New Geometric Methods of Mixture Models for Interactive Visualization

Jia Li
Bruce Lindsay
Xiaolong (Luke) Zhang

The Pennsylvania State University
University Park

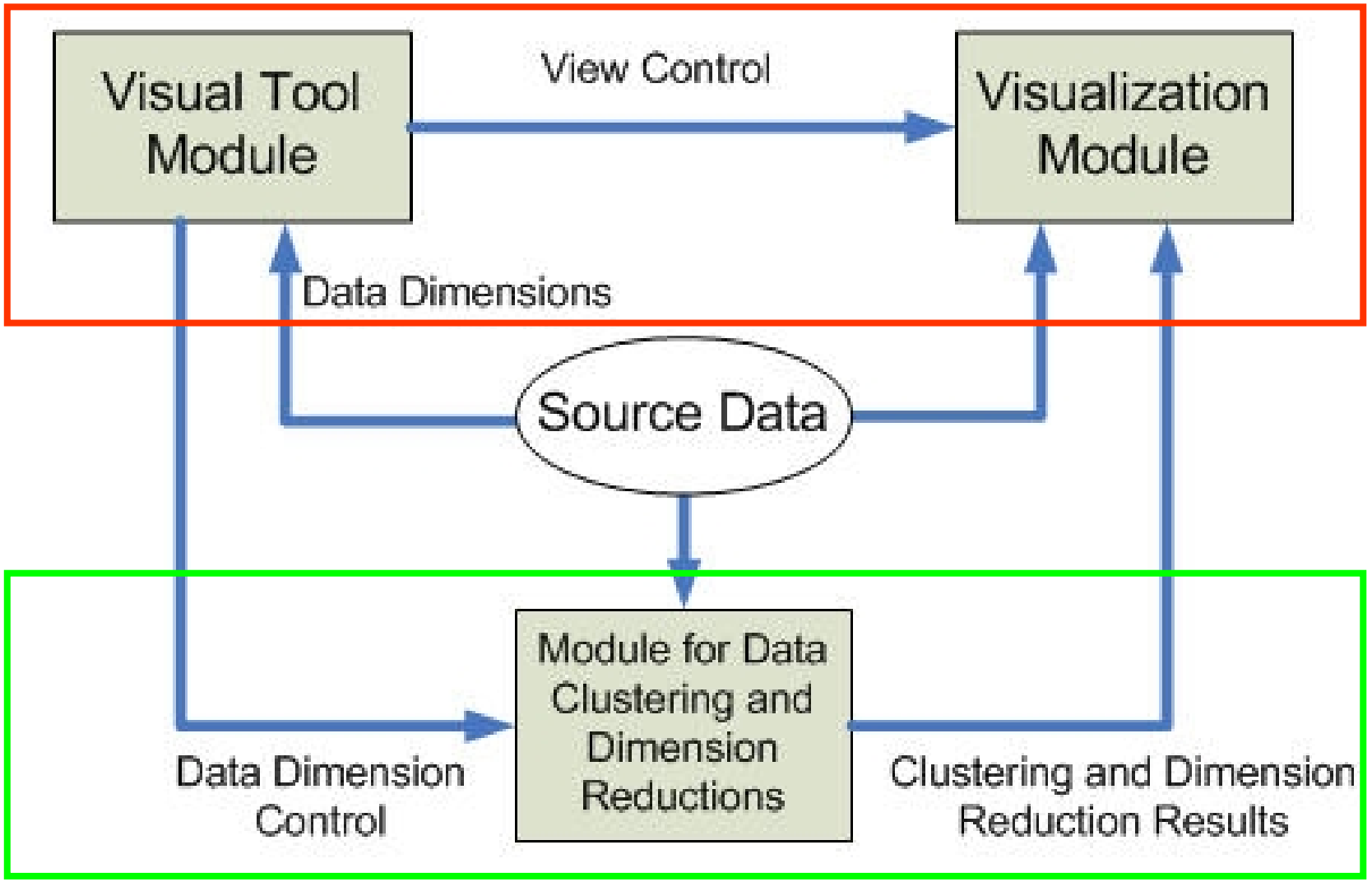
Outline

- Project overview
- Geometric methods of mixture models
  - Modal EM (MEM)
  - Hierarchical mode association clustering (HMAC)
- Application example: weather forecast

Project Overview

Big Picture

- Goal
  - Advance visualization and visualization with geometric methods of mixture models
- Approach
  - Developing theories and algorithms to uncover geometric features of mixture density
  - Developing tools for data clustering and dimension reduction
  - Developing visualization systems by integrating statistical tools to analyze large-scale, high-dimensional, and temporally evolving data.
    - Meteorology datasets, engineering design datasets
  - Understanding the use of our systems


Visual Tool Module
View Control
Visualization Module
Data Dimensions
Source Data
Module for Data Clustering and Dimension Reductions
Data Dimension Control
Clustering and Dimension Reduction Results

Research Team

- Researchers
  - Jia Li, Associate Professor of Statistics
  - Bruce Lindsay, Professor of Statistics
  - Xiaolong (Luke) Zhang, Assistant Professor of Information Sciences & Technology

- Partners
  - Fuqing Zhang, Professor of Meteorology and Statistics
  - Tim Simpson, Professor of Industrial & Manufacturing Engineering

Geometric Methods of Mixture Models

- For more details about the models
  - http://www.stat.psu.edu/~jiali/hmac/

Clustering by Mixture Models (Conventional)

