

A Clustering Algorithm for the Automated Storm Identification of Space-Based Optical Lightning Data

A. Mielke¹, S. Davis², D. Suszcynsky², A. Jacobson²

¹*Los Alamos National Laboratory, Safegaurds Systems, MS E541, Los Alamos, NM. 87545* ²*Los Alamos National Laboratory, Space and Atmospheric Sciences Group, MS D466, Los Alamos, NM 87545*



ABSTRACT

The Fast On-Orbit Recording of Transient Events (FORTE) satellite is a joint Los Alamos National Laboratory (LANL) and Sandia National Laboratories (SNL) experiment that was launched into a nearly circular low-earth orbit on August 29, 1997. The payload consists of broadband Very High Frequency (VHF) receivers and a two-sensor Optical Lightning System (OLS). One of the OLS sensors, the Lightning Location System (LLS), is a narrow band ($777.6 \text{ nm} \pm 0.5 \text{ nm}$) 128×128 pixel charge coupled device (CCD) array that is autonomously triggered, and provides imaging and geolocation of lightning events to within a pixel size of $10 \text{ km} \times 10 \text{ km}$. This paper presents a data-clustering algorithm which uses FORTE LLS event locations to both (a.) discriminate between lightning and energetic-particle/glint events and (b.) identify regions of high event density that are associated with storm activity. In addition to the utilization of basic statistical and data-clustering techniques, data driven thresholds are employed in the identification of probable storm regions. The application of automatic data discovery and analysis techniques allows for an efficient, flash/storm-level analysis of the more than 30 million events recorded by FORTE LLS, including a statistical characterization of seasonal, diurnal, and geographical variations in lightning and storm activity.

FORTE: Fast On-orbit Recording of Transient Events

MISSION

- Testbed for Next Generation Nuclear EMP Sensor Technology.
- Space-based Lightning Detection.

PLATFORM

Altitude: ~ 825 km
Inclination: 70 degrees
Launched: August 29, 1997

SENSORS

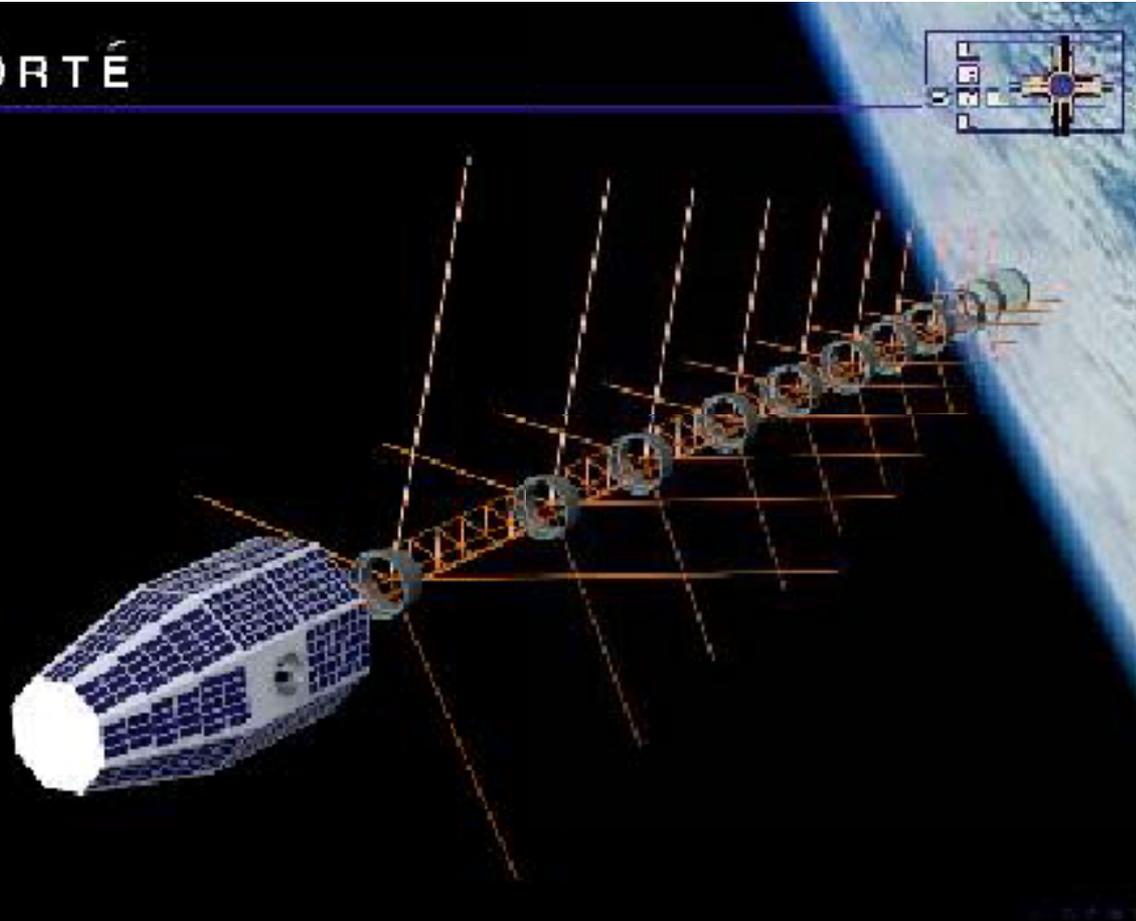
Type: Broadband VHF receivers
- (26 – 300 MHz)
- 1 μ s or better resolution

