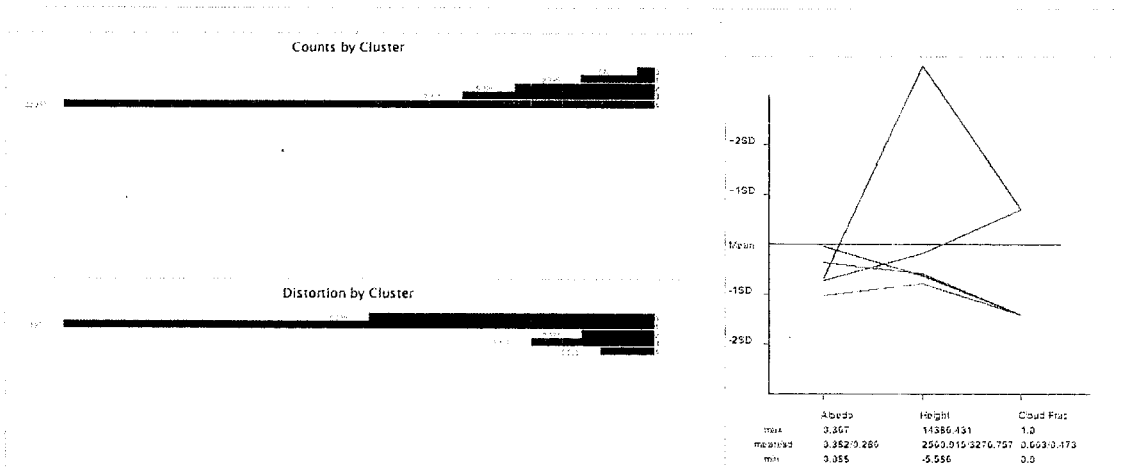
Can we model these relationships physically/statistically?

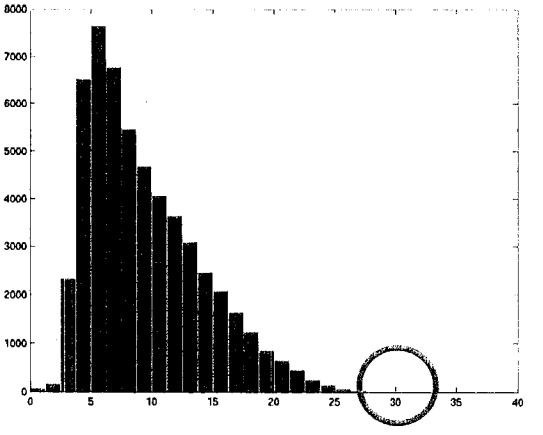
Example Outlier Detection



For each grid cell, calculate the maximum squared difference between cluster centroids and grid cell centroid.

Latitude 17, Longitude 55





Summary/Conclusions

- Accept coarser spatial/temporal resolution (one degree/monthly) to achieve replication and estimate empirical distributions.

- Calculations based on summaries approximate calculations based on original data. Mean squared errors (distortions) are known, and can be propagated.

- Meant for exploratory data analysis and comparisons of global data sets. *Not* a replacement for original data.

- **Wanted:** *collaborators to "test" and suggest modifications to the MISR Level 3 products. Pariticipate in climate model validation and climate research using MISR (and possibly other EOS) data.*

- MISR web site: `http://www-misr.jpl.nasa.gov`.