Performance Assessment of the Optical Transient Detector and Lightning Imaging Sensor. Part II: Clustering Algorithm

Douglas M. Mach, Hugh J. Christian

University of Alabama in Huntsville, Huntsville, AL 35899

Richard J. Blakeslee, Dennis J. Boccippio, Steve J. Goodman

NASA/MSFC, Huntsville, AL 35805

William Boeck

Niagara University, Niagara Falls, NY 14109

Abstract

We describe the clustering algorithm used by the Lightning Imaging Sensor (LIS) and the Optical Transient Detector (OTD) for combining the lightning pulse data into events, groups, flashes, and areas. Events are single pixels that exceed the LIS/OTD background level during a single frame (2 ms). Groups are clusters of events that occur within the same frame and in adjacent pixels. Flashes are clusters of groups that occur within 330 ms and either 5.5 km (for LIS) or 16.5 km (for OTD) of each other. Areas are clusters of flashes that occur within 16.5 km of each other. Many investigators are utilizing the LIS/OTD flash data; therefore, we test how variations in the algorithms for the event-group and group-flash clustering affect the flash count for a subset of the LIS data. We divided the subset into areas with low (1-3), medium (4-15), high (16-63), and very high (64+) flashes to see how changes in the clustering parameters affect the flash rates in these different sizes of areas. We found that as long as the cluster parameters are within about a factor of two of the current values, the flash counts do not change by more than about 20%. Therefore, the flash clustering algorithm used by the LIS and OTD sensors create flash rates that are relatively insensitive to reasonable variations in the clustering algorithms.

1. Introduction

Lightning consists of a series of multiple electrical breakdown pulses producing high current channels with path lengths measured in kilometers [*Uman*, 1987]. Lightning pulses have been typically clustered into basic units called "flashes" [e.g., *Livingston and Krider*; 1978, *Goodman et al.*, 1988; *Ushio et al.*, 2001]. In small storms with low levels of lightning activity, classic individual flashes are easily discerned. In very active large storms, however, the division between the flashes become difficult [e.g., *Williams et al.*, 1999]. In most cases the definition of the flash depends heavily on the instrument used to detect the lightning. Each lightning detecting system detects only a fraction of the lightning event such that limitations within the systems bias the results. The instrument results are then further biased by the logical, but still somewhat arbitrary, limits placed on the data by the human analyst. The definition of the flash is thus influenced by both human and instrument biases.

The Lightning Imaging Sensor (LIS; 1997-present) and Optical Transient Detector (OTD; 1995-2000) are low Earth orbit (LEO) instruments that detect optical pulses from lightning flashes during both day and night [*Christian et al.*, 1992; 1996; 1999; *Boccippio et al.*, 2000; 2002]. The LIS is a component of the Tropical Rain Measuring Mission (TRMM) satellite while OTD was part of the Microlab satellite. The data from the instruments are currently being used for scientific study [e.g., *Nesbitt et al.*, 2000; *Ushio et al.*, 2001; *Christian et al.*, 2003]. Rather than using the raw filtered optical pulse data, most investigators prefer using the clustered data produced by the

LIS/OTD software algorithms. The most frequently used data set is the LIS/OTD "flash" data.

The manner in which the flashes were created from the raw LIS/OTD data is a very critical bit of information because many of the scientific results depend directly on the LIS/OTD "flash" data [e.g., *Christian et al.*, 2003]. The definitions and determinations of the "flash" units are buried deep within the LIS/OTD clustering software (which is detailed in the Algorithm Theoretical Basis Document (ATBD) for the Lightning Imaging Sensor (LIS) [*Christian et al.*, 2000]) and these determinations are not obvious to the typical user of the data. It is imperative that the algorithms and their sensitivity to variations in their values be presented in the open literature, given the fact that we and others are already using the clustered data.

In this paper, we first define the temporal and spatial limitations of the LIS/OTD data. We detail the current algorithms used to cluster the raw LIS/OTD data in an appendix. We then take a subset of the data and vary the parameters of the clustering algorithms to determine how variations affect the output product. The results of this study will be applicable to instruments beyond LIS and OTD because these algorithms will likely be the basis of future algorithms for other orbital lightning instruments.

2. LIS/OTD Data Clustering Limitations

As in most systems, the clustering of LIS and OTD data is limited by the resolution of the instrument. For LIS and OTD, there are three instrumentation characteristics that impact the data and the clustering algorithm: 1) pixel integration time, 2) pixel spatial resolution, and 3) signal to noise ratio. Each of these three limitations will affect the absolute limits to the LIS/OTD clustering algorithms. No LIS/OTD clustering algorithm can be of a finer resolution than these limits.

2.1 Pixel Integration Time

The LIS and OTD instruments detect light by the summation of the photon count incident on a pixel over the integration time of the sensor. In an integrating sensor, such as LIS/OTD, the integration time specifies how long a particular pixel accumulates charge between readouts. The median optical lightning pulse width when viewed from above is around 400 μs [*Christian and Goodman*, 1987; *Christian et al.*, 1989; *Mach et al.*, 2005] which would indicate that an integration time of 1 ms would be most appropriate to minimize pulse splitting and maximize lightning detectability. Technological limitations, however, forced a 1.9 ms integration time with a readout time of 0.1 ms (for a net frame-to-frame time of 2.0 ms). The integration time design limitation results in the possibility of multiple lightning pulses occurring in the same LIS integration window. From a clustering point of view, any LIS/OTD algorithm will be limited to a time resolution of 2 ms.

2.2 Pixel Spatial Resolution

The LIS and OTD instruments are limited in their spatial resolution to the size of the CCD pixels. With the altitude, physical pixel size, and lens system used for the LIS, the geometrically projected pixel size is about 4 km on a side. Therefore, the resolution of any LIS clustering algorithm will be limited to the size of the LIS pixel convoluted with the size of the cloud illumination. Although OTD uses the same lens system and physical pixel size as LIS, the significantly higher OTD orbit (~750 km for OTD verses ~350 km for LIS) sets the pixel size to about 8 km on a side [*Boccippio et al.*, 2000;

Boccippio et al., 2002]. This means that no LIS or OTD clustering algorithm will distinguish individual lightning channels.

2.3 Signal-to-Noise Ratio

The lightning signal-to-noise ratio improves as the integration period approaches the pulse duration. Yet, if the integration period becomes too short, the lightning signal tends to be split between successive frames which actually decreases the signal-to-noise ratio. For daylight operations, the lightning signal is swamped by the strong cloud background illumination. Even with all of the filtering designed to extract the lightning signal out of the cloud background level, many weak lightning flashes will be missed [*Boccippio et al.*, 2002]. The estimated lightning pulse detection efficiencies for LIS and OTD range from around 93% for LIS during the night to 44% for OTD during the day [*Boccippio et al.*, 2002]. The detection efficiency places another limitation on any clustering method used to classify the LIS and OTD data as it is impossible to cluster something that is not detected.

3. LIS/OTD Data Clustering Method

Once a lightning signal is detected by either the OTD or LIS on-board hardware and software, the data is sent to the ground for processing into flashes and other lightning related parameters by the LIS and OTD software [*Christian et al.*, 2000]. The goal of our software design was, given the LIS and OTD data, to create data units that were as close to the "classic" definition of a flash as possible. Although the two clustering algorithms (one for OTD and one for LIS) differ in some ways, they are fundamentally the same. When the two algorithms differ in a way that impacts clustering, we will describe both algorithms separately. The only differences between the two algorithms that will affect clustering are in the areas of distance and time separations allowed between different groups and flashes. In cases where the clustering algorithms use the same (or very similar code), we will describe them both as a single (LIS/OTD) algorithm.

3.1 Clustering Element Definitions

The LIS/OTD clustering algorithm is based on a tree or parent-child relationship between the individual levels of the clustering. Each item of the lower cluster level is associated with one and only one parent (higher) item, while each parent item can be associated with many child or lower level items (see Figure 1). There are no cross associations nor are there any partial associations (such as a "fuzzy set" approach). An example of how lightning signals are clustered in the current algorithm is illustrated in the Appendix.

3.1.1 Events

An Event is the basic unit of LIS/OTD data. It is defined as the occurrence of a single pixel exceeding the background threshold during a single frame. In other words, each pixel output from the on-board instrument hardware and software produces a separate event. Although an event can be thought of as a single optical pulse due to lightning, it is possible that multiple pulses occurring within the 2 ms frame may contribute to an event. Therefore, we avoided the label of "pulse" or "stroke" (or other similar name) to describe the basic unit of the LIS/OTD data.

3.1.2 Groups

A lightning discharge will usually illuminate more than one LIS/OTD pixel during a single frame. When these multiple events are adjacent to each other (a side or corner of the events touching), they will be placed in a single Group. The formal definition of a group is one or more simultaneous events (i.e., events that occur in the same time integration frame) that register in adjacent (neighboring or diagonal) pixels in the focal plane array. Our goal for the event-group clustering algorithm was to match as closely as possible the common definition of a lightning pulse. However, because of the spatial and temporal limitations of the raw OTD/LIS data, groups will not always correspond to a single lightning pulse.

3.1.3 Flashes

A lightning flash consists of one to multiple optical pulses that occur within a specified temporal and spatial range. A LIS/OTD Flash corresponds to several related groups in a limited spatial and temporal range. A flash may include as few as one group with a single event or it may consist of many groups, each containing many events. We use a simple group-flash clustering algorithm that is based on the principle that groups that are close in time and space are more likely part of the same flash than groups that are not close in time and space.

The LIS algorithm uses a Weighted Euclidean Distance (WED) method [*Hartigan*, 1975] to test whether two groups should be assigned to the same flash. Figure 2 shows the WED relationship. The spatial and temporal weighting constants are presently assigned values of 5.5 km and 330 ms. Two groups are said to belong to the same flash if the WED between the groups is less than 1.0. In other words:

$$\mathrm{WED}^2 = (X/5.5)^2 + (Y/5.5)^2 + (T/330)^2 , \qquad (1)$$

where X is the East-West and Y is the North-South distances in kilometers between the two group centroids and T is the time difference between the two groups in milliseconds. Note that for groups of more than one event, the WED is calculated between the centroids of each group. The clustering algorithm performs a group-by-group comparison of the WED differences between groups, and assigns groups to the same flash if their temporal and spatial properties are consistent with the above criteria. Effectively, a group only has to be spatially and temporally near one other group in a flash to be part of a multi-group flash. As a result, there is no absolute time limit to a flash. That is, as long as subsequent groups are produced with a WED less than 1.0, all groups will be assigned to a single flash.

The OTD algorithm uses a simple box distance where the spatial parameter is 16.5 km and the temporal parameter is 330 ms. Any group centroid that is within 16.5 km and 330 ms of any other group centroid will be placed into the same flash. Note that the OTD algorithm does not use the WED to determine the clustering of groups. As long as two group centroids are less than 16.5 km apart in the East-West direction, 16.5 km apart in the North-South direction, and less than 330 ms apart in time, they are placed in the same flash.

3.1.4 Areas

Lightning is produced in thunderstorm cells that have dimensions of about 10 km in diameter. Many storms, however, are multicellular and may extend over large areas and exist for many hours. Individual storms generally last much longer than LIS or OTD will view them. Therefore, our flash clustering algorithm does not have an explicit time limitation. Flashes that occur within a set distance of each other are clustered into what we call Areas. There is an implicit time limit to an area of ~80 seconds for LIS and ~160 seconds for OTD because LIS and OTD view a location on the earth for only about those times.

We define an area as a near contiguous region on the surface of the Earth that has produced lightning (defined as a set of flashes) during a single orbit of LIS or OTD. An area thus defined consists of a set of flashes separated in space by no more than 16.5 km. An area may include many flashes consisting of many groups and events or it may contain as few as one event (i.e., one flash consisting of one group which in turn consists of one event). There is no interflash or absolute time limit rule imposed on the area definition because, as noted previously, the LIS and OTD viewing times are much shorter than a storm life cycle. For isolated storms, it is simple to assign flashes to a single area/storm. For multicellular systems, however, the assignment of flashes to a single cell is more difficult. If a flash can be assigned to more than one area, the oldest area is assigned the flash.