Photodiode (PDD)
- 15 μ S resolution

CCD Imager (LLS)
- 10 km location accuracy

Data: Optical/VHF Waveforms
Event times
Event location

FORTE



FORTE Lightning Location System

- **Type:** CCD array
- **Effective Array Size:** 128 x 128 pixels
- **Field-of-view:** ~1200 km diameter (80 deg.)
- **Spectral Response:** 10 Å FWHM filter centered on 777.4 nm
- **Geolocation accuracy:** 1 pixel = roughly 10 km x 10 km
- **Maximum Pixel Rate:** 400 pixels/sec
- **Integration Time:** 2.38 ms
- **Modes:**
 - Pixel event mode, threshold triggered
 - Full frame mode
 - Background event mode
- **Anti-glint feature:** If a particular pixel lights up for more than 2 consecutive sample periods, the event is ignored.

Motivations/Study Goals

- The purpose of this study is to improve upon the filtering of data from the FORTE LLS through the use of a data clustering algorithm.
- The more than 30 million events detected by the FORTE LLS over the past 3.5 years provide a unique data set with which to study seasonal and geographical differences in lightning activity. Due to the large number of events detected, any such statistical study needs to be automated.
- We also desire a means by which to automatically “discover” probable regions of storm activity for storm-level studies with other FORTE instruments
- The FORTE LLS can be used in corroboration with other FORTE instruments to provide event locations for events that otherwise would have unknown locations.

Previous Clustering Efforts

- Former clustering algorithm identifies any LLS event as “clustered” if there are any other LLS events within x kilometers of it.
- Typical values of x are 30 - 50 km

Limitations of Previous Clustering Efforts

- Speed: For each LLS event, the distance to every other event must be calculated.

Number of distance calculations $\propto n^2$

- In the new algorithm, the center-most event is used as the basis for distance measurements

Number of distance calculations $\propto n$

- The old algorithm does not identify events as being associated with a cluster. Therefore, there is no possibility of identifying storms with the algorithm

Assumptions of Clustering Algorithm

- For a pass of FORTE over any individual storm, the noise background is constant
- Storm sizes are 30 - 100 km
- Random noise may be assumed to be Poisson distributed

Background: The Poisson Distribution

- Distribution models random events - “Randomly throwing darts at a dartboard”
- The mean, $\mu = \text{total \# events} / \text{\# possible outcomes}$
- The variance, $\sigma^2 = \mu$

Method

1. Select LLS data for a single FORTE overpass of a specific geographical region (see Figure 1)
2. Remove any data with signal levels characteristic of glint. Glint is characterized by abnormally high signal levels and pixel saturation.

