2012 22nd International Lightning Detection Conference
2 – 3 April • Broomfield, Colorado, USA
4th International Lightning Meteorology Conference
4 – 5 April • Broomfield, Colorado, USA

IDENTIFICATION OF THE MULTIPLE GROUND CONTACTS FLASHES WITH LIGHTNING LOCATION SYSTEMS

Stéphane Pédeboy
Météorage SAS, Pau, France
E-mail (corresponding author): stephane.pedeboy@meteorage.com.

1. INTRODUCTION

Most of the international and national lightning protection standards in the world recommend the use of the annual number of dangerous events for the lightning risk assessment (see IEC 62305-2). This lightning parameter can be derived from empirical equations using the thunderstorm days measured by human observers but nowadays remote sensing systems like Lightning Locations Systems (LLS) can provide high resolution lightning data leading to more reliable and accurate statistics. They tend to replace the number of thunderstorm days (N_k) parameter since it is then possible to estimate the annual numbers of dangerous events directly from the return stroke locations and compute the ground flash density (N_g) as the number of lightning flashes striking one square kilometer per year. Usually LLS detect and locate subsequent strokes before grouping them in flashes on the inter-strokes distances and time interval basis. The stroke and flash data can then provide many application and services in the field of lightning prevention or protection.

However the different ground contacts (GC) produced mostly in negative cloud-to-ground flashes are not estimated by LLS, affecting the accuracy of the risk assessment based on N_g considering only one striking point per flash. Combining the findings from several studies which were carried out over the last decades Rakov (2007) recommend applying a correction factor to N_g ranging from 1.5 to 1.7 in order to take into account the multiple striking points in flashes. These results are based on video record analyses that were ran on limited observation areas and mostly done during the intense lightning season.

The goal of this study is to develop a simple and efficient method that identifies GC in flashes in order to estimate the correcting factor on a larger data set covering a long period and area. To note that a GC identification method was suggested by Stall et al. (2009) using inter-strokes distances and characteristics like rise time, peak current and stroke order but not used on a large data set to get statistics.

This paper shows that it is possible with a statistical clustering method to achieve the goal to get some statistics on ground terminations. The method was tested with a ground truth data set to validate its efficiency. Then the French national LLS so called Météorage, was used to provide high quality lightning data. With this data, statistics on ground terminations are produced. In addition, the relative stroke location accuracy was assessed for Météorage.

2. STATISTICAL CLUSTERING METHOD

Data clustering is based on statistical data analysis aiming to group individual observation in a data set into consistent subsets, so called 'clusters', so that the observations in each cluster share similarity, most commonly depending on distances to the center of the clusters. The goal of clustering is to identify natural groups of observation or data within a large data set.

2.1 *The k-means algorithm*

Among all the clustering data analysis method, the k-means algorithms is one of the most popular because it is easy to deploy and converges rapidly to solutions. Given a data set of n points, the k-means algorithm uses a local search approach to partition the points into k clusters. To do so, a set of k initial cluster centers is chosen arbitrarily. Each point is then assigned to the closest center, and the centers are recomputed as centers of mass of their assigned points. This is repeated until the process stabilizes. It can be shown that no partition occurs twice during the course of the algorithm, and so the algorithm is guaranteed to terminate (D. Arthur et al. ,2006).

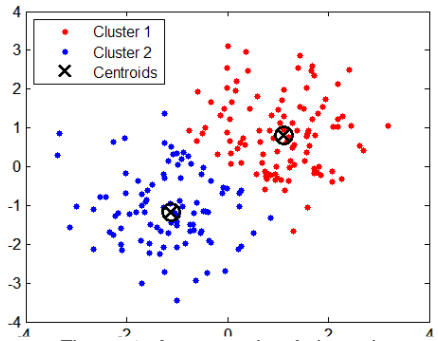

Figure 1: An example of clustering

An example based on clustering with the k-means method is presented in Fig. 1. The points representing the observations to be classified are distributed according to their coordinates. In this example, two data subsets representing clusters are

clearly identified (in red and blue) just like the human eyes would have naturally done it as a first guess. Then the points are associated with their closest cluster center and color coded. The centers of the clusters (centroids) were computed as the mean value of all the coordinates of the corresponding points and are represented by the black crosses.

2.2 *Discussion on the method*

By definition, one flash is made of n geo-localized time sorted return strokes. It can be seen as a finite dataset of n strokes (observations) to be grouped in k ground contacts (clusters) according to their proximity to the closest ground contact. Then as a first guess the k-means clustering algorithm seems to be simple to adapt to lightning data for GC identification.