This clustering algorithm will create cell-sized areas for isolated storms. For multicellular storms and storm systems, the LIS/OTD flash-area clustering algorithm can produce areas that are as large as the whole storm complex. However, based on the order of the flash occurrence, individual storms and sets of storms smaller than the whole complex can be clustered into one or more areas. An area will not always correspond to a single storm or the whole storm complex but a complicated function of the order of occurrence of the flashes.

4. Algorithm Variations

For each clustering level (event-group, group-flash, and flash-area), we have chosen logical but arbitrary rules. Any variation to these clustering rules will affect the statistics of the LIS/OTD data used by us and other researchers. Obviously, there are an infinite number of ways to vary the LIS and OTD clustering algorithms. For this study, we will only vary the algorithms slightly about their current values to see how these changes affect the summary results.

The most common product in the LIS data used by researchers is the "flash rate"; thus, we will concentrate on those algorithm changes that most affect the flash rate. In our case, this will actually be the flash counts because LIS/OTD view a region for a limited time period. The flash-area algorithm will not affect the overall flash rates (will only affect flash rates for individual areas); we will only investigate changes to the event-group and group-flash algorithms.

It was impractical to run the tests on the complete LIS dataset (over 2 million areas with over 10 million flashes), we have chosen a subset of six days spread across the LIS test dataset. One day was randomly chosen from each year of the LIS (1998 to 2003) dataset. We also made sure that the days were spread over the seasons in an attempt to remove any seasonal bias in our data subset. We did this to attempt to mimic the full LIS dataset with a much more manageable sized dataset. Figure 3 shows the flash density for all LIS data from January 1998 through February 2006. Figure 4 shows the flash density for the data subset using the current LIS algorithm.

There are 15505 flashes in the test data subset (based on the current LIS algorithm). The sample contains flashes from all seasons and both continental and oceanic storms. The sample is dominated by continental storms as is the full LIS dataset. Table 1 summarizes the dataset using the current LIS algorithm. Variations in the algorithm may affect low and high flash rate storms differently; therefore, we further divide the data into flash count categories ranging from 1 flash per area (very low flash rates of 1 per ~95 seconds) to areas with greater than 256 flashes (flash rates of almost 3 per second). Figure 5 graphically displays the distribution of events, groups, flashes, and areas for our test dataset and the full LIS dataset (January 1998 to February 2006) as a function of flash rate. We will compare the changes to these distributions as a function of the variations in the LIS clustering algorithm.

4.1 Variations in the Event-to-Group Algorithm

Our goal for the event-group clustering algorithm was to match as closely as possible the common definition of a lightning pulse. In the current algorithm, we have chosen to cluster into a group only those events that occur within the same frame and that occur within adjacent pixels. We cannot make the algorithm "tighter" as we will simply produce groups that consist of only one event, thereby eliminating the "group" as a class of LIS data.

4.1.1 Pixels in Adjacent Time

It is possible that if a lightning pulse begins near the end of one integration window, it could continue into the next. This single pulse would then create two different groups in the current algorithm. It is also possible that a lightning pulse could last longer than 2 ms (e.g., from a continuing current flash). In this case, the pulse could extend over 3 or more integration periods. In addition, a pulse could be below the LIS threshold for one or more of those integration windows and cause gaps in its time history.

To explore these possibilities, we will extend the event-group clustering in time (i.e., allow a group to have more than one time). The first step is to allow groups to extend in time, but demand that there be no time gaps. This would allow any pulses that were split across integrations times to be recombined. The next step is to allow for one or two consecutive frame gaps in the event-group clustering algorithm. This would recombine any groups that are continuous in time that tend to have weaker optical emissions during sections of the pulse and cause gaps in the LIS/OTD event list. Other than the changes to the event-group clustering, the rest of the algorithm (group-flash, and flash-area) remains the same.

Figure 6 shows the effect on the group/flash/area clustering count of allowing groups to extend in time. The number of areas remained nearly the same (within 1%). As expected, the number of groups decreased by almost 50% because groups could extend in time and have 2 frame (~4 ms) gaps. Although the total number of groups decreased when groups can extend in time, the number of flashes actually increases by up to 10%. The reason for the somewhat unexpected increase in flash counts is that the combined groups cause some flashes to be split in the group-flash algorithm. Some of the combined group centroids become too far away from the original parent flash because the group-flash algorithm uses group centroids (and not event locations) to cluster the groups into flashes. For example, Figure 7 shows a set of three events that occur at different times (in the order 1, 2, and 3). In the current algorithm, each event would be assigned to a different group. The first group (event 1) would be combined with the second group (event 2) and then third group (event 3) to form a single flash. If events 2 and 3 were first combined into single group (allowing for the groups to have a time extent), then their centroid would be too far away from the centroid of the first group/event to be part of that flash. The combined group (events 2 and 3) would then be assigned a new flash.

Figure 8 shows how the flash counts changed (as we increased the time extent of the groups) depending on the approximate storm flash rate (areas flash count). Note that for all flash rate areas except for the lowest, the number of flashes increases as the groups were allowed to extend in time. The maximum group time extent for the test sample ranges from 10.5 ms (groups can extend in time but have no gaps) to 676 ms (groups can have up to four consecutive frames of no data).

4.1.2 Pixels in Non-Adjacent Locations

The current event-group algorithm demands that all events in a group be from adjacent pixels in the LIS/OTD field of view. Some groups can be quite extensive (as seen in Figure 9). If the optical signal from one or more of the pixels in the pulse were to fall below the threshold for detection, due to either a dim section of channel or intervening non-illuminated cloud, there could be "holes" in the lightning signal as seen by the LIS or OTD. These holes can split a pulse into two or more groups. To account for this, we could allow non-adjacent, but simultaneous pixels at the same time to be included in a single group. It is very unlikely that two different lightning pulses would occur at different locations in the same 2 ms time window. By allowing non-adjacent, but simultaneous pixels to be included, we may recombine pulses split by low signal strength.

To test the impact of spatial gaps, we change the algorithm to allow for single, double, and triple pixel gaps. Note that for groups near the edge of the field of view, these gaps can be much larger than 6 km. We apply the original temporal rule for the event-group clustering, that is, all events must occur in the same ~2 ms integration window. As in the temporal variation test, the group-flash and flash-area clustering rules will be the same as the current LIS algorithm.

Figure 10 presents the effect on the overall group/flash/area counts of allowing gaps in pixel space (all at a single time). Again, the group and area counts decrease and the flash counts increase, although all changes are less than 1% of the original values. The reason for the somewhat unexpected increase in flash counts is the same as for extending groups temporally (see Figure 7). Figure 11 shows the effect of letting groups have spatial gaps on the flash counts for the various area flash counts (storm flash rates). Note that with the exception of one value, all changes are less than 2%.

4.2 Group-to-Flash Variations

The goal of the group-flash clustering algorithm is to match as closely as possible the common definition of a flash. This is made somewhat difficult by the lack of a "common" definition of a flash in the literature for a sensor such as LIS. The main concern for the group-flash clustering is tracking spatial or temporal links in the flash progression. If the flash "goes dark" (i.e., signal drops below the LIS threshold) for a significant distance or time, single large flashes can be broken into several smaller ones. In the other direction, a very high flash rate storm can fool an algorithm into clustering several distinct flashes into one.

For this clustering variation test, we choose commonly accepted numbers to define the temporal and spatial limits of a flash. Variations in both the 5.5 km and the 330 ms limits for a flash will affect the resultant flash counts and flash rates. We will vary the time parameter from 100 to 1000 ms in 100 ms steps and then vary the spatial parameter from 0 km (overlap/touching) to 30 km (greater than the area limit) in steps of 5 km to establish how that affects the LIS flash data.

4.2.1 Time Variation

The effect on flash counts for variations in the group-flash time gap from 100 to 1000 ms is shown in Figure 12. Note that the plot shows only the relative changes in the flash and area counts. We did not display the group counts as they remained the same (the event-group algorithm was the one in the current LIS algorithm). The values are not very different from the current LIS algorithm as long as the group-flash time gap is kept between 200 ms and 1 s. When the group-flash time gap is lowered to 100 ms, the number of flashes for the dataset increases by over 60%.

Figure 13 illustrates how the various flash rate areas contributed to the overall flash count changes presented in Figure 12. As one would expect, the areas with the highest flash rates were most affected by changing the group-flash time gap limit. For all areas except the highest flash counts, once the group-flash time gap limit was raised above 300 ms, the flash counts changed by less than 5%. Only the highest flash rate areas (flash counts greater than 64) showed any major effect of changing the group-flash time gap limit from 200 ms to 1 s.

4.2.2 Distance Variation

Figure 14 shows the overall effect of variations in the group-flash spatial clustering algorithm. Again, the plot shows only the relative changes in the flash and area counts as a function of group spacing. Groups are not shown as they are not affected by the group-flash algorithm. As the allowable spacing between groups is lowered from the current 5.5 km to 1.1 km, the number of flashes rapidly increases to almost 4.5 times the value at 5.5 km. That is, there are 4.5 times as many flashes if the group-flash spacing was lowered to 1.1 km. Note that 1.1 km is less than the nominal pixel spacing (~5 km). As the group-flash spacing is allowed to increase beyond the current 5.5 km, the number of flashes decreases as one would expect; however, the fractional change from the current algorithm is no more than 10%, even for group centroid spacing of 11 km.

The total number of areas decreases if the group-flash spacing is lowered from the current value of 5.5 km to 1.1 km; however, the maximum change is less than 10%. The number of areas increases as the allowable spacing between groups in a flash is increased. The change in the number of areas is less than 1% for even a group spacing of 11 km. Figure 15 shows how the various flash rate areas contributed to the overall flash count changes show in Figure 14. As one would expect, the areas with the highest flash rates were most affected by changing the group-flash spacing limit. For all areas except the highest flash counts, once the group-flash spatial gap limit was raised above 5.5 km, the flash counts changed by less than 20%. For allowable group gaps of less than 5.5 km, all flash rate storms, with the exception of the low flash rate set (less than 3 flashes per minute), increased their flash counts by large amounts, up to 15 times the value at 5.5 km. The low flash rate set showed a decrease in the flash rate by almost 60%.

5. Conclusions

5.1 Variations in the Event-to-Group Algorithm

Reasonable variations in the event-to-group clustering algorithm have a small but measurable effect on the flash counts. As a group is allowed to extend in time (first with no dark frames and then with 1-4 dark frames), some groups in the test dataset extend in time to over 400 ms. However, the number of flashes increases by only about 10%. Even with the allowance of a four frame (~8 ms) gap, over 80% of the groups still occur in a single LIS/OTD time frame. Variations in the spatial event-to-group algorithm create even smaller changes in the resultant flash rates. Even allowing groups to have visible gaps of up to 5 pixels (which corresponds to a nominal gap of 25 km) changed the overall flash rates by less than 4%.