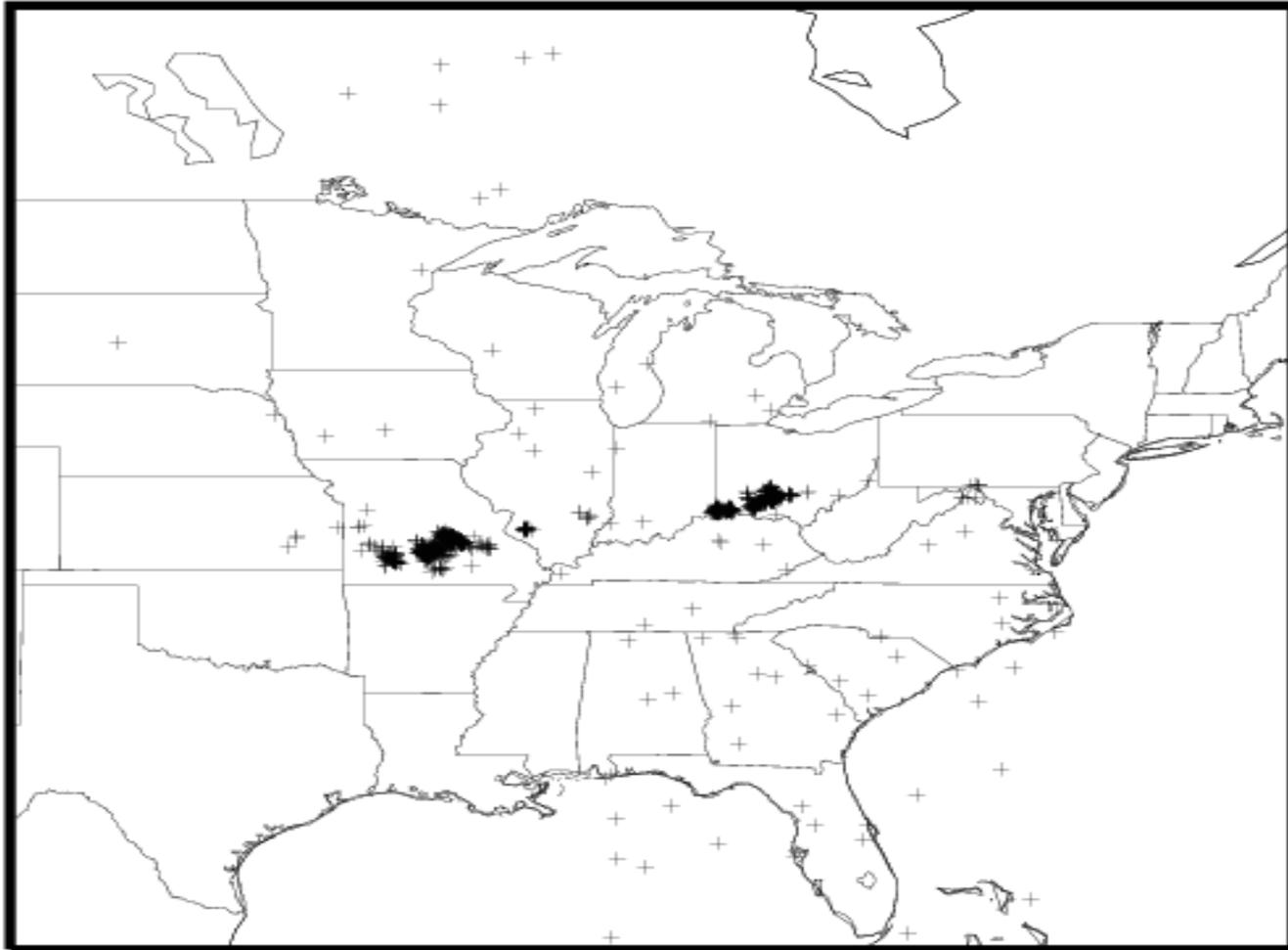


Figure 1: Plotted as black plus signs are 1280 FORTE LLS locations of for one overpass of the satellite. Event times are from 8/1/99, 05:27:16 - 05:38:34 UT.