- Cluster: a component in a mixture model

$$f_k(x) = \phi(x \mid \mu_k, \Sigma_k)$$

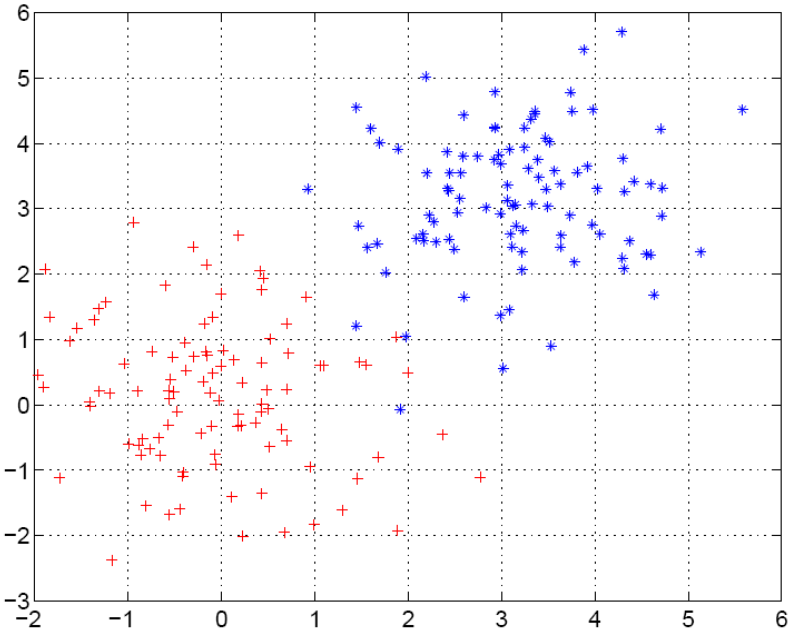
- Overall data

$$f(x) = \sum_{k=1}^{K} a_k f_k(x) = \sum_{k=1}^{K} a_k \phi(x \mid \mu_k, \Sigma_k)$$

An Example

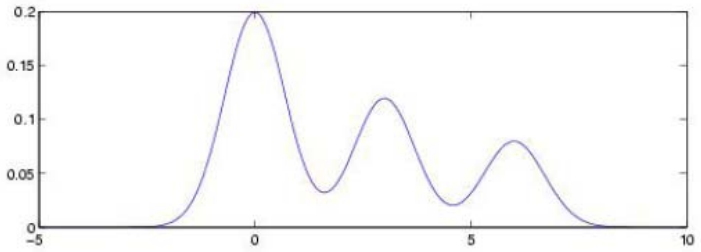
- Fit a two-component Gaussian mixture via the EM algorithm
- Cluster by MAP (Maximum A Posteriori)

$$\arg\max_k a_k f_k(x)$$

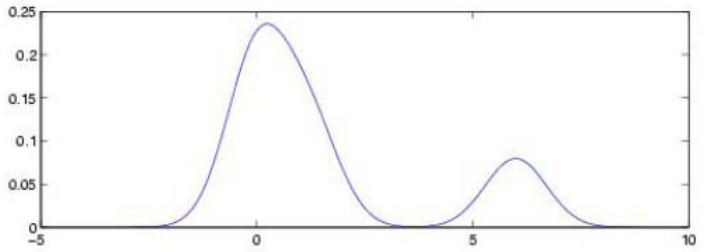


Heuristic Justification

- Challenge: Associating every cluster with a single mixture component may not be a good practice.



Good correspondence



Poor correspondence

Framework of Modal Clustering

- Form a density estimation by kernels or mixture models.
- Associate every point to a mode (local maximum).
  - Hill climbing to a mode without crossing ''valleys''.
- Group data according to common modes.

Compare the Roles of Mixture Models

- Conventional approach
  - Dual roles:
    - Density estimation
    - Each component captures one cluster
  - Potential issues:
    - Violation of the parametric distribution assumed for each component
    - Poorly separated components
- Modal clustering
  - Single role: Density estimation
    - Kernel density can be used (free of initialization)
  - Geometric heuristics
    - Every cluster is a hill (bump) of the density.

Model EM (MEM)

- Let a mixture density be $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$.
  - $x \in \mathcal{R}^d$
  - π_k is the prior probability of mixture component k.
  - $f_k(x)$ is the density of component k.

- Given any initial value $x^{(0)}$, MEM solves a local maximum of the mixture by alternating two steps.

Model EM

1. Let

$$p_k = \frac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}\,,\ k = 1, \ldots, K.$$

2. Update

$$x^{(r+1)} = \operatorname*{argmax}_x \sum_{k=1}^{K} p_k \log f_k(x)\,.$$

Mode Association Clustering (MAC)

- The MAC Algorithm

1. Form kernel density $f(x \mid S, \sigma^2) = \sum_{i=1}^{n} \frac{1}{n} \phi(x \mid x_i, D(\sigma^2))$, where $S = \{x_1, x_2, ..., x_n\}$.
2. Use $f(x|S, \sigma^2)$ as the density function. Use each x_i, $i = 1, 2, ..., n$, as the initial value in the MEM algorithm to find a mode of $f(x|S, \sigma^2)$. Let the mode identified by starting from x_i be $\mathcal{M}_\sigma(x_i)$.
3. Extract distinctive values from the set $\{\mathcal{M}_\sigma(x_i), i = 1, 2, ..., n\}$ to form a set G. Label the elements in G from 1 to $|G|$.
4. If $\mathcal{M}_\sigma(x_i)$ equals the kth element in G, x_i is put in the kth cluster.

Hierarchical Mode Association Clustering (HMAC)

- Gradually increase kernel bandwidth:

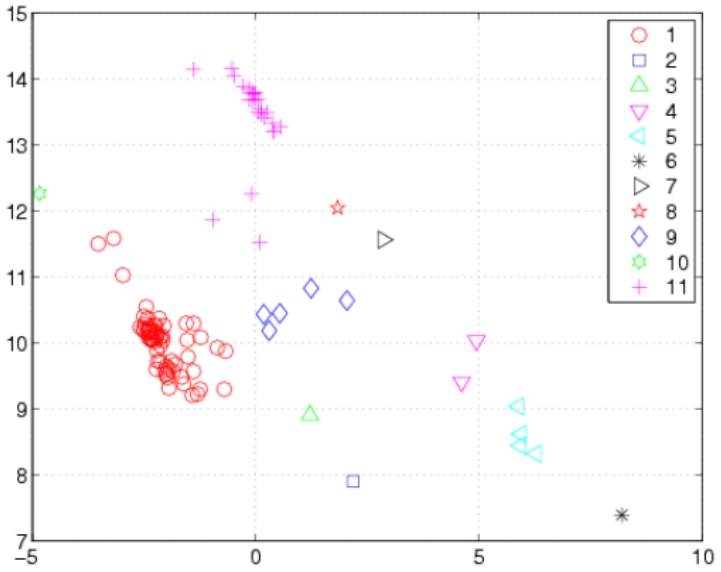
$$\sigma_1 < \sigma_2 < \sigma_3 \cdots$$

- Kernel density at level i: $f(x \mid S, \sigma_i^2)$
  - $\sigma_i \uparrow \longrightarrow$ smoother density, fewer modes
- Starting points at level *i* are the modes acquired at the previous level *i* - 1.
- The hierarchy by design:

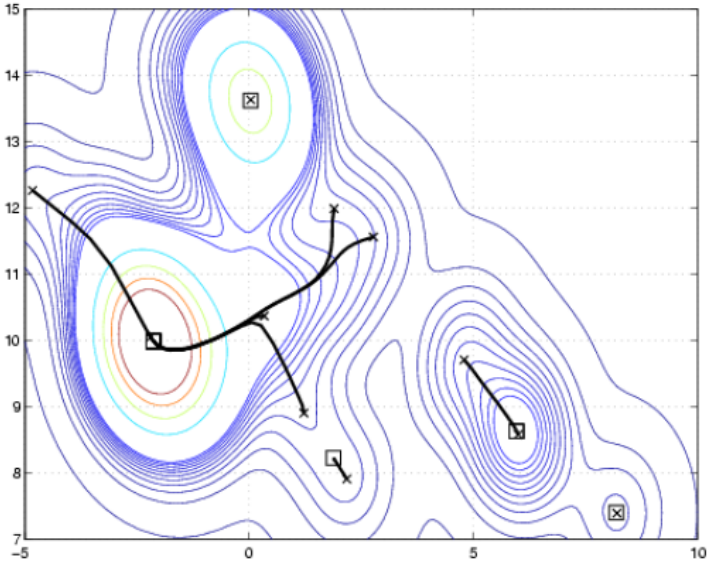
$$x_i \rightarrow \mathcal{M}_{\sigma_1}(x_i) \rightarrow \mathcal{M}_{\sigma_2}(\mathcal{M}_{\sigma_1}(x_i)) \rightarrow \cdots$$

An Example

- Glass identification data set
  - 105 samples
  - Only take the first two principal components of the original 9 dimensions

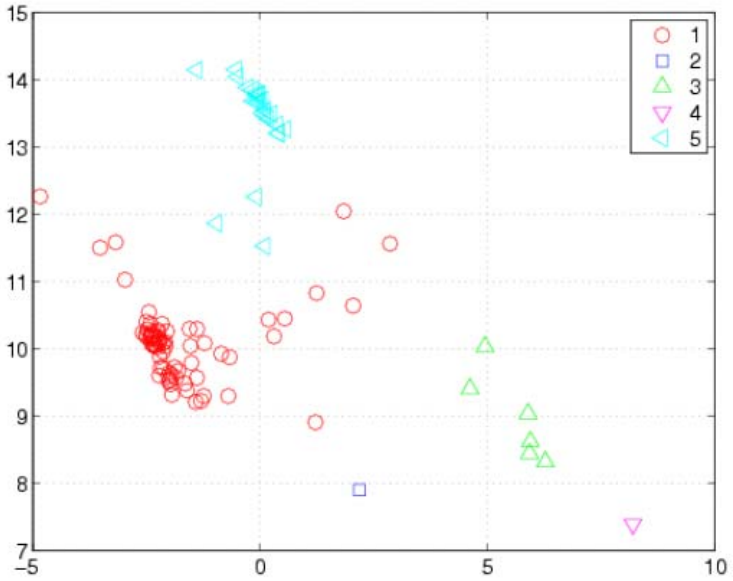


Clustering result at level 2

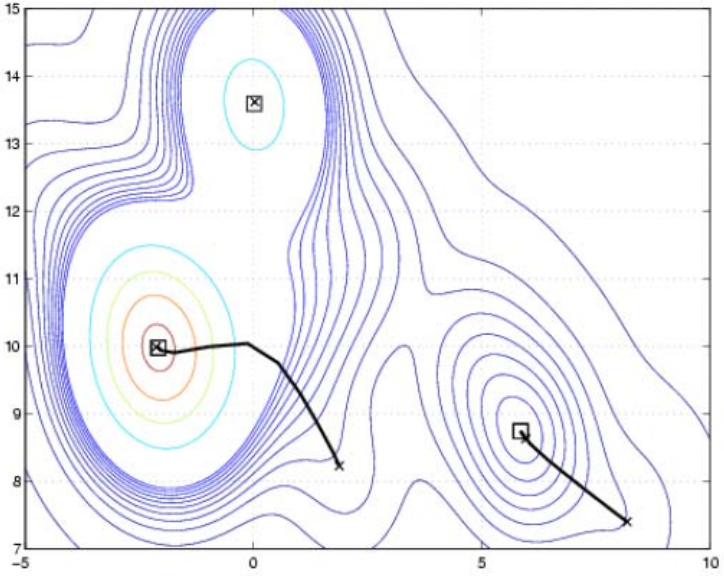


At level 3, merge the modes from level 2.

MEM paths are shown.



Clustering result at level 3



At level 4, merge the modes from level 3.

Hierarchical Mode Association Clustering (HMAC)

- Advantages of MAC and HMAC
  - MAC requires no model fitting and uses a nonparametric kernel density estimation.
  - The density of a cluster is not restricted to be parametric, for instance, Gaussian, but ensures uni-modality.
  - MAC is similar to mixture-model-based clustering in the sense of characterizing clusters by smooth densities.
  - HMAC seems to combine the complementary merits of bottom-up clustering such as linkage and top-down clustering such as mixture modeling and k-means.
  - MAC can be performed hierarchically or in a one level manner.
  - For hierarchical clustering, HMAC can either generate a nested structure as a conventional dendrogram, or a non-nested hierarchy.
  - Ridgelines between pairs of clusters can be computed to numerically profile the separation between clusters, a useful feature for diagnosis.

Application Example: Weather Forecast

Weather Forecast

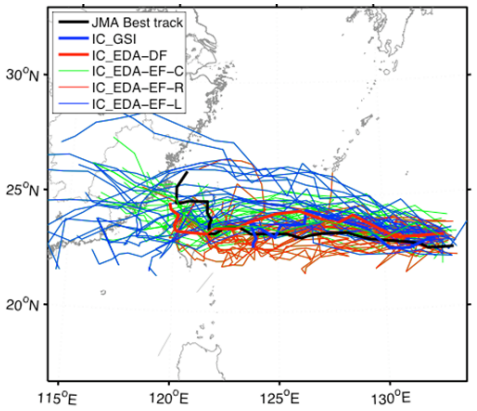
- Data
  - Multi-dimensional prognostic variables
    - Dry air: the mass, moving velocities (x, y and z), potential temperature, …
    - Geopotential and mixing ratios for cloud, rain, ice, water vapor, and snow
    - .
- Human analytic activities
  - Satellite images, forecasting models, 3D views, spreadsheet, …
- Challenges in interacting with data and models and making quick decisions about what may happen next.