5.2 Variations in the Group-to-Flash Algorithm

One would expect variations in this algorithm to have the greatest impact on the actual flash counts because the group-to-flash clustering algorithm directly affects the flash count. This contention was supported only for variations that decreased the sizes of flashes. For variations that increased the size of flashes, the changes in the flash count and rates were much smaller. In the current algorithm, if two groups are separated in time by more than 330 ms, the two groups cannot be part of the same flash. Decreasing this interval had a major impact on the flash counts, but only when the value was cut to less than one third of its original value. When the group-to-flash time interval was decreased to 100 ms, the number of flashes increased by over 60%. However, if the group-to-flash time interval was either decreased to 250 ms or increased to as much as 1000 ms (one second), the changes in the flash counts were no more than 5%. For areas with flash counts of less than 64 (nominal flash rates of 45 flashes per minute or less), increasing the allowable time between groups in a single flash to as much as one second did not significantly change the flash counts/rate. For these areas, most flashes were separated by more than one second. Only in areas with flash rates near one per second or greater (i.e., very high) did the changes in the group-to-flash time limit make a major difference in the flash rates. Even for these areas, increasing the allowable time interval between groups in a flash by almost half an order of magnitude (factor of 3) decreased the flash counts by only 25%. Setting the allowable time interval between groups in a single flash between 250 and 1000 ms would not greatly affect the flash rates derived from LIS or OTD data.

In the current LIS algorithm, groups that have centroids separated by more than 5.5 km cannot be put into the same flash. Although the nominal pixel spacing is around 5 km, most groups have more than one event associated with them. That means the centroid for a group can be calculated to a resolution less than the LIS pixel spacing. If the group-to-flash clustering algorithm is changed to only allow groups with centroids within 1 km of each other, the number of flashes is almost 5.5 times the value for the current algorithm. The number of flashes becomes less than 10% different than the current algorithm only when the group centroid spacing is increased to greater than 4.5 km. Clearly, decreasing the allowable spacing between groups to values less than the pixel spacing has a major detrimental affect on the flash counts and rates. Increasing the allowable group spacing to up to twice the current value, however, only decreases the flash counts (and rates) by about 10%. If the spacing limit is between 4.5 and 11 km, the flash counts will be within 10% of the current value at 5.5 km.

6. Summary

We have described out current algorithm for clustering the LIS/OTD events into groups, flashes, and areas. The values we selected for the various parameters are based on the current knowledge of lightning, but they can still be considered to be somewhat arbitrary. Events are the basic unit of measurement of lightning data from LIS/OTD. Groups are collections of adjacent events that occur within the same 2 ms window. Flashes are collections of groups that occur within 330 ms and either 5.5 km (for LIS) or 16.5 km (for OTD) of each other. Areas are collections of flashes that occur within 16.5 km of each other. We set out to determine the effect of variations in the parameters because variations in the clustering parameters could affect the research results based on the LIS/OTD data.

We found that extending the group definition to include events at different times or with temporal or spatial gaps will not significantly affect the overall flash counts derived from LIS/OTD data. The flash counts and rates are not strongly sensitive to variations in the current event-to-group clustering algorithm. This is because a maximum of about 20% of the groups are close in time or space to other groups. Therefore, almost any reasonable event-to-group clustering algorithm will reproduce the lightning flash statistics of *Christian et al.* [2003].

We also found that as long as the group-to-flash clustering algorithms are not made more restrictive than the current values, the affect on the flash counts (and rates) are minimal. Other orbital lightning sensors that use other clustering algorithms, such as the Fast On-orbit Recording of Transient Events, (FORTE) [e.g., *Suszcynsky et al.*, 2001; *Davis et al.*, 2002] should achieve similar flash count results as long as their clustering algorithms use values that are no more restrictive than ours.

Overall, it would take radical changes in the LIS/OTD clustering algorithm to make major changes in the flash counts or rates derived with the current algorithms. Variations in the algorithm parameters of up to a factor of two often change the flash counts by less than 10%. The LIS/OTD datasets seem to be somewhat insensitive to even some major changes in the clustering algorithm. This robust clustering algorithm should increase the confidence in the LIS/OTD flash rate results.

7. Appendix: Example clustering Series

To illustrate the general clustering algorithm, we simulated a series of optical pulses within the LIS FOV and demonstrated how they cluster given the current clustering algorithm. By changing the appropriate spatial and temporal constraints, this example can also be applied to OTD data. For the purpose of this demonstration, it was assumed that there were no events prior to the events at time 0 and that the pixel grid was 4 km wide when translated into ground coordinates. In the actual LIS and OTD data, the latitude/longitude grid in earth-based coordinates and the pixel grid will not be the same size or co-registered. In addition, the simulated times will begin from the start of the simulated orbit.

7.1 Time = 0 ms

The first time frame is shown in Figure 16. Three simulated events (designated 1, 2, and 3) occurred at this time integration. They are collected into a single group (designated a) because the events were simultaneous and registered in adjacent (i.e., neighboring or diagonal) pixels. The group was assigned a new parent flash (designated A) and the new flash was assigned a new parent area (designated α).

7.2 Time = 100 ms

The next frame with data is shown in Figure 17. At this time (100 ms after the first one), there were three more events (designated 4, 5, and 6). As in the previous case, these three new events were all assigned to a new group (called b). These events were not assigned to group a because they occurred at a different time. The time difference between groups a and b was 100 ms, and the minimum ground distance between these groups was 4 km (calculated from the center of the two nearest events from each group). This WED distance between a and b is small enough to have them both assigned to the same flash. As a result, group b was assigned to the first flash A and therefore, to area α.

7.3 Time = 350 ms

The next frame with data is shown in Figure 18. The time was 350 ms after the time of the first events, but only 250 ms after the time of the last events. At this time there were four more events (labeled 7, 8, 9, and 10). Events 7 and 8 were adjacent to each other and were assigned to a new group (designated c). Events 9 and 10 were not adjacent to events 7 and 8, but are adjacent to each other. They were assigned to another new group (called d). The time difference between group b and group c was 250 ms, and because events 4 and 8 share the same pixel, the minimum ground distance between these groups was 0 km. This WED is small enough to assign group c to flash A and area α. Although group d also occurred within 250 ms of group c in flash A, its distance from any part of group c was approximately 20 km. Note that two groups that are separated by more than 5.5 km cannot possibly meet the WED criteria and therefore would not be assigned to the same flash. As a result, group d was assigned to a new flash (designated B). All parts of flash B (i.e., group d) were greater than 16.5 km away from any part of area α, so flash (B was also assigned a new area (called β).

7.4 Time = 400 ms

Figure 19 illustrates the next integration time with data. The time was 400 ms after the first events and 50 ms after the latest events. Two more events (labeled 11 and 12) occurred at this time. These two events were at the same time, but they are not adjacent to each other; they were assigned to two new groups (called e and f). The two new groups were less than 330 ms from the time of the last group of flash B and were within 5.5 km of flash B; thus, the two groups were assigned to flash B and area ß.

7.5 Time = 700 ms

The last frame with events (for this example) is shown in Figure 20. At 700 ms after the first events and 300 ms after the last events, there were two new events (designated 13 and 14). The events were not adjacent and they are assigned to two new groups (called g and h). Group g overlapped the parts of flash A; however, it has now been more than 330 ms since the last group associated with flash A. Therefore, group g was assigned to a new flash (labeled C). As parts of flash C were less than 16.5 km from parts of area α, and as there is no time limit for areas, flash C was assigned to area α. Group h is not within 5.5 km of any current flash; it is assigned another new flash (called D). Flash D is also not within 16.5 km of any currently active area and it was assigned another new area χ.

7.6 Summary Data

In the example data processing sequence just described, there were fourteen events, eight groups, four flashes, and three areas. The example showed how the

OTD/LIS algorithms would cluster events into groups, flashes, and areas. Some of the summary data statistics that would be generated from the LIS/OTD processing algorithm are shown in Tables 2 (areas), 3 (flashes), and 4 (groups) for this example.

References

Boccippio, D. J., W. Koshak, and R. Blakeslee, K. Driscoll, D. Mach, and D. Buechler (2000), The Optical Transient Detector (OTD): Instrument characteristics and cross-sensor validation, *J. Atmos. Oceanic Technol.*, *17*, 441-458.

Boccippio, D. J., W. J. Koshak, and R. J. Blakeslee (2002), Performance assessment of the Optical Transient Detector and Lightning Imaging Sensor, I, Predicted diurnal variability, *J. Atmos. Oceanic Technol.*, *19*, 1318–1332.

Christian, H. J. and S. J. Goodman (1987), Optical observations of lightning from a high altitude airplane, *J. Atmos. Oceanic Technol.*, *4*, 701-711.

Christian, H. J., R. J. Blakeslee, and S. J. Goodman (1989), The detection of lightning from geostationary orbit, *J. Geophys. Res.*, *94*, 13329-13337.

Christian, H. J., R. J. Blakeslee, and S. J. Goodman (1992), Lightning imaging sensor for the Earth Observing System, *Tech. Rep. NASA TM-4350*, NASA, Washington, D. C..

Christian, H. J., K. T. Driscoll, S. J. Goodman, R. J. Blakeslee, D. M. Mach, and D. E. Buechler (1996), The Optical Transient Detector (OTD). *Proc. 10th Int. Conf. on Atmospheric Electricity,* ICAE, 368–371, Osaka, Japan.

Christian, H. J., R. J. Blakeslee, S. J. Goodman, D. M. Mach, M. F. Stewart, D. E. Buechler, W. J. Koshak, J. M. Hall, W. L. Boeck, K. T. Driscoll, and D. J. Boccippio (1999), The Lightning Imaging Sensor. *Proc. 11th Int. Conf. on Atmospheric Electricity,* Guntersville, AL, NASA, 746–749.

Christian, H. J., R. J. Blakeslee, S. J. Goodman, and D. M. Mach (Eds.) (2000), Algorithm *Theoretical Basis Document (ATBD) for the Lightning Imaging Sensor*

(LIS), NASA/Marshall Space Flight Cent., Alabama, 2000. (Available as http://eospso.gsfc.nasa.gov/atbd/listables.html, posted 1 Feb. 2000).

Christian, H. J., R. J. Blakeslee, D. J. Boccippio, W. L. Boeck, D. E. Buechler, K. T. Driscoll, S. J. Goodman, J. M. Hall, W. J. Koshak, D. M. Mach, and M. F. Stewart (2003), Global frequency and distribution of lightning as observed from space by the Optical Transient Detector, *J. Geophys. Res., 108 (D1)*, 4005, 10.1029/2002JD002347.

Davis, S. M., D. M. Suszcynsky, and T. E. L. Light (2002), FORTE observations of optical emissions from lightning: Optical properties and discrimination capability, *J. Geophys. Res., 107, (D21)*, 4579, 10.1029/2002JD002434.

Goodman, S. J., D. E. Buechler, and P. D. Wright (1988), Lightning and precipitation history of a microburst-producing storm, *Geophys. Res. Lett., 15*, 1185-1188.

Livingston, J. M., and E. P. Krider (1978), Electric fields produced by Florida thunderstorms, *J. Geophys. Res., 83*, 385-401.

Mach, D. M, R. J. Blakeslee, J.C. Bailey, W. M. Farrell, R. A. Goldberg, M. D. Desch, and J. G. Houser (2005), Lightning optical pulse statistics from storm overflights during the Altus Cumulus Electrification Study, *Atmos. Research*, *76*, 386-401.

Nesbitt, S. W., R. Zhang, and R. E. Orville (2000), Seasonal and global NOx production by lightning estimated from the Optical Transient detector (OTD), *Tellus*, *52B*, 1206-1215.

Suszcynsky, D. M., T. E. Light, S. Davis, J. L. Green, J. L. L. Guillen, and W. Myre (2001), Coordinated observations of optical lightning from space using the FORTE photodiode detector and CCD imager, *J. Geophys. Res., 106, (D16)*, 17,897–17,906.

Uman, M.A. (1987), *The Lightning Discharge*, Academic Press, New York, NY.

Ushio, T, S. J. Heckman, D. J. Boccippio, and H. J. Krider (2001), A survey of thunderstorm flash rates compared to cloud top height using TRMM satellite data, *J. Geophys. Res., 106*, 24089-24095.

Williams, E., R. Boldi, A. Matlin, M. Weber, S. Hodanish, D. Sharp, S. Goodman, R. Raghavan, and D. Buechler (1999), The behavior of total lightning activity in severe Florida thunderstorms, *Atmos. Research., 51*, 245-265.