However the k-means algorithm mainly relies on the distance between observations and clusters centroids. Transposed to the flash data it means the stroke location accuracy is a critical parameter for an efficient and precise clustering. It determines the capacity of the method to discriminate clusters in the cloud of the stroke observations. Also important is the stroke detection efficiency since each not located subsequent stroke may increase the probability to miss a new contact point in a flash.

Thottappillil et al. (1992) has found that the separation distances between GC in flashes range from 0.3 to 7 km, with a geometrical mean of 1.7 km. The relative location accuracy error of the newest Vaisala's remote sensing technology is estimated to be about 250 m when propagation corrections are in use (see Cummins et al., (2010). Furthermore Honma et al. (2011) showed the relative location accuracy can be decreased down to 90 m when both propagation and waveform onset time corrections are taken into account.

As a result, the newest LLS technology performance in term of location accuracy is in the same range as the new termination separation distance. In this case it is theoretically feasible to use a clustering method based on the inter-strokes distances to classify subsequent strokes in groups achieving in most of the cases a good identification of GC.

3. APPLICATION TO FLASH DATA ANALYSIS

The k-means algorithm presented above is implemented in a program that is mainly designed to handle lightning data sets collected by the French national lightning location system (LLS) or by other systems running a compatible technology. However it might be possible to use this program with any kind of system that provides high quality lightning observations.

3.1 *The lightning data*

The lightning dataset contains flash and stroke data. The subsequent strokes must have been grouped in flashes before the clustering algorithm can be applied. The most common flash grouping algorithms use the following parameters:

- The maximum inter-stroke distance that is the distance between the first stroke and a subsequent stroke (set to 10km for Météorage).
- The maximum inter-stroke interval that is the time difference between subsequent strokes (set to 500ms for Météorage).
- The maximum flash duration that is the time difference between the first and the last stroke of the sequence (set to 1s for Météorage).
- The type of the event, knowing that only Cloud-to-Ground strokes can be grouped together.

As a result each flash is made of one or several strokes. The input data for the k-means method consists of the flash id, the latitude, longitude and the semi-major axis of the 50% confidence ellipse. This parameter helps the users to rate the stroke location accuracy as described in Cummins et al. (1998).

3.2 GC identification algorithm

The application of the clustering method for GC identification has led to the development of a software that analyzes individual flashes and their subsequent strokes, grouped as described above. Several iterations are necessary to fully analyze the flash and come up with a final clustering representing ground contacts.

Iteration consists of assigning strokes to the closest ground contact in respect to the minimum geometrical distance. If this distance is greater than a maximum distance limit, the stroke creates a new ground contact. Before the iteration ends the ground contact positions are updated according to the mean locations of the strokes having being affected. A new iteration can then start and the process is repeated until the mean ground contacts positions do not vary anymore, meaning all the strokes are durably assigned to their ground contact. Then the algorithm exits.

When starting the first iteration, it is assumed that the first stroke creates the first ground contact.

The maximum distance limit is a key point in this method as it determines the distance above which strokes create new terminations. This parameter needs to be consistent with the LLS location accuracy and must be chosen carefully. If the value is too large then the number of ground contacts will be underestimated, because strokes that really belong to different contacts will be assigned to the same group. Conversely a smaller limit tends to overestimate ground contacts splitting strokes in

different clusters because of the location accuracy of the system.

It is interesting to note that the ground contact locations are determined as the mean of all their members' subsequent strokes positions. This tends to minimize the striking points location errors that may be introduced by poor stroke location accuracy. In addition, the initial assumption that consists in starting with one ground contact which position is equal to the first stroke location is not an issue since the ground contact positions are iteratively recomputed, updating the assignment of strokes to a new GC if necessary.

In order to get rid of the poorest located strokes, often strokes with small peak currents, the mean GC positions value is inversely weighted by the stroke semi-major axis parameter. With this, the location accuracy of each individual stroke is taken into account leading to a better robustness of the clustering method and therefore to a better GC location accuracy.