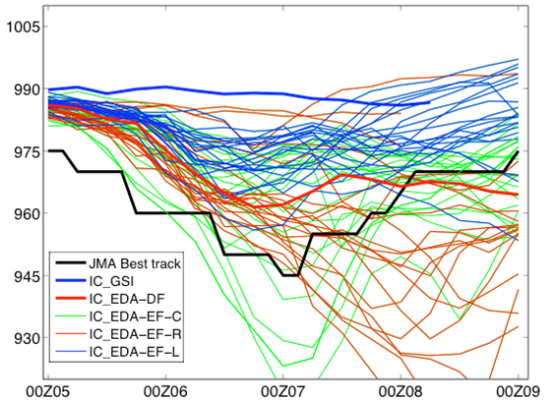
After Typhoon Morakot

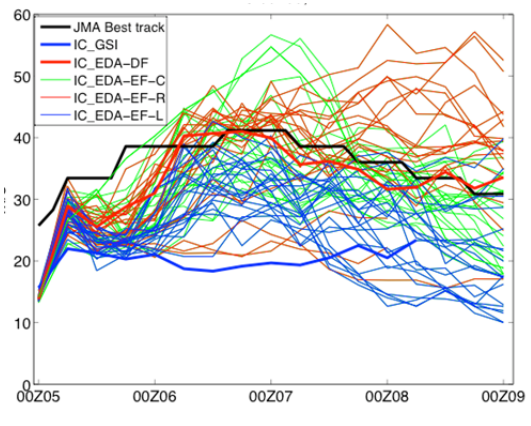
From boston.com

Task 1: Ensemble-Based Analysis and Forecast

- An ensemble
  - A collection of forecasting models
- Ex. Typhoon Morakot







Challenges

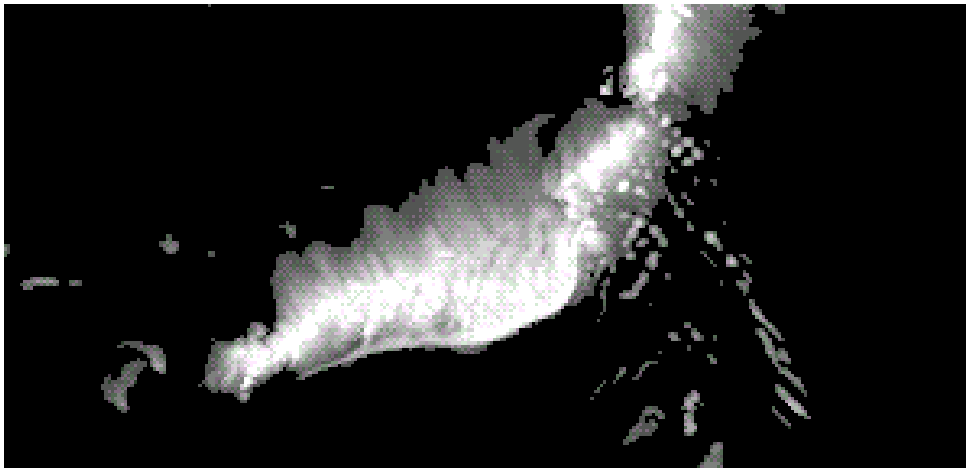
- Analyze 60 models
  - Examining the variations of the resulting forecasts from different models
  - Estimating the uncertainty of the prediction.
  - Delivering the best forecast
- Project
  - Where will the Typhoon go and how severe will it be?

Our Approach

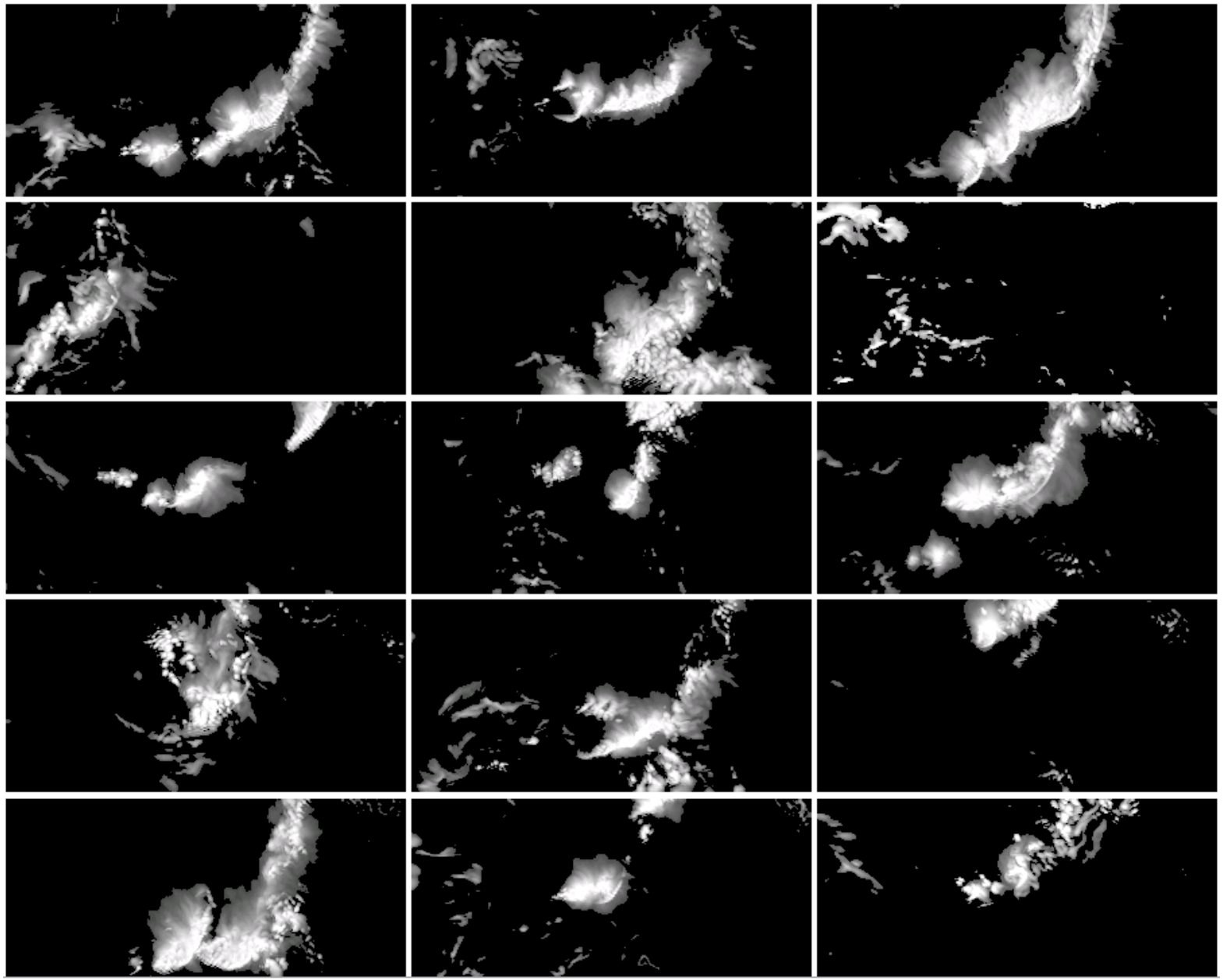
- Clustering model data
  - Different clustering criteria
    - Path + sea level pressure + wind strength
    - Sea level pressure + wind strength
    - Path only
    - Sea level pressure only
    - Wind strength only
- Interactive tools
  - Choose different clustering methods
  - Examining the modes and ranges of model clusters
  - Examining the temporal evolution of model clusters