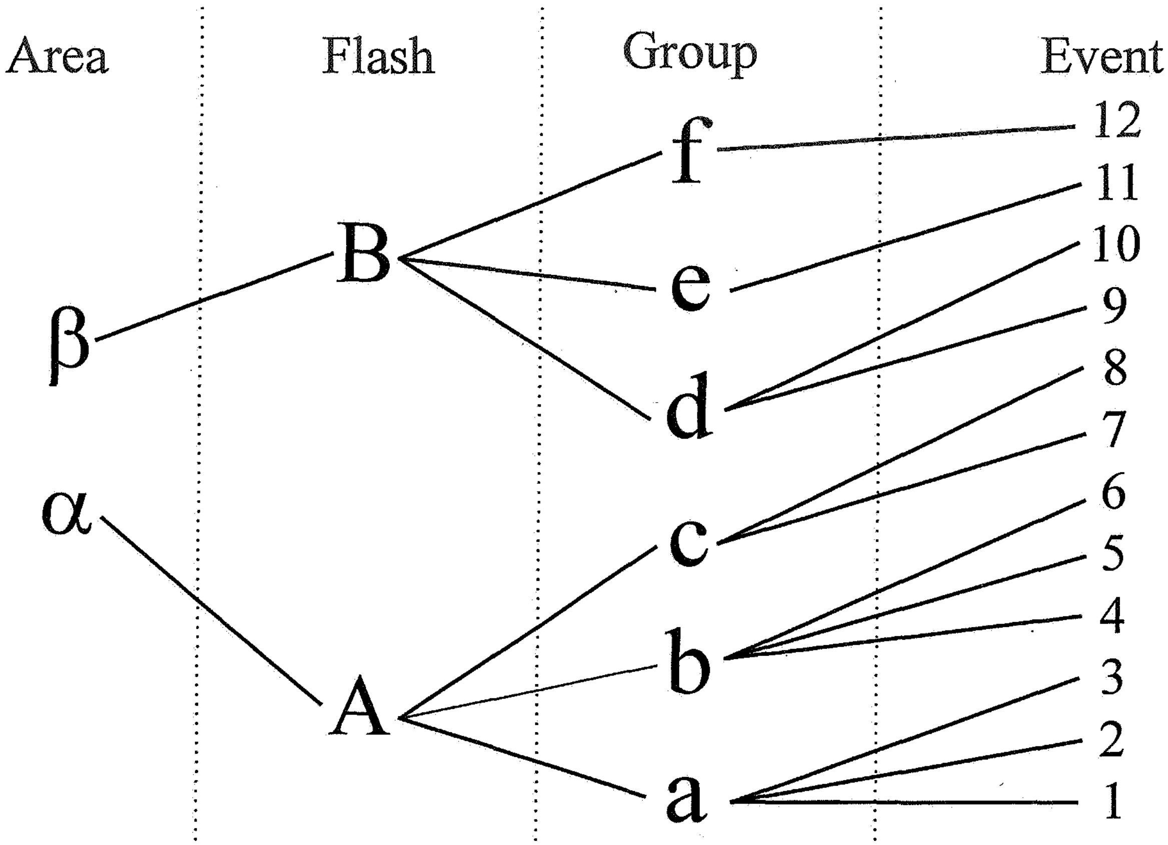


Figure 1. Clustering relationship between areas, flashes, groups, and events. The structure can be described as a 'parent-child' or 'tree' relationship where a single parent can have multiple children but each child only has a single parent.

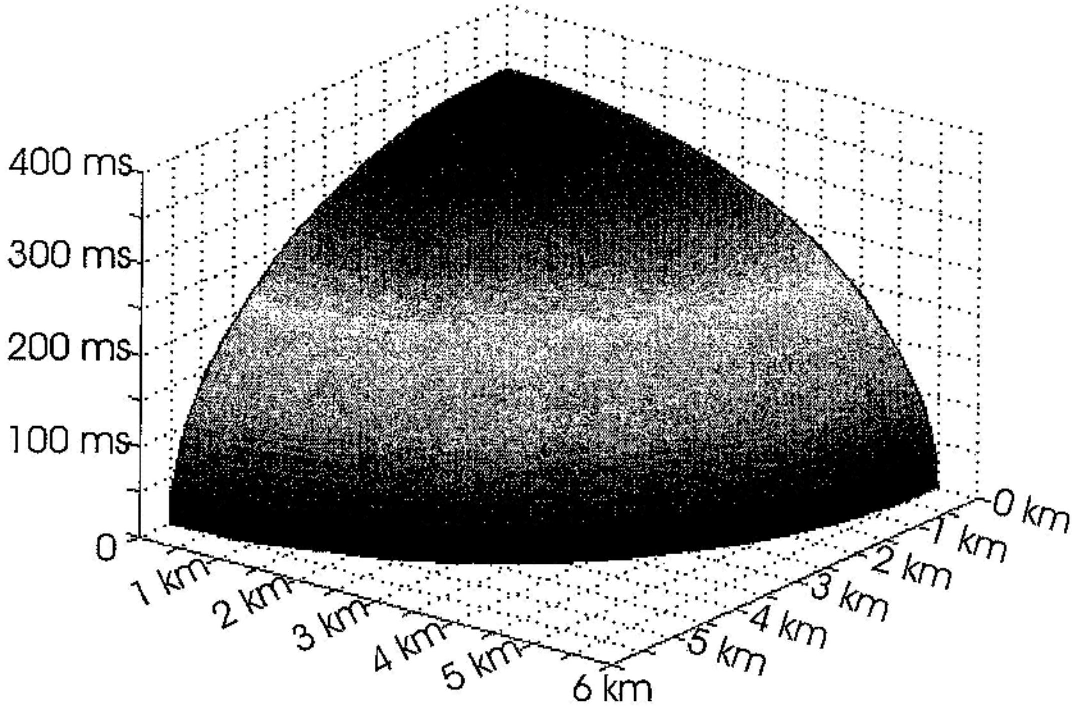


Figure 2. Weighted Euclidean Distance (WED) used for the group-flash clustering algorithm. If the WED between the two group centroids is within the sphere, the two groups will be clustered into the same parent flash.

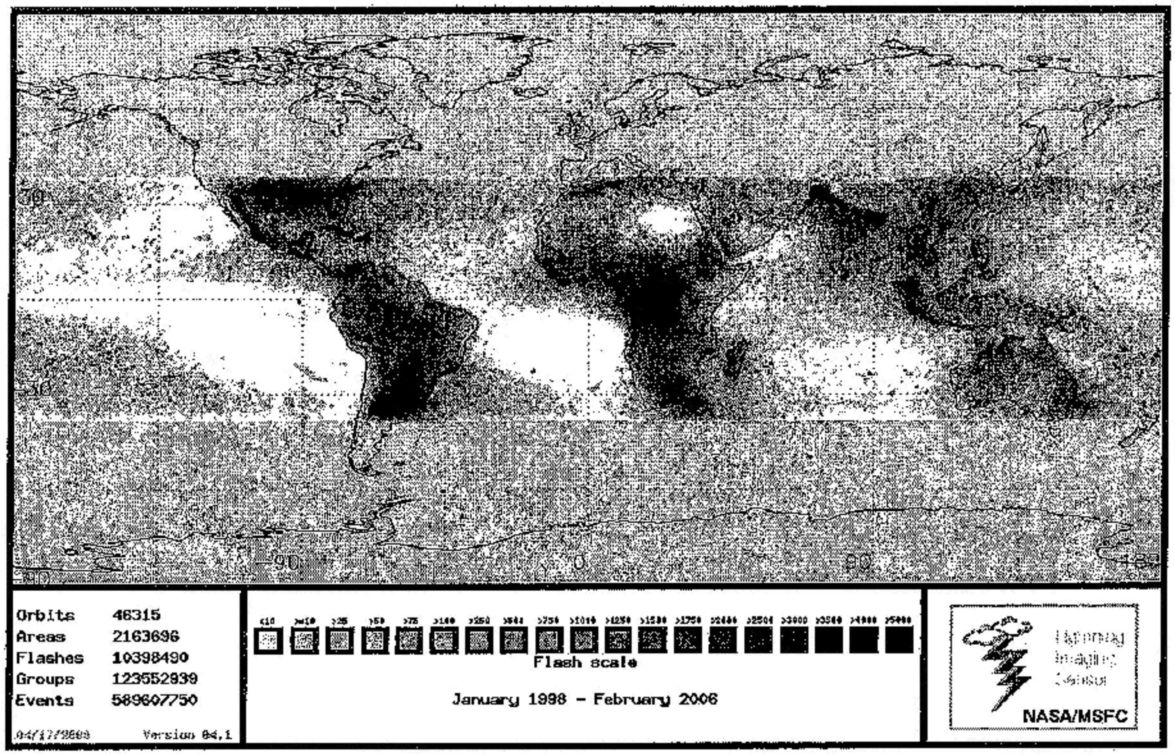


Figure 3. Distribution of LIS flash data from January 1998 to February 2006. Note that most lightning is over land with strong concentrations in Central Africa and South-Central South America.

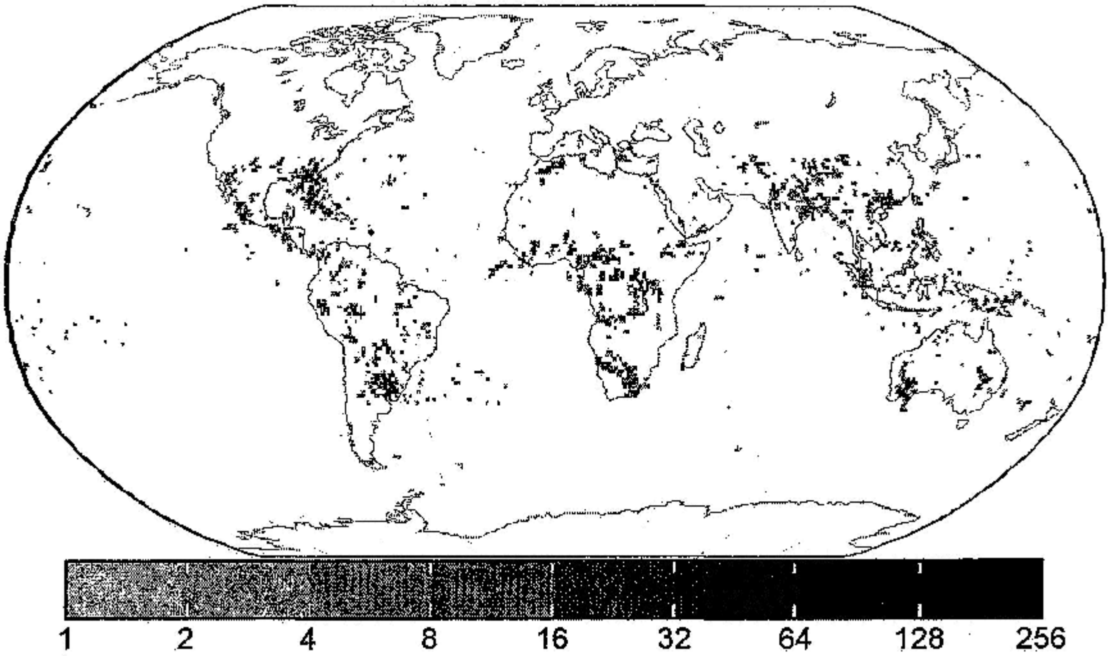


Figure 4. Distribution of LIS flash data for our test data subset. Note that the distribution approximates the distribution of the larger LIS dataset.