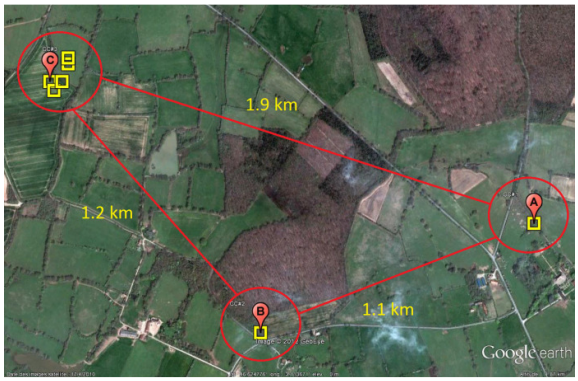

Figure 2: Example of clustered flash with 3 GC

The Fig. 2 shows an example of results after the GC identification with the k-means method was applied. This flash consist of 8 subsequent strokes represented as the yellow squares and exhibits three different GC referenced as A, B and C surrounded by the red circles. GC separation distances range from 1.1 to 1.9 km. Two GCs are made of only one stroke whereas the third one consists of the remaining 6 strokes.

This flash was extracted from the Météorage lightning database. One can note the good clustering of strokes in ground termination C showing the small size of the relative stroke locations errors. Also, the two others striking points (A and B) are clearly separated from the third one.

4. TEST AND VALIDATION OF THE METHOD

Before using the clustering tool to compute statistics on a large lightning data set, the clustering algorithm was tested on two smaller ground truth data sets. That data set consists of flashes and correlated to video records. As a result the number of ground contact was visually determined for each flash. The results obtained with the clustering program were compared to the video data results.

The analysis focused on the GC detection efficiency that is the fraction of flashes having a correct number of identified ground contacts with respect to the video observation.

4.1 *NLDN Flash data set*

All the flashes in the dataset were first correlated with data derived from video observation records collected around Tucson (USA) in 2004. As a result, a total of 39 flashes with a number of ground contacts ranging from 2 to 6 were available for analysis. The mean stroke location error is estimated to be 1000 m with a standard deviation of 1360 m by using the semi-major axis of the confidence ellipsis. The data was collected by the NLDN in the Tucson area which is at the edge of the network and the performance of this network in this region is not as good compared to the "interior" of the country (Biagi et al., 2007). Further the data set is from 2004, a time period before propagation corrections and LS technology was implemented. In order to be consistent with the location accuracy quality level the maximum distance for grouping flashes in GC is set to 1km.

The comparison of the results given by the clustering method with the video observations shows only 21 flashes have a correct number of ground contacts, leading to 54% of GC detection efficiency. It rises up to 72% when the ground contacts created by strokes with poor location quality are removed. The GC detection efficiency looks poor since the clustering algorithm managed to detect only half of the striking points, but this was expected because of the stroke location accuracy.

This is a perfect illustration of how the k-means algorithm is sensitive to LLS performance. A lot of poor stroke locations tend to mislead the algorithm creating fake ground terminations. In addition, as the maximum grouping distance had to be relaxed the discrimination is less precise and some GC (3) was grouped together instead of being separated.

4.2 *EUCLID Flash data set*

Similarly to the previous dataset, the flash data was correlated with video records collected in Austria in 2009. The data set consist of 28 cloud-to-ground flashes with ground contacts ranging from 2 to 5. The location accuracy is better than for NLDN as the mean semi major axis is equal to 550m (median is 300 m) with a standard deviation of 800m.

The analysis came up with 82% of GC detection efficiency since 23 flashes having a correct number of ground contacts were identified by the clustering method. This result looks much better compared to the previous ones obtained with NLDN because the location accuracy is higher. The observations were made in the center of the Austrian LLS where LS7000 sensors are employed. Note the waveform onset time correction was not applied at that time and so it is expected to get even a better GC

detection efficiency with lightning data observed since 2011.

4.3 *Discussion of the results*

The results are encouraging since the clustering method managed to identify up to 82% of GC in flashes in the best case. As expected the k-means algorithm is very sensitive to the stroke location accuracy. The GC identification algorithm performs much better with a dataset collected with the newest technology, limiting the number of fake clusters created by poorly located strokes.

The maximum separation distance is a parameter of the clustering algorithm used to group strokes together in GCs. It is recommended to set this parameter to the average stroke location error in order to assign correctly strokes into groups, to get the most relevant GC information possible. Then this parameter tunes the algorithm permitting to adapt the effectiveness of the lightning data sets processing according to the stroke location accuracy. However the GC discrimination might therefore lead to an underestimation of the number of ground contacts when stroke data with poor location accuracy are treated.

As a conclusion, the identification of GC flashes based on the k-means algorithm is encouraging. The stroke location accuracy has to be taken into account in order to limit the underestimation of the number of ground contacts. Improvements in remote sensing techniques tend to significantly reduce the location errors and increase the stroke detection efficiency. As a result the GC identification with such a statistical partitioning method will give better results. Another way to increase the GC detection efficiency is to use, like Stall et al. (2009), position independent stroke parameters like peak current, rise time and stroke order.