Demo: Multi-Scale Clustering

Task 2: Understanding Hurricane Cloud Maps



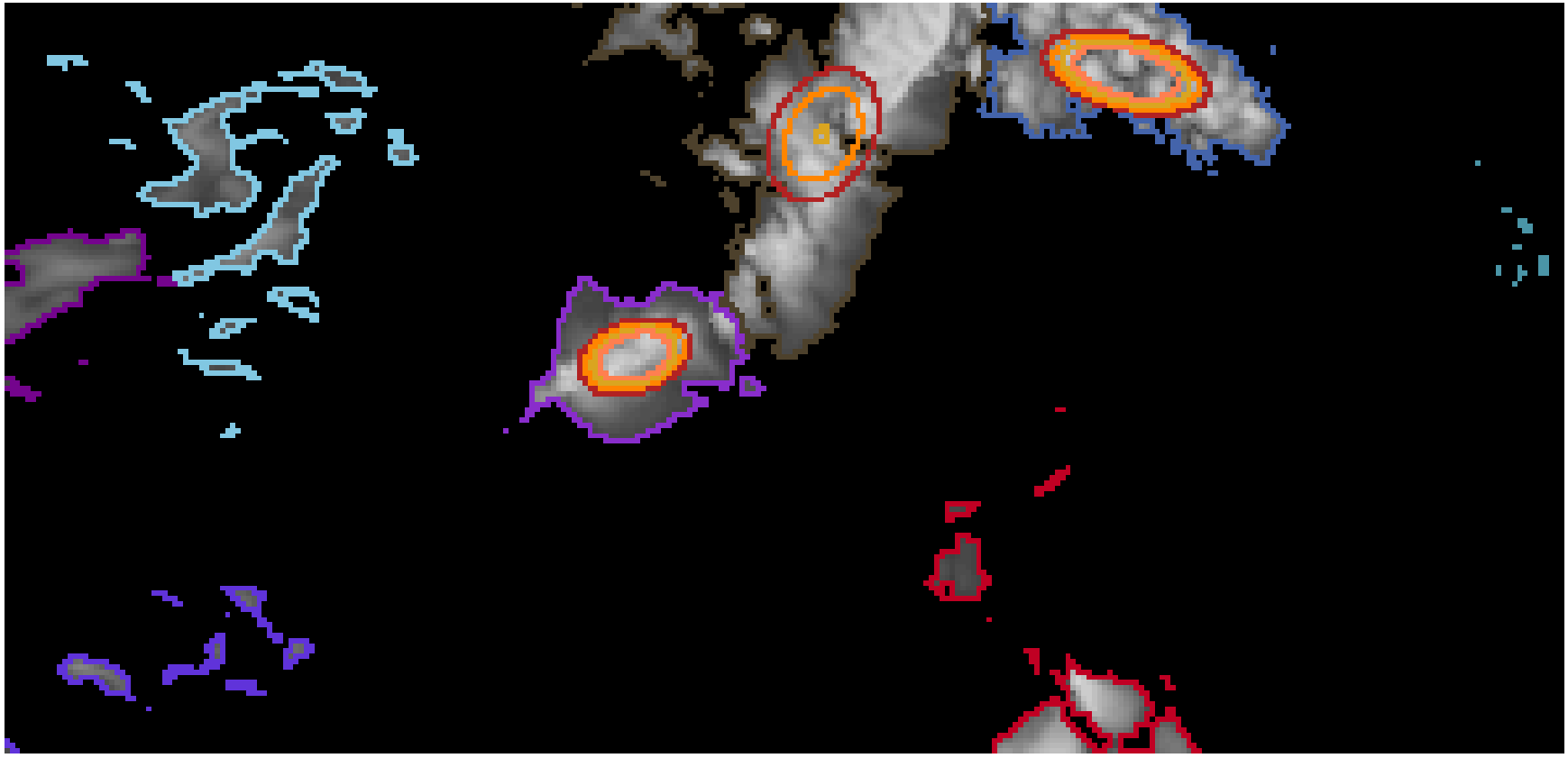
- Questions to answer
  - How many cloud clumps are there?
  - What are their trends?