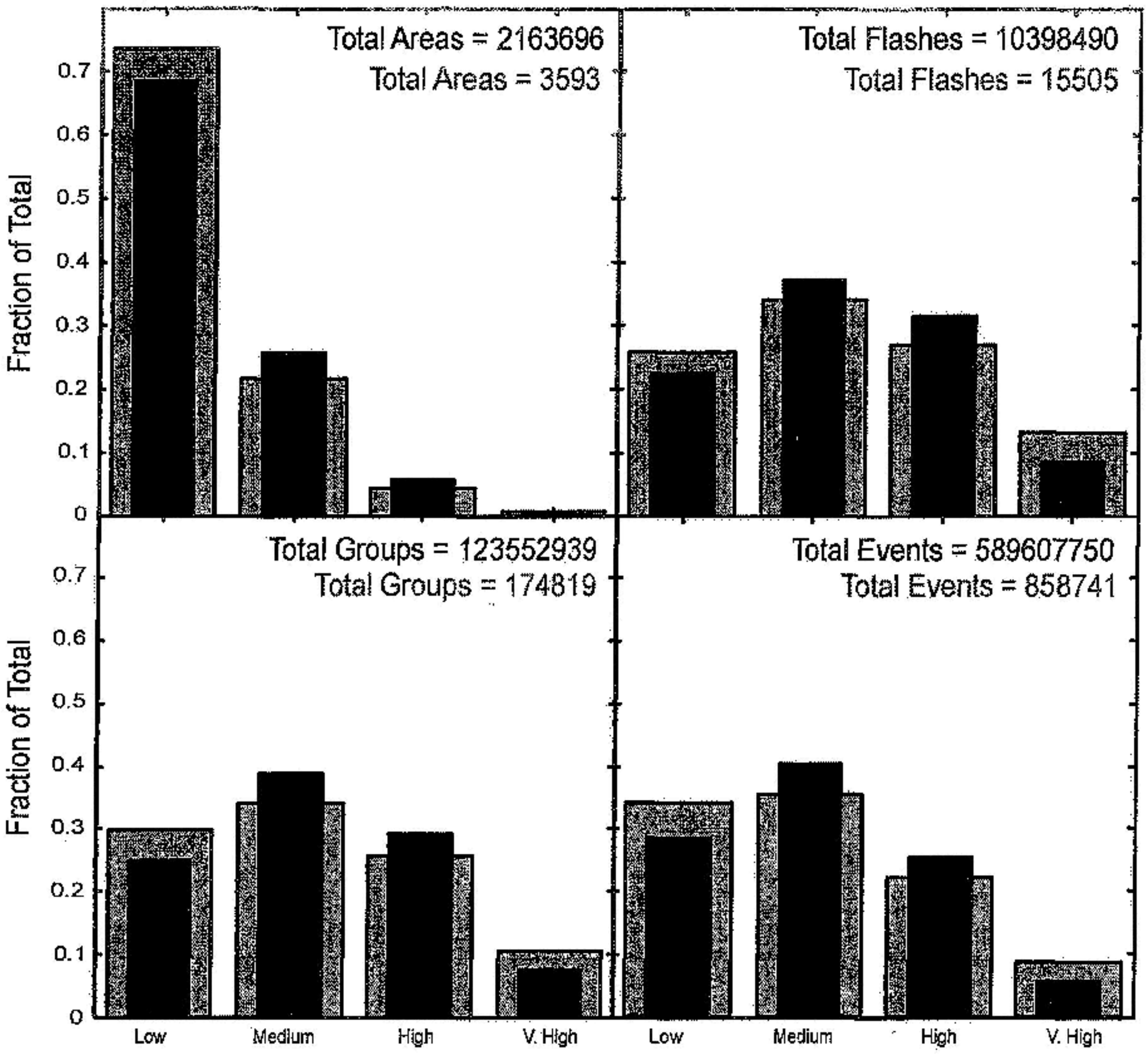


Figure 5. Distribution of the test and full LIS data based on nominal flash counts per area. Areas with flash counts of 1-3 are considered "Low" flash rate storms while areas with flash counts of 4-15 and 16-63 are considered "Medium" and "High" flash rate storms, respectively. Areas with flash counts of greater than 63 are considered to be "Very High" (V.High) flash rate storms. The black bars are for the whole LIS dataset (from January 1998 to February 2006) while the wider gray bars are for the test dataset. Note that the two distributions are similar, but not exactly the same.

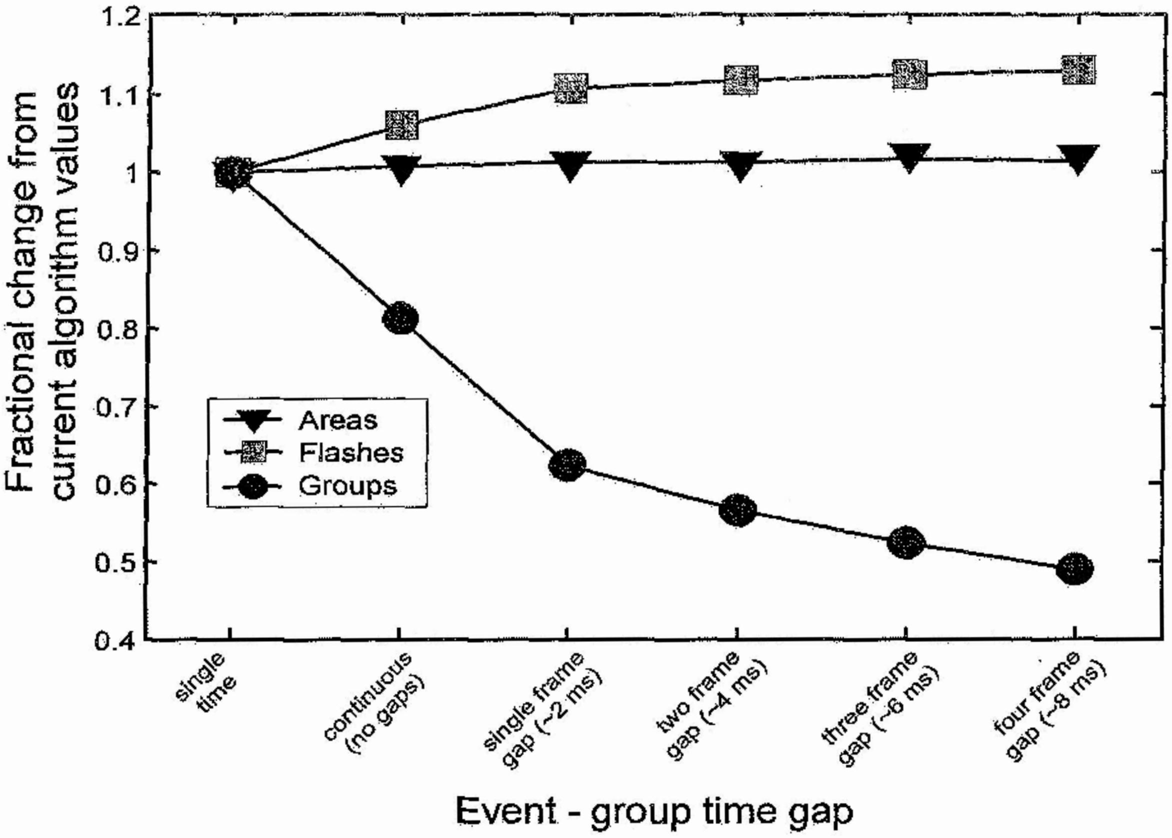


Figure 6. Fractional change in number of areas/flashes/groups as the allowable group time gap increases from single time only (current algorithm) to multiple times with no frame gaps to multiple frames with a ~8 ms (4 frame) gap.

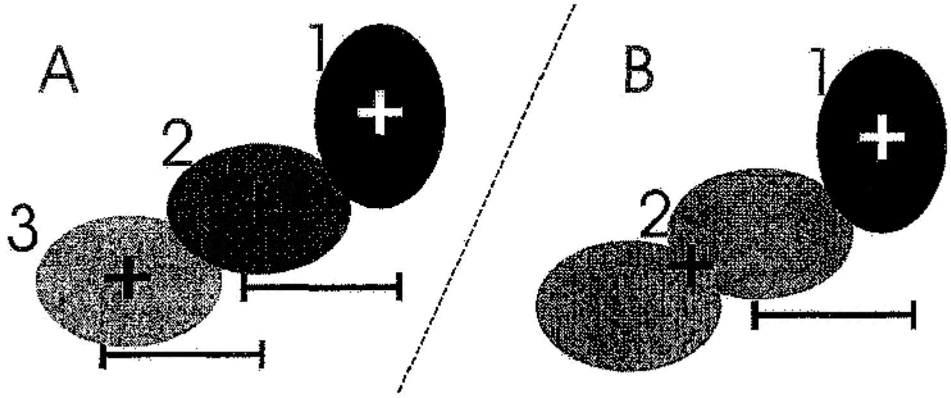


Figure 7. Effects of allowing groups to extend in time on the group-flash clustering. The horizontal bars represents the maximum distance allowed between two groups assigned to a single flash. In the current algorithm (example A on the left), the events 1, 2, and 3 would be each assigned to a different group. Group 1 would then be assigned to a flash with group 2 and then group 3 also being assigned to the flash. In the case on the right (example B, groups can extend in time), event 1 is assigned to a group while events 2 and 3 are assigned to another group. The resulting centroids of the two groups are too far away from each other to be assigned to the same flash.

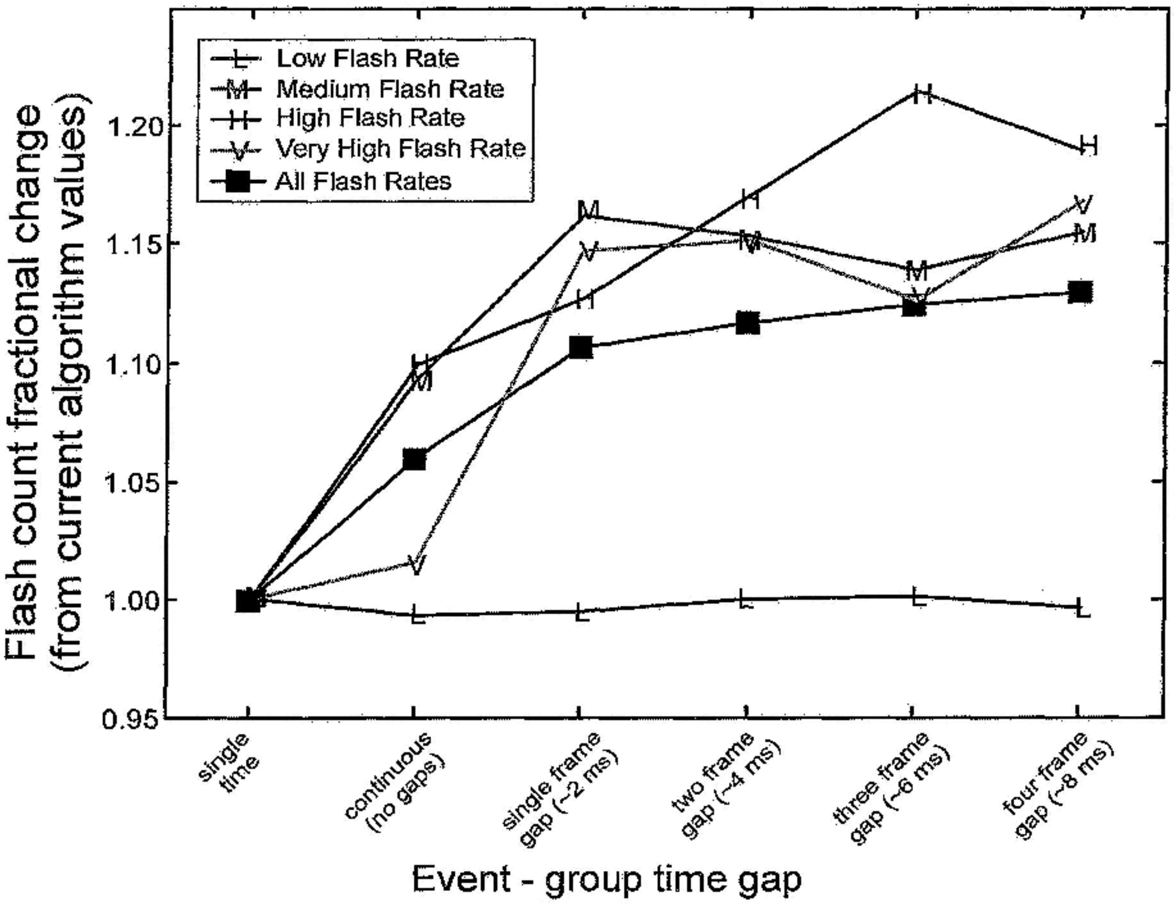


Figure 8. Flash count fractional change for low through very high flash rate areas as a function of allowable group time gap.