3. Convert latitudes/longitudes to kilometers from center event.
4. Bin the data into 100-km square bins

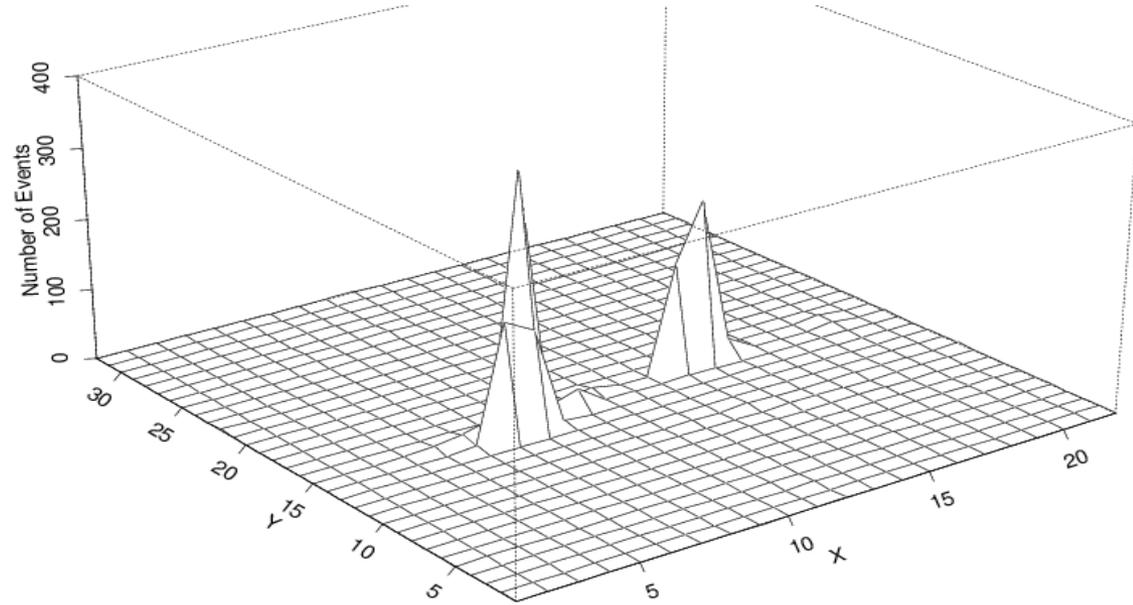


Figure 2: The data from Figure 1 binned into 40-km square bins, with x and y distances relative to an arbitrary point.

5. Choose a threshold. The threshold is chosen by computing the Poisson mean and standard deviation, then choosing a threshold a certain number of σ 's above μ . In this example, the threshold is set to $\mu + 3\sigma$. If the threshold is too small, it is set to a default value.
6. Any bins containing more events than the threshold are recorded as clusters. Adjacent bins exceeding the threshold are merged into larger clusters.