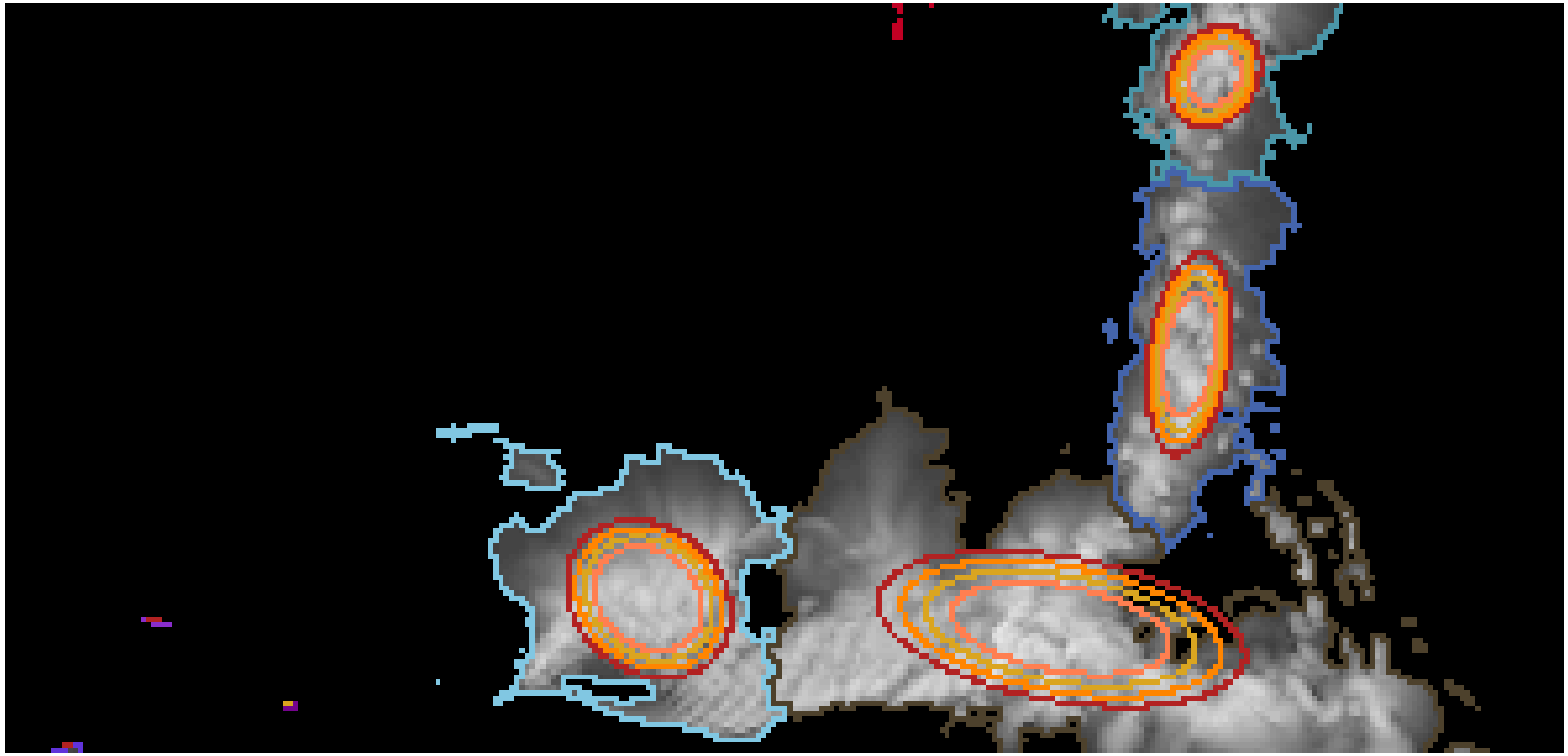


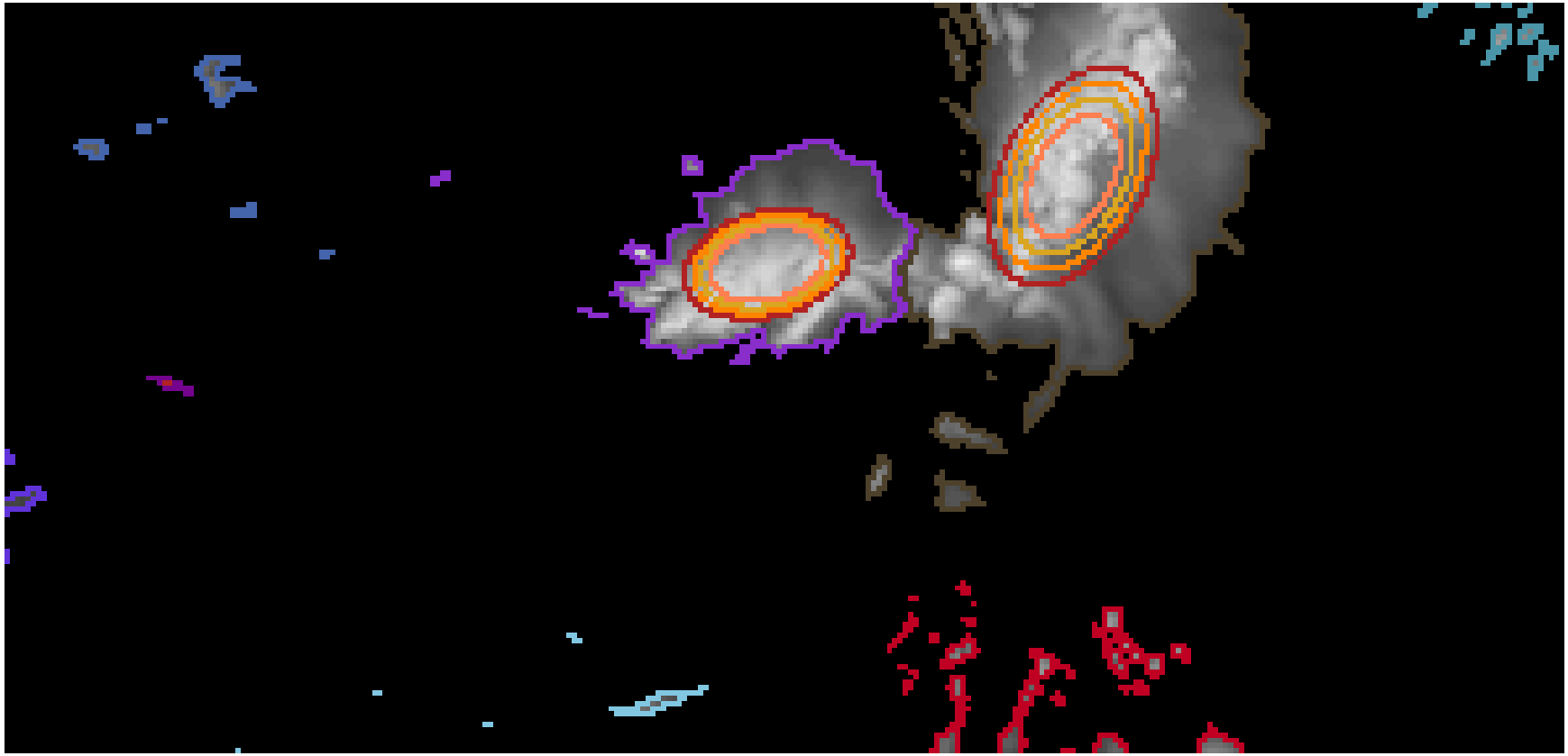
Our Approach

- Clustering cloud image pixels with MAC
  - Location + brightness

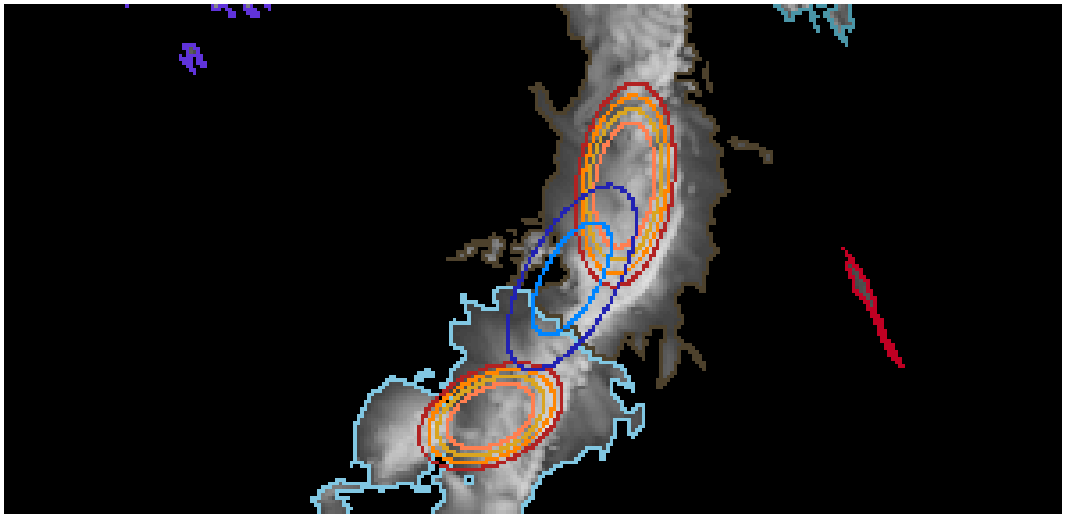
Some Results

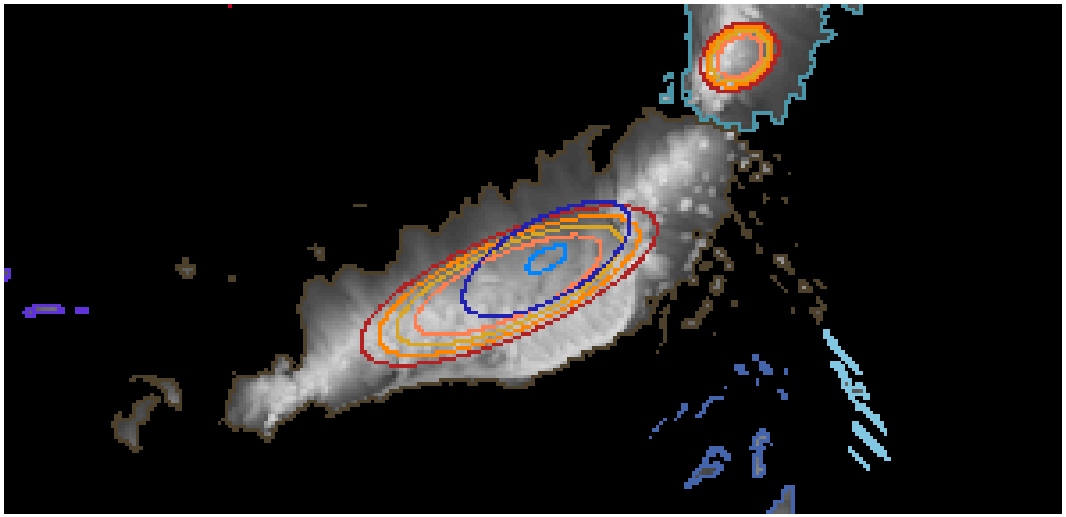


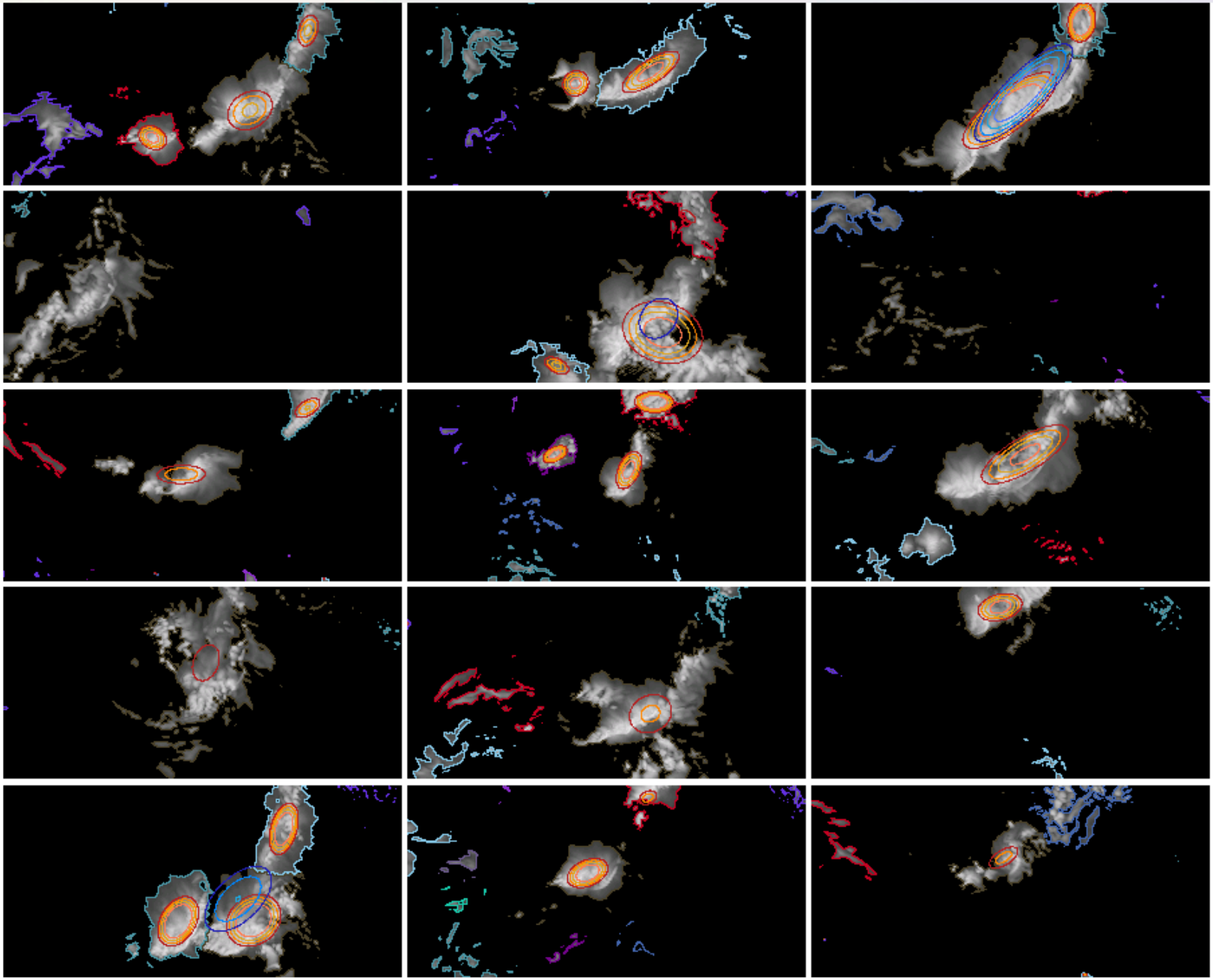




Comparison between Our Method and Single Gaussian Modeling