5. STATISTICAL FLASHES ANALYSIS

According to the previous test and validation results, it is possible to run an analysis on a large lightning data set to get relevant statistics on separated contact points in flashes. The dataset consists of CG flashes collected by Météorage in the center of France where the system has the best performance.

Recently the overall system performance of the French LLS was significantly improved with two major changes. In 2009 the system was upgraded with the Vaisala's LS7001 sensors model described in Cummins et al. (2011) and in July 2011 the system was configured to handle a new method for correcting the onset time of arrival measurement presented by Honma et al. (2011). Thanks to these changes it is likely the Météorage LLS observes lightning data with a very high quality.

The analyzed data period covers August and September 2011. A filtering was applied to this raw dataset in order to remove outlier flashes and the fraction of cloud-to-cloud flashes which could have been misclassified as cloud-to ground by the system because of a weakness of their peak current. As a result all flashes containing only one stroke with a semi-major axis greater than 1 km or a with peak current amplitude lower (absolute value) than -5 kA were removed from the original dataset. As a result the mean location accuracy estimated with the semi-major axis is equal to 450 m with a standard deviation of 700m. However, this value seems to underestimate the actual location accuracy since the minimum semi-major axis limit is set in the system at 400m. In addition a previous study on the relative stroke location accuracy performed in August 2011 with the PEC method (see Cummins et al (2010)) gave a value about 150 m with the new the onset time correction applied. Therefore the maximum stroke grouping distance was set to 300 m for this statistical analysis.

5.1 *The number of GC per flash*

A total of 8081 negative CG flashes was located during this two months period. They produced about 18618 subsequent strokes leading to a mean number of strokes per flash equal to 2.30.

The statistics on the number of GC per flash are presented in the Table 1 below:

Table 1: number of GC per flashes

GC per flash	Total Flashes	Total GC	% of total
1	4292	4292	53%
2	2299	4598	28%
3	949	2847	12%
4	418	1672	5%
5	93	465	1%
6	23	138	0%
7	7	49	0%
TOTAL	8081	14061	

The first column represents the ground contacts per flash that ranges from 1 to 7 ground terminations. The corresponding number of flashes in the next column shows a decreasing as the GC multiplicity rises. Most of the flashes have only one ground termination (53%), this proportion being roughly divided by a factor 2 as the number of GC increases. It is interesting to note that 47% of the CG flashes strike the ground in more than one place. Finally a total of 14061 ground contacts were counted by the clustering method leading to an average of 1.74 ground contacts per flash.

These results are in pretty good agreement with several statistics from video records observations which can be found in literature. Saba et al. (2006) observed in southeastern Brazil 51% of flashes exhibiting more than one ground contact with an average number of striking points per flash of 1.70. Rakov et al. (1994) came up with a similar result, about 50% of flashes with an average of 1.67 striking points per flash in Florida, as well as Kitagawa et al. (1962) in New Mexico that found a proportion of 49% of flashes producing more than one ground contact.

As a remark this result might suffer from a seasonal effect as the analyzed data period is summer 2011. The reason for the small period is that in July the new onset time correction was taken into operation It seems reasonable to run the same analysis on a whole year period in order to check if the GC multiplicity is seasonal or not.

5.2 *The separation distance*

The separation distances between ground terminations computed with the clustering method was also analyzed. For each flash having two or more separated striking points, the distance between the two closest GC was computed for each pair of terminations in the flash. The figure below shows the distribution of striking points distances.

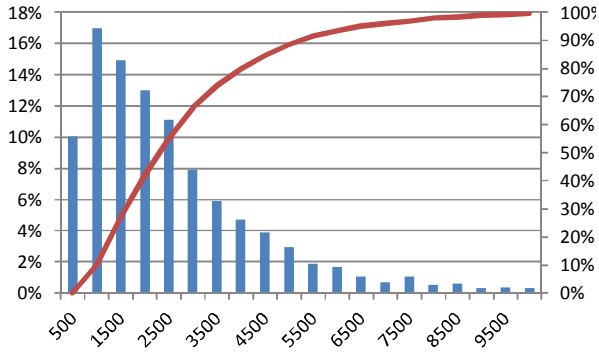

Figure 3: Distribution of the GC separation distances in GC

About 3800 flashes were analyzed leading to separation distance ranging from 300 m up to 9.9 km with a mean value of 2.2 km and a standard deviation of 1.8 km. Of course the lower limit depends on the maximum stroke grouping distance fixed in the algorithm (300 m), and the upper limit of the 9.9 km depends on the maximum inter-strokes distance used in the flash grouping algorithm.