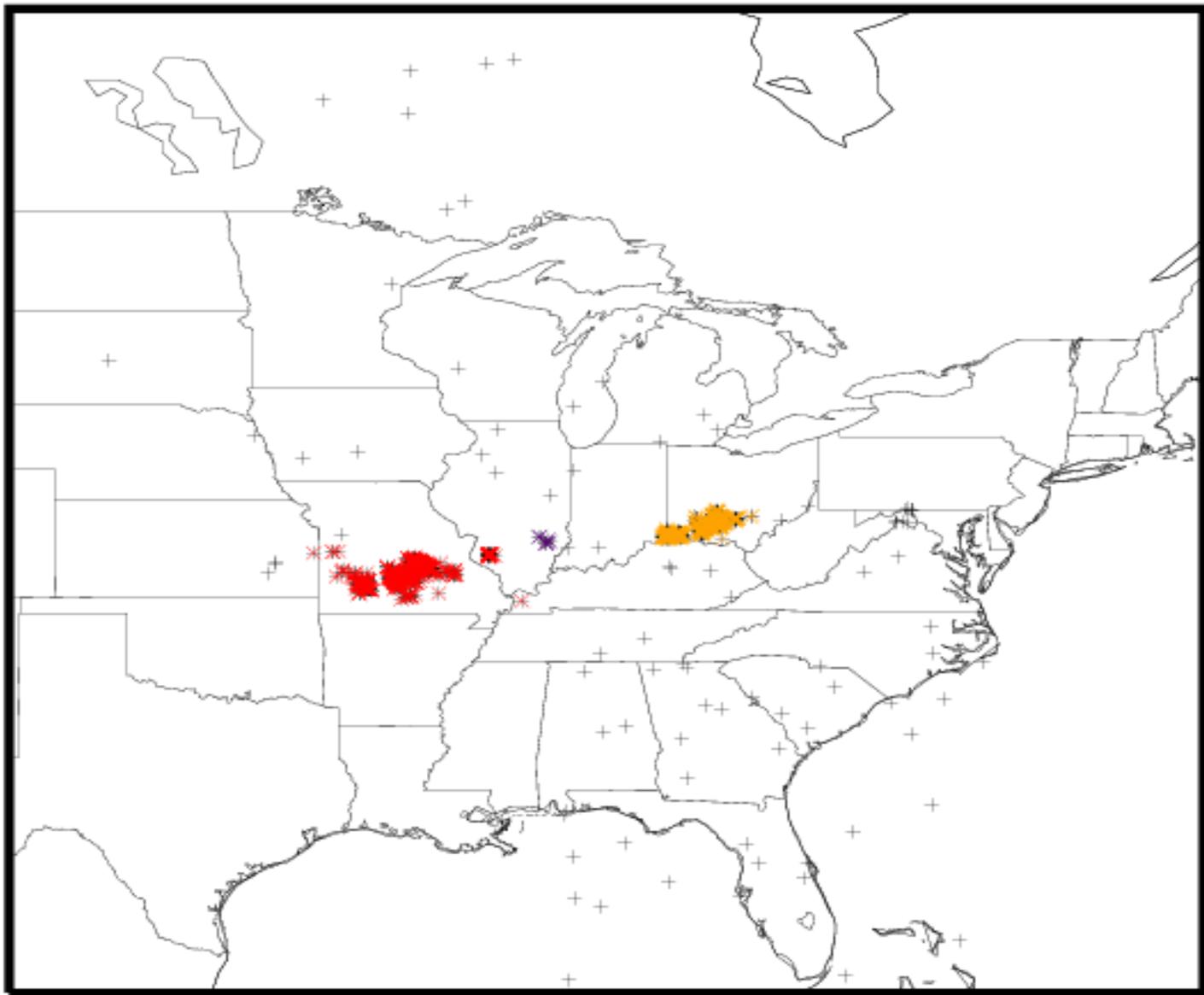


Figure 3: Same data as Figure 1, except with the 3 clusters colored.

Algorithm Validation

- We have optimized the clustering algorithm by varying the bin size and threshold parameters.
- By looking at 50 passes of the satellite in each of 3 geographical regions, we measure how many valid clusters are missed, and how many false clusters are identified.
- The algorithm is more likely to miss clusters of events than it is to falsely identify clusters.
- On average, the algorithm identifies ~75 % of LLS data as being associated with a cluster.

Missed/False-Alarm Rate Table

<u>Geographical Region</u>	<u>% Missed</u>	<u>% False</u>
Continental U.S.	2.19	4.74
Central America	17	0
Maritime (Pacific)	1.64	0

Overview

- Although the clustering algorithm correctly identifies clusters, we must also ask, to what extent are these clusters associated with storm activity?
- We use GOES cloud imagery to validate the storm identification by evaluating whether LLS event locations overlay with cloud data

Data

- We chose a data set that included 225 clusters in 50 passes of FORTE over North America during the month of August, 1999.

Results

- We chose a data set that included 225 clusters in 50 passes of FORTE over North America during the month of August, 1999.
- Of this data, 98 % of the clusters were associated with storm activity based upon our definition
- All of the clusters not overlaying cloud imagery were taken during local daylight hours, and significant levels of glint were present in the raw data.
- We conclude that clusters taken during daytime conditions should be treated as suspect unless accompanied by data from other FORTE instruments