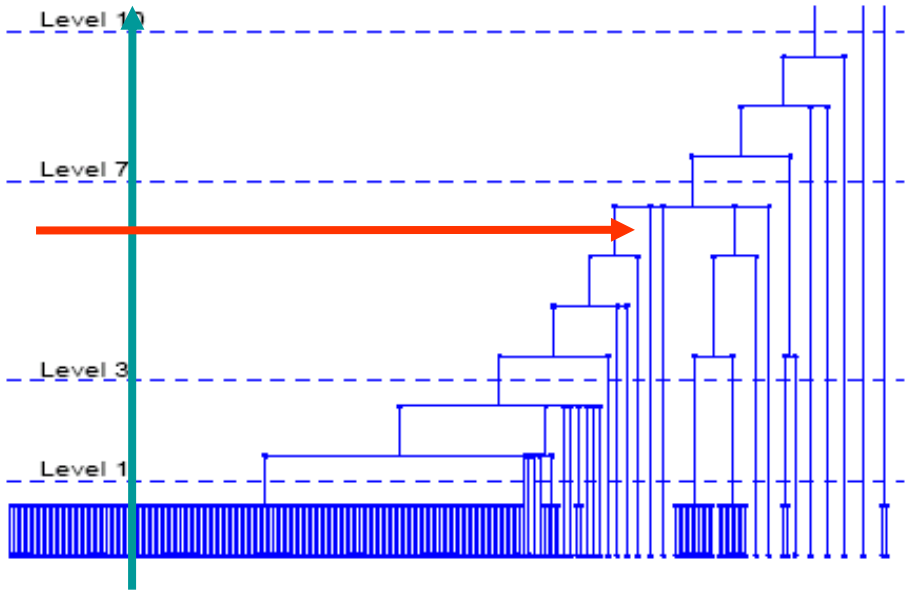


Preliminary Feedback

- Promising results
- Eager to see an interactive system to put various tools together

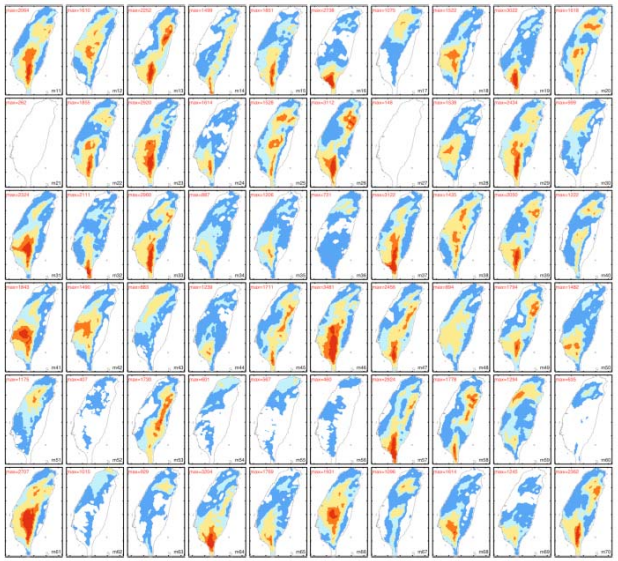
Next Step

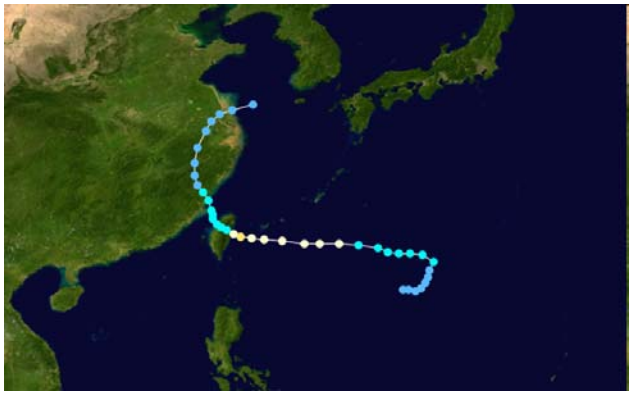
- Improving algorithms
  - Mixture modeling for density-estimation of high dimensional data
  - Scale-sensitive vs. scale-independent
  - Speed of clustering
    - Real-time analysis



Next Step

- Developing visual tools and testing how helpful our approach is
  - Typhoon Morakot data
    - Which typhoon models are similar?
    - What areas are most likely to be flooded
    - .
  - Cloud maps
    - Where will a cloud clump look like in 5 hours?
    - .





For More Information

- http://gmmv.ist.psu.edu/