One can note separation distances around 1000 m are the most likely to be observed. The decreasing is fairly continuous and does not show any secondary peak. This seems to indicate that the spatial distribution of the ground terminations is homogeneous. The small peak at 7500 m can be explained by a lack of data at this range. Finally, half of the distances are less or equal to 1.8 km that is in very good agreement with Thottappillil et al. (1992) who measured a geometrical mean of 1.7 km.

5.3 Stroke relative location accuracy

The relative location accuracy is a parameter that determines the repeatability of the measuring system in locating lightning strikes. Applied to the strokes locations, the relative location accuracy can be estimated with the pre-existing channels (PEC) terminations assuming all the return strokes in a PEC strike the ground at the same place. A method based on PEC was first described by Cummins et al. (2010) and used to assess the relative location accuracy. This method relies on a PEC property that states the strokes in flashes with an order greater than 5 are not likely to create a new GC. A limitation of this method is it must deals with flashes having a minimum strokes multiplicity of 5 that are not so numerous in nature. In addition it is known that the return stroke peak current tends to decrease as the stroke order increases making them difficult to detect with LLS.

The clustering algorithm used in this study can identity GC, from which PEC can be derived regardless of the stroke order. Distances between stroke locations and GC positions are computed for each individual stroke in multi-stroke termination points. The analyzed data set consists of a total of 1597 contact points. The data set is the same as in Table 1, excluding the GC composed of only one stroke since in this case the location of the stroke is the same as the ground termination.

The amount of data distributed per GC order is presented in Table 2:

Table 2: Number of multi-stroke GC as a function of terminations order in flashes

GC#1	GC#2	GC#3	GC#4	GC#5	GC#6	GC#7
863	420	220	85	4	3	2

It is interesting to note that the number of multiple stroke GC decreases rapidly as a function of the GC order.

In order to estimate the relative stroke locations errors, it is assumed that the position given to GC by the k-means method is the ground truth as it results from the average of all the members' strokes positions.

Fig. 4 shows the distribution of the median values of these errors with respect to the striking point order. The values on top of the bars are the median errors in meters.

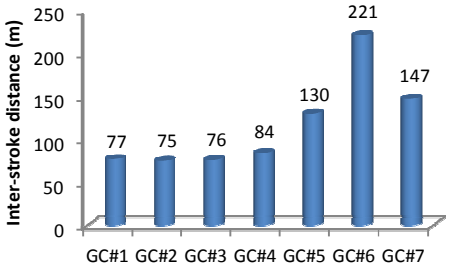

Figure 4: distribution of the inter-stroke distances in GC

The median errors are quite stable for the first fourth GC orders, ranging from 75 m to 84 m. Errors for GC with order greater or equal to 5, increase significantly up to 221 m. The reason for this is mainly the small number of GC in these bins.

However this method introduces a bias leading to an underestimation of the relative stroke location error. Since stroke positions are also used in the GC mean position calculation the distance between strokes and GC positions are both dependent on the location errors. In order to take into account this dependency the computed relative stroke location error is increased by a conservative factor of 35%. This value computed for GC consisting of two strokes decreases as the GC stroke multiplicity increases (e.g. 23% for a 3 strokes GC). As a result one can consider the relative median stroke location error value is about not bigger than 110m for the Météorage LLS what is consistent with the results by Cummins (2010) and Honma (2011). It is interesting to note the propagation corrections, not yet in use in France, should improve this parameter.

Note that this is an analysis of the relative location accuracy. It is important to understand it is different from the absolute location accuracy itself since it does not show the systematic bias that might affect the absolute location accuracy errors. However it is a good parameter to assess LLS performance.

5. CONCLUSION

The annual lightning flash density N_g is an important parameter used by the lightning protection community in the risk assessment. Today it is usually based on flash data set collected by LLS that do not determine the multiple ground contacts in flashes, leading to an underestimation of the actual risk.