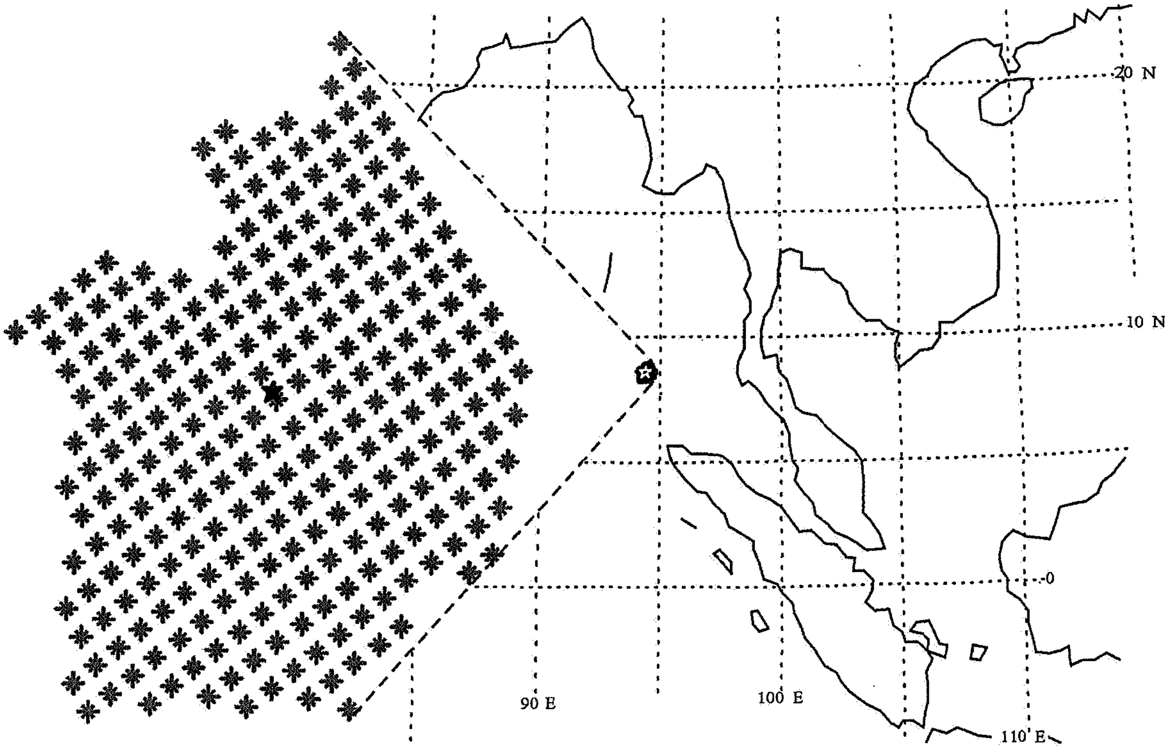


Figure 9. Large group found in test dataset. All events (as indicated by '*') occurred at the same LIS frame time. The star indicates the centroid of the group. Although none were this large, there were other groups/flashes in the area that contained this group.

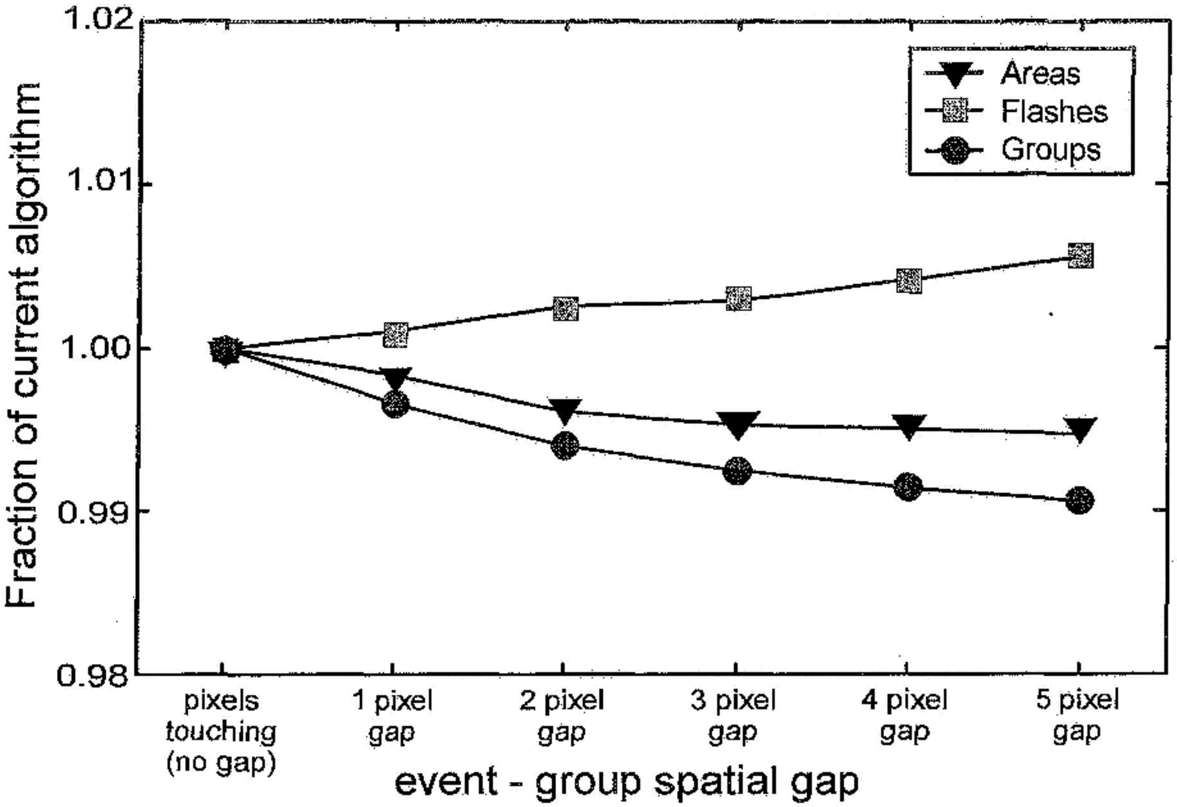


Figure 10. Fractional change in the overall group/flash/area counts as a function of pixel gaps allowed in the event-group clustering algorithm. Note that all changes were less than 2%.

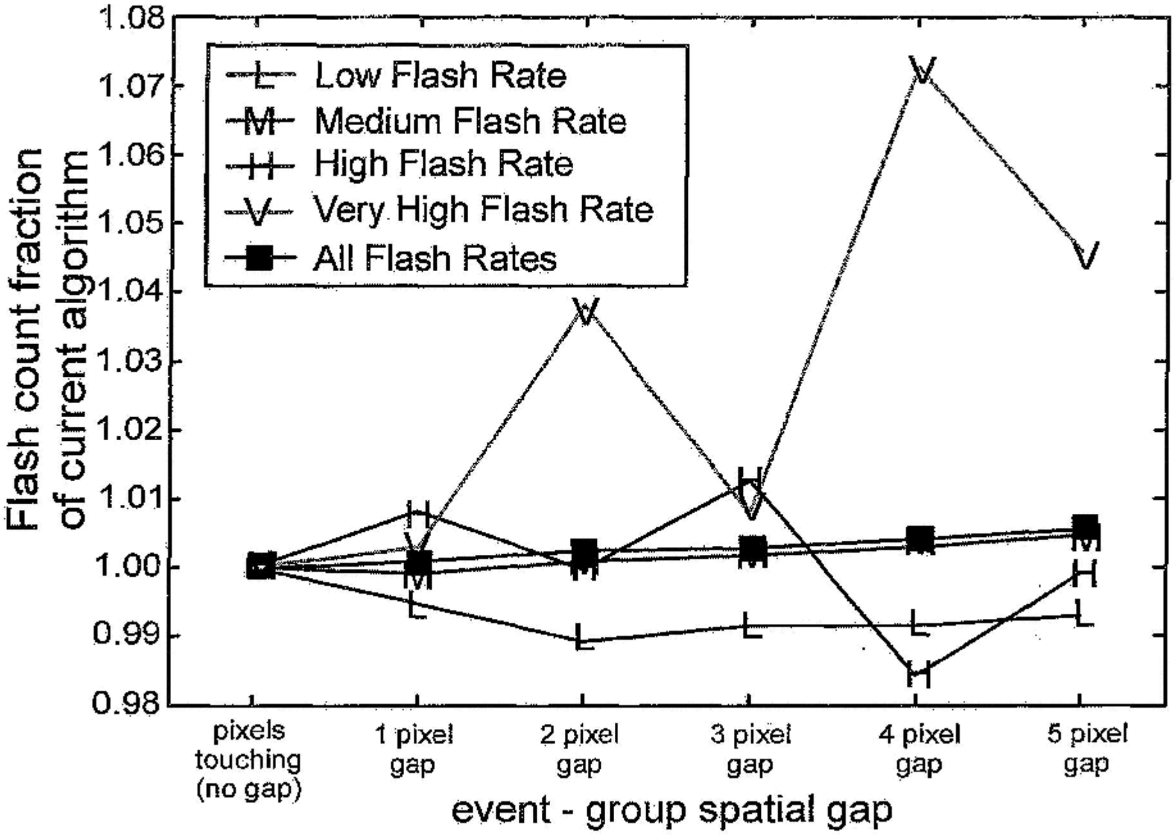


Figure 11. Flash count fractional change for low through very high flash rate areas as a function of allowable group pixel gap. Note that all changes were less than 8%.