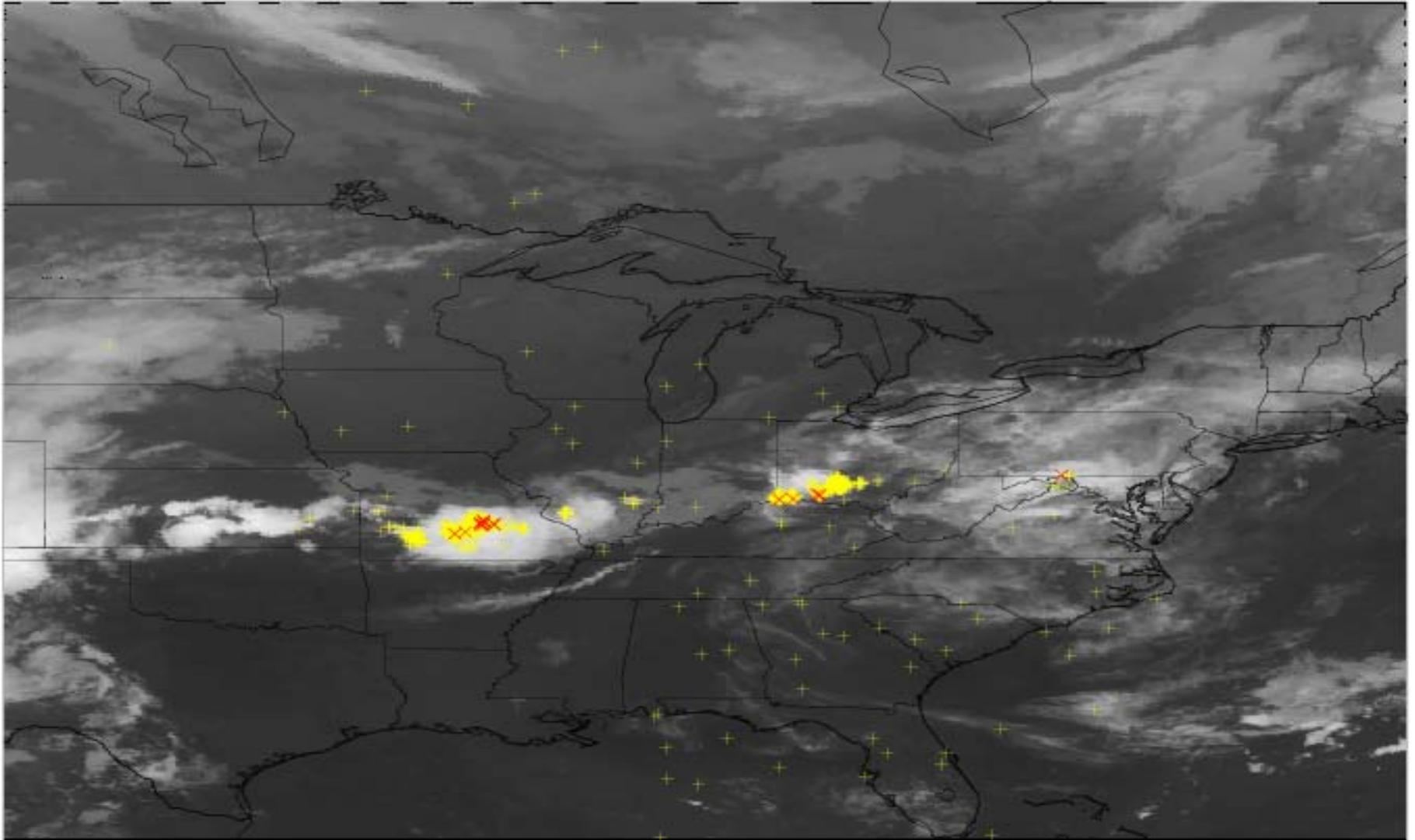


Figure 4. A sample pass over the Continental U.S. LLS data is plotted in yellow. Time coincident FORTE VHF/LLS events in red. GOES-8 cloud image from 8/1/99 05:32 UT.

Conclusions

- The clustering algorithm accurately identifies clusters of FORTE LLS data, erring on the side of missing valid clusters instead of falsely identifying non-clustered data.
- Under nighttime conditions, clusters of LLS data overlay cloud imagery, and are therefore assumed to be associated with storm activity.

Future Work

- Study LLS event rate as a function of cluster size, location, season.
- Use clustered LLS data in corroboration with other FORTE instruments for storm characterization.
- Use clustered LLS data in corroboration with ground-based instruments for automated storm identification.
- Study cluster size as a function of location, season.

Future Work: Data Post-Processing

- There are several post-processing issues that will be addressed in the near future:
 1. Merging of smaller clusters into larger ones - We have seen in some cases that a single storm can have more than one cluster of LLS data associated with it. We desire to merge clusters so that in general, storms are represented by only one cluster.
 2. Automated storm parameterization using cloud imagery - Using cloud imagery, we can filter out clusters not associated with storm activity. Also, it should be possible to come up with estimates of cloud-top temperature / height by looking at the cloud region circumscribed by LLS clusters. This may lead to a study of LLS flash rate as a function of cloud-top temperature, among other things.