Thanks to improvements in LLS performances it is possible to determine a correcting factor, compensating for multiple ground terminations in flashes from the lightning data itself. A new method using the k–means algorithm was developed and tested to classify and assign return strokes to different clusters defining separated ground contacts. Despite the fact that this method relies mainly on the performances of the LLS, it is shown that if the method is applied to data produced with the current state of sensor technology the results are quite good. The method was tested with lightning data sets correlated with video records acting as ground truth to check the ability of the k-means algorithm to determine the multiple GC in flashes. The results are quite encouraging since up to 82% of the GC flashes could be identified depending on the stroke location accuracy of the system collecting the data.

Then a statistical analysis was performed on a large dataset of observations from summer 2011 by the French national LLS that is a system with high performance. The average number of ground terminations for CG flashes was found to be about 1.74 and the fraction of the flashes striking the ground in more than one place is about 47%. These results are in quite a good agreement with those from several other authors that ran video records observations in some countries, despite the possible errors in the GC accounting due to some poorly located strokes that are badly assigned, either creating a fake new contact point or joining an existing termination.

With the deployment of the latest remote sensing technology it will become possible to apply the method to a highly accurate data set and then to tune the program more efficiently to get better estimation of the number of ground contacts. However this method relies significantly on location accuracy and it is necessary to find some parameters which do not depend on this LLS parameter, maybe the electromagnetic waveforms characteristics could be used in conjunction with the distance to even increase the accuracy of ground contact estimates.

This analysis also provides a way to estimate the relative stroke location accuracy of the Météorage's LLS to be about 110 m, confirming already published results.

This study has shown the capability of this method to identify the different contact points in flashes. It offers the possibility to study the GC flashes as a function of season and or region making available some statistics for large periods and areas. The future work will consist of comparing the results of this clustering method to video correlated flash data observed with systems using propagation and onset time measurement corrections. In addition, it is necessary to analyze a whole year period of data before stating on the final correcting factor.

6. ACKNOWLEGMENTS

I would like to warmly thank Ken Cummins from the University of Arizona and Wolfgang Schulz (ALDIS) for their review and contribution of lightning data and their comments that helped a lot in the tuning and validation of the k-means algorithm presented in this paper.

5. REFERENCES

Arthur D. and S. Vassilvitskii, 2006: How Slow is the k Means Method?. Standford University

Biagi, C. J.; Cummins, K. L.; Kehoe, K. E.; Krider,E. P., National Lightning Detection Network (NLDN) performance in southern Arizona, Texas, and Oklahoma in 2003-2004. Journal of Geophysical Research, Volume 112, IssueD5, 2007.

Cummins, K. L., M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer (1998), A Combined TOA/MDF Technology Upgrade of the U.S. National Lightning Detection Network, *J. Geophys. Res.*, *103*(D8), 9035–9044, doi:10.1029/98JD00153.

Cummins K. L., M. J. Murphy, J. A. Cramer, W. Scheftic, N. Demitriades, A. Nag, 2010: Location accuracy improvements using propagation corrections: A case study of the US NLDN, 21st ILDC.

Cummins, K.L., N. Honma, A.E. Pifer, T. Rogers, M. Tatsumi (2011), Improved detection of winter lightning in the Tohoku region of Japan using Vaisala's LS700x Technology, 3rd Intl. Symp. on Winter Lightning, Sapporo, Japan, June15-16, 2011.

Honna H., K. L. Cummins, M. J. Murphy, A.E. Pifer, T. Rogers, 2011: Improved lightning locations in the Tokohu region of Japan using propagation and waveform onset corrections, 3rd International Symposium on Winter Lightning (ISLW)

Kitagawa N., M. Brook, E. J. Workman, 1962: Continuing currents in cloud-to-ground lightning discharges, JGR, 67, 637-647.

Rakov V. A., M. A. Uman, and R. Thottappillil, 1994: Review of lightning properties from electric field and TV observations, JGR, 99, 745-750.

Rakov V.A. (2007), Lightning phenomenology and parameters important for lightning protection, IX SIPDA, 26th-30th November 2007.

Saba M. F., M. G. Ballarotti and O. Pinto Jr, 2006: Negative cloud-to-ground lightning properties from high speed video observations, JGR, vol 111, D03101.

Stall C, K.L. Cummins, E.P. Krider, J. Cramer (2009), Detecting Multiple Ground Contacts in Cloud-to-Ground Lightning Flashes. Journal of Atmospheric & Oceanic Technology, vol. 26 (11), pp. 2392-2402, November 2009.

Thottappillil R., V. A. Rakov, M. Uman, W. Beasley, M. Master and D. Shelukhin, 1992: Lightning subsequent stroke electric field peak greater than the first stroke peak and multiple ground terminations, JGR, 97(D7), 7503-7509.