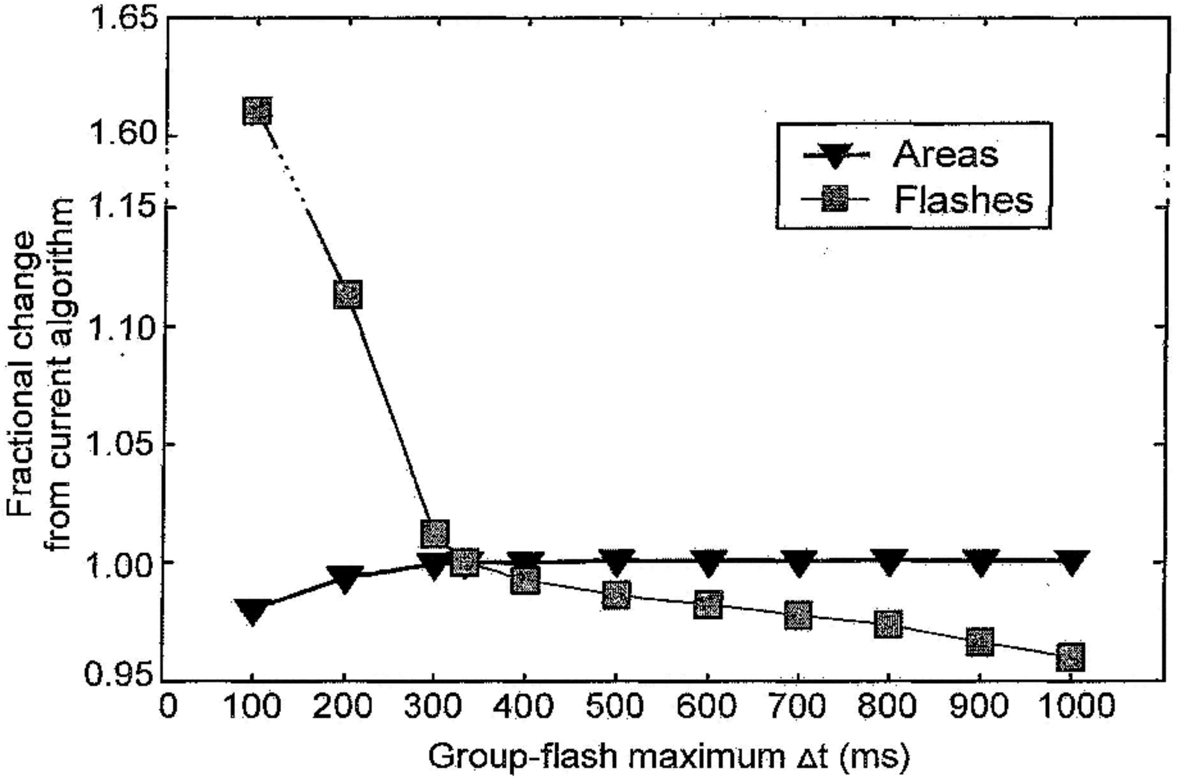


Figure 12. Fractional changes in flash and area counts when changing the group-flash time interval from 100 ms to 1 s. As long as the group-flash time gap is between 200 and 1000 ms, the change to the flash count (and global flash rate) should be less than 10%. The current LIS algorithm uses 330 ms as the group-flash time gap limit. Note the gap in the vertical scale between 1.15 and 1.60 to show both detail in the data from 200 to 1000 ms and to show the extreme value at 100 ms.

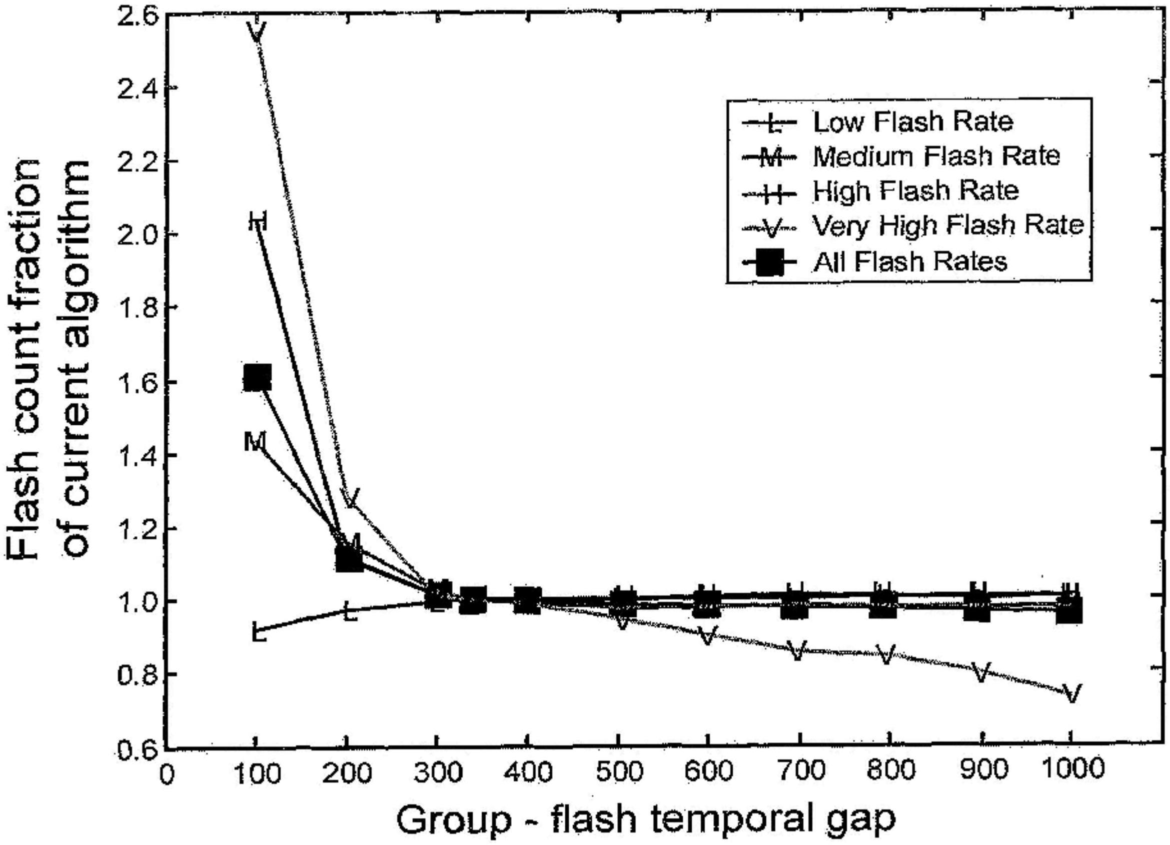


Figure 13. Fractional changes in flash counts as a function of area flash count. Note that for all flash counts except the highest (greater than 64 flashes per area), the changes to the overall flash rate was less than 10% for group-flash time gap limits between 200 and 1000 ms.

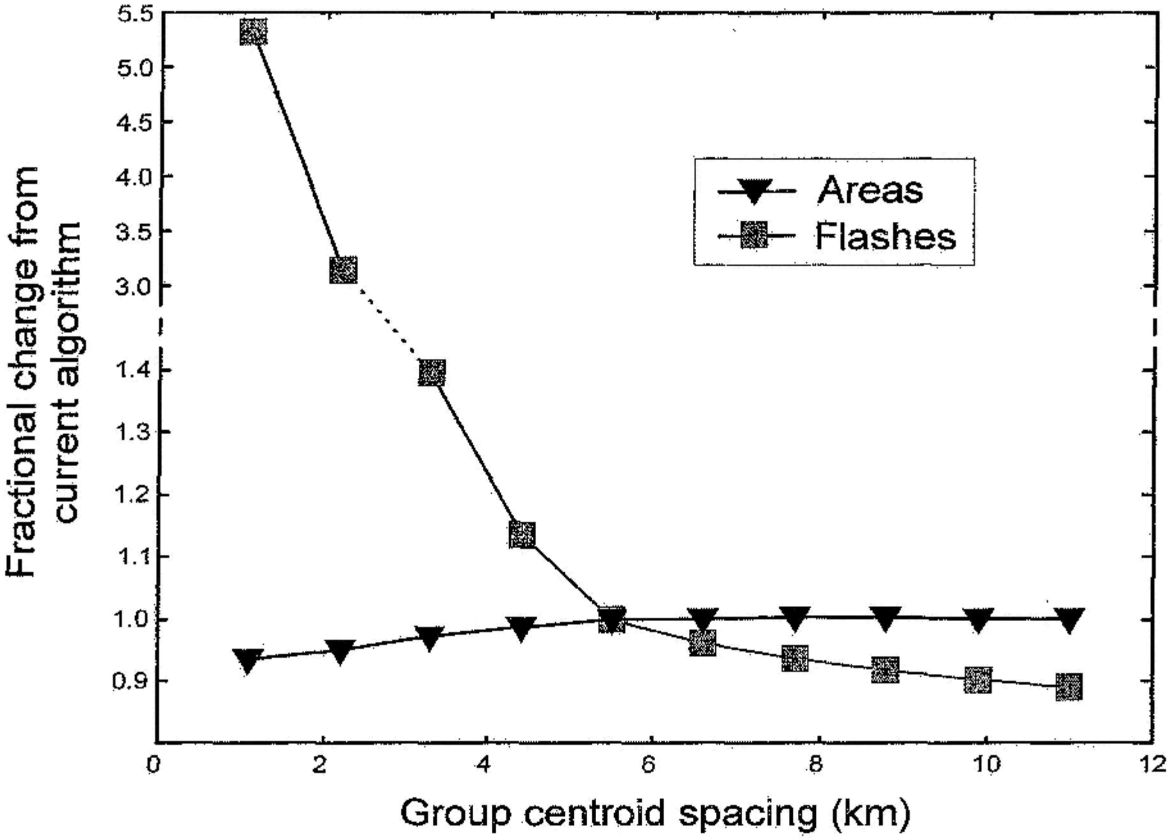


Figure 14. Fractional changes in flash and area counts when changing the group-flash spacing from 1.1 to 11 km. As long as the group-flash spatial gap is between 4 and 12 km, the change to the flash count (and global flash rate) should be less than 10%. The current LIS algorithm uses 5.5 km as the group-flash spatial gap limit. Note the gap in the vertical scale (and scale change) between 1.4 and 3.0 to show both detail in the data from 3 to 12 km and to show the extreme values below 3 km.

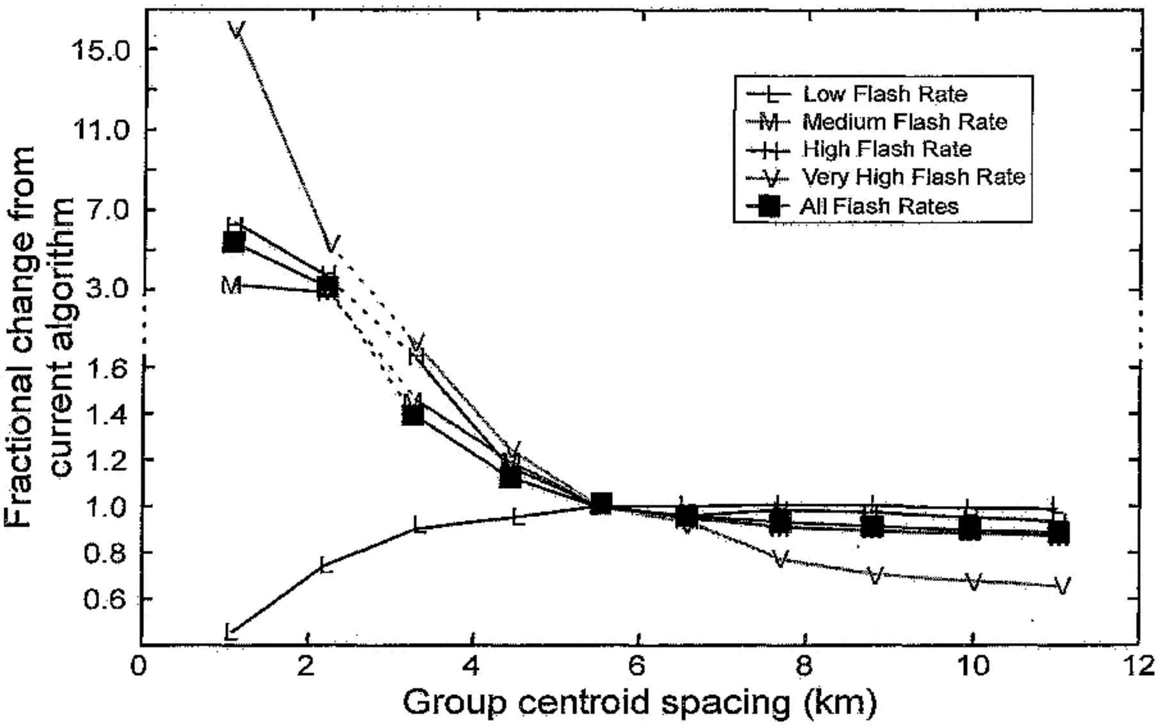


Figure 15. Fractional changes in flash counts as a function of area flash count. Note that for all flash counts except the highest (greater than 64 flashes per area), the changes to the overall flash rate was less than 20% for group-flash spatial gap limits between 4 and 12 km. Note that there is a break in the vertical scale (and change of scale) of the plot between 1.6 and 3.0. This is done to show the severe changes in flash counts due to the small allowable group spacing while at the same time show details of the much smaller changes when the flash group spacing is allowed to increase above the current 5.5 km.

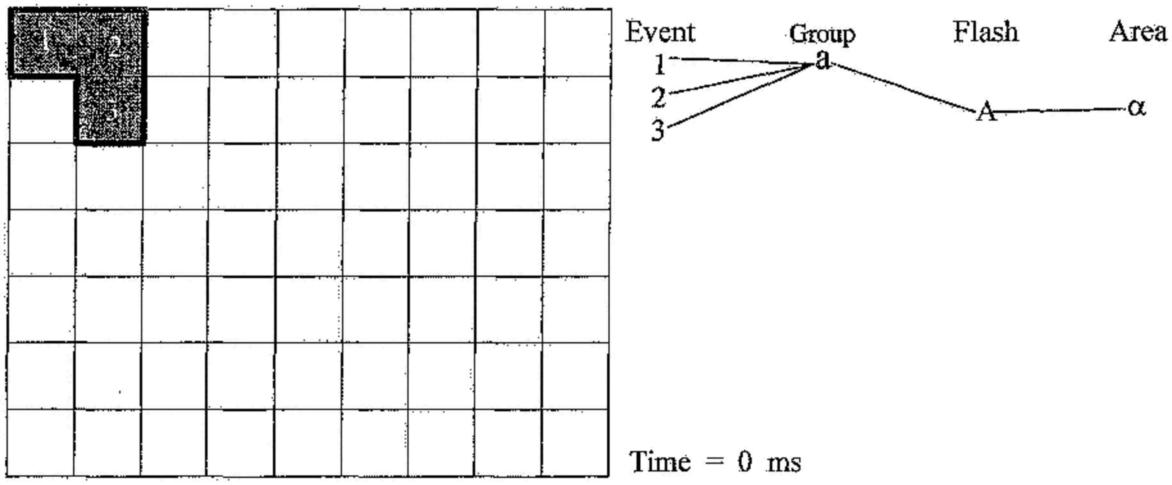


Figure 16. First three events. They are clustered into a single group, flash, and area.

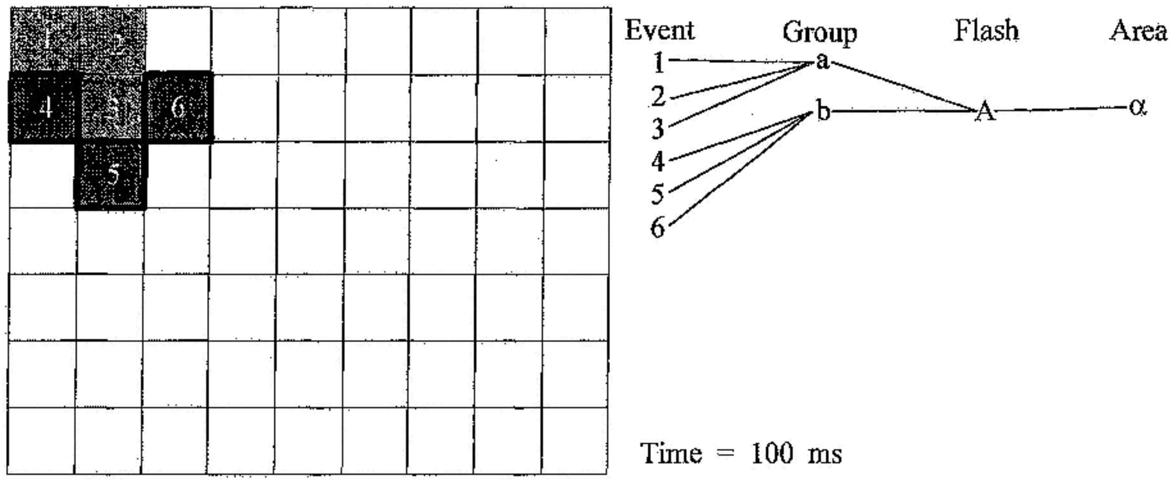


Figure 17. Second three events. They are clustered into a new group (b) but are clustered to the previous flash and area.

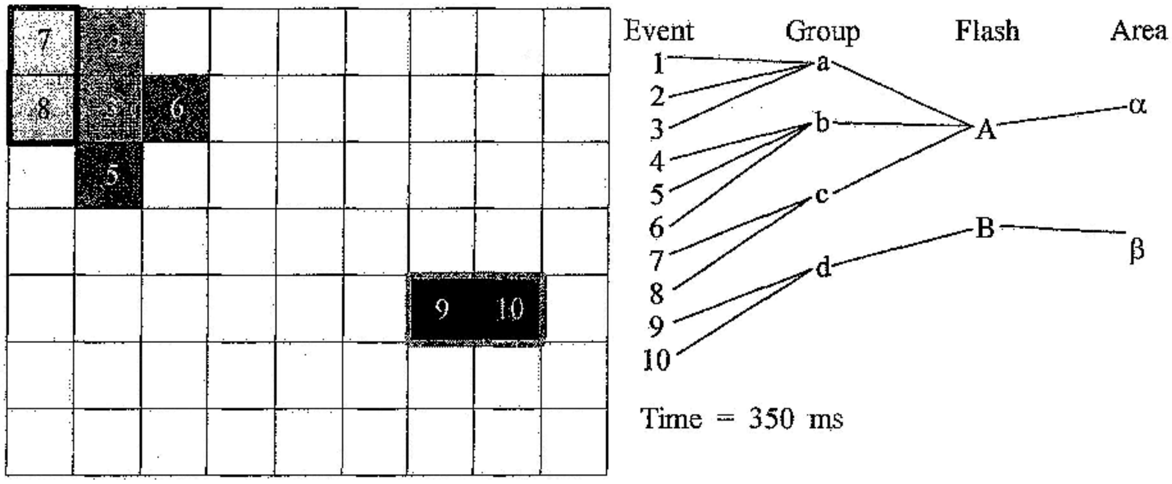


Figure 18. Next four events. Two are clustered into a new group (c) which is clustered into the previous flash and area. The other two events are clustered into a new group (d) which forms a new flash and area.

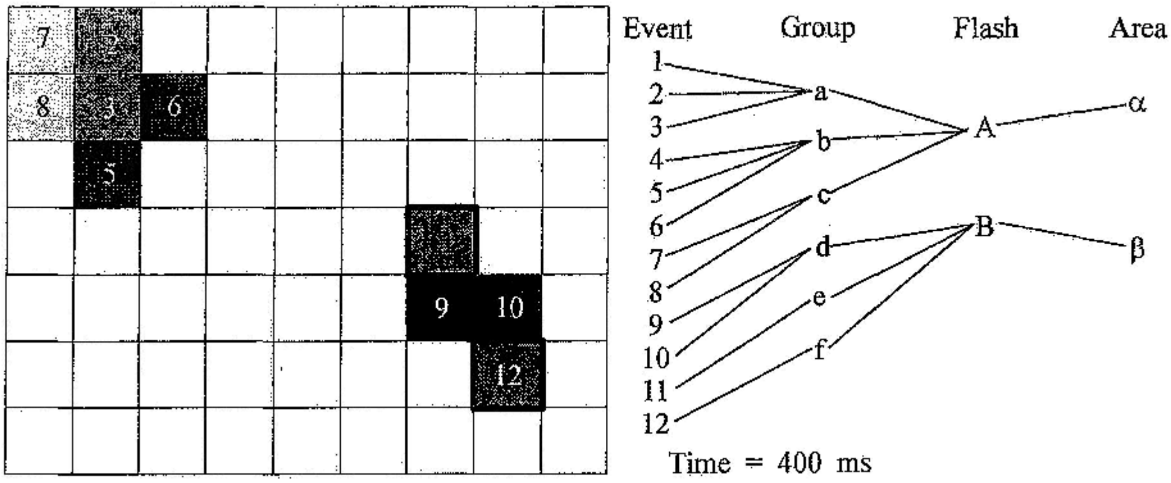
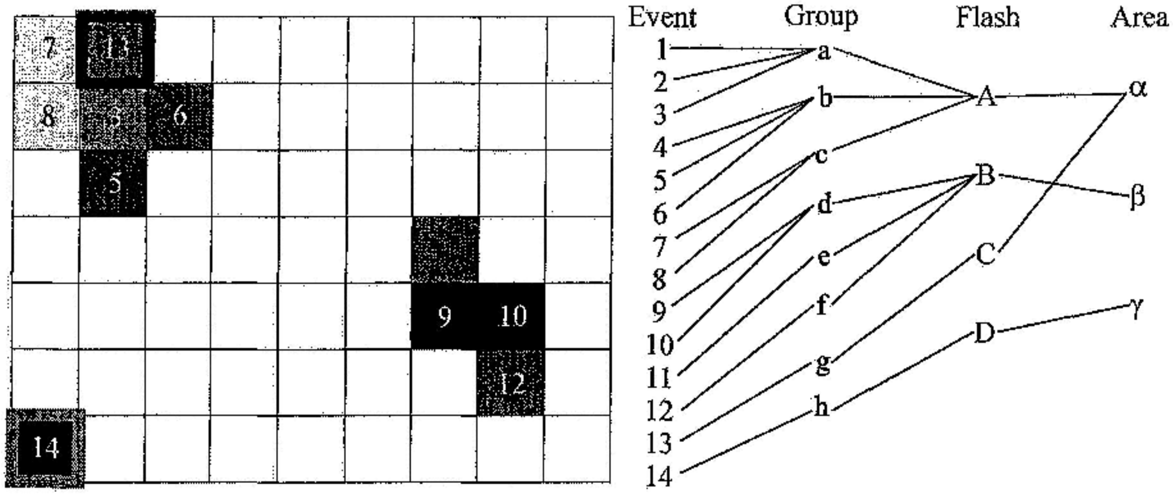


Figure 19. Two more events. Each creates a new group but both are clustered into the second flash (designated B) and area.



Time = 700 ms

Figure 20. Two final events. Each becomes a new group. The first creates a new flash (C) which clusters into the first area. The second event/group of this time frame creates a new flash and area.

Qualitative Flash Rate	Mean/Median Flash Rate (flashes/min)	Flash Counts	Events	Groups	Flashes	Areas
Low	1.1/0.7	1-3	291829	52210	4004	2641
Medium	4.8/4.3	4-15	304164	59367	5260	778
High	20/17	16-63	190041	44966	4189	152
Very High	67/51	64+	72707	18276	2052	22
Total	3.1/1.4	1-256	858741	174819	15505	3593

Table 1. Distribution of LIS test dataset as a function of flash count per area.

Area ID	Start Time	Delta Time	Event Count	Child Count	Child IDs
α	0	700	7	2	A,C
β	350	50	4	1	B
γ	700	0	1	1	D

Table 2. Data for the three areas created for this example. The data is similar to that which is stored in the LIS/OTD data files.

Flash ID	Parent ID	Start Time	Delta Time	Event Count	Child Count	Child IDs
A	α	0	350	6	3	a, b, c
B	β	350	50	4	3	d, e, f
C	α	700	0	1	1	g
D	γ	700	0	1	1	h

Table 3. Data for the four flashes created for this example. The data is similar to that which is stored in the LIS/OTD data files.

Group ID	Parent ID	Group Time	Event Count	Child Count	Child IDs
a	A	0	3	3	1, 2, 3
b	A	100	3	3	4, 5, 6
c	A	350	2	2	7, 8
d	B	350	2	2	9, 10
e	B	400	1	1	11
f	B	400	1	1	12
g	C	700	1	1	13
h	D	700	1	1	14

Table 4. Data for the eight groups created for this example. The data is similar to that which is stored in the LIS/OTD data files.