Approbation of automatic allocation of thunderstorm foci in the south of Western Siberia according to WWLLN data according to the FRIS-Tax clustering algorithm

S Yu Karanina¹, M Yu Belikova¹, N A Kocheeva¹ and A V Karanin¹

¹ Gorno-Altaisk State University, 1 Lenkin street, Gorno-Altaisk 649000, Altai Republic, Russia

kreichetova@yandex.ru, belikovamy@yandex.ru, nina_kocheewa@mail.ru, vedmedk@bk.ru

Abstract. The article shows the ability of the FRiS-Tax algorithm to separate lightning discharge clusters and combine them into classes. Lightning discharge clusters are a model of an individual lightning cell. Classes (groups) of clusters of lightning discharges that are close in space and in time of action can be associated with a model of a multi-cell thunderstorm or a thunderstorm focus. Clustering of lightning discharges is carried out using normalization values specifying the linear dimensions of a thunderstorm cell (ρ*) and its lifetime (τ*). The experimental sample consisted of 3934 lightning discharges registered by WWLLN in the south of Western Siberia, August 7–8, 2017. The results of lightning discharge clustering are compared for two parameterization conditions: ρ₁*=50 km, τ₁*=30 min and ρ₂*=15 km, τ₂*=20 min. For the first set of parameterization conditions, the resulting lightning discharge clusters represent zones of possible activity of a thunderstorm cell for one to two hours. Classes formed from these clusters describe the behavior of long lasting and vast thunderstorm foci. The clusters obtained for the second set of parameterization conditions better correspond to the linear sizes and lifetime of individual thunderstorm cells. An adjustment of the clustering algorithm is required to combine such clusters into classes corresponding to a multi-cell thunderstorm.

1. Introduction
To solve individual tasks of monitoring of lightning activity [1-3] and the fundamental tasks of the Global Electric Circuit [4] it is necessary to distinguish groups of lightning discharges characterized by temporal and spatial proximity. Such groups of lightning discharges are associated with a convective cell in an active thunderstorm state (thunderstorm cell) [5]. A thunderstorm cell is defined as an indivisible element of a thunder cloud. Thunderstorms consisting of a single convective cell are called a single-cell thunderstorm. If a thunderstorm consists of several cells moving as a whole, then it is called a multi-cell thunderstorm. Multi-cell thunderstorms are also called thunderstorm foci. The dimensions of a single-cell thunder cloud vary from 5 to 20 km. The life cycle of a convective thunderstorm cell at the stage of maturity takes from 20 minutes to 1 hour.

The linear dimensions of a multi-cell thunder cloud vary from 10 to 1000 km. At the same time, thunderstorm cells in such a cloud exist independently and can be in different stages of development. As a rule, after the appearance of the first thunderstorm cell and its maturity stage, one or more daughter cells are initiated within 15-30 minutes at 20-30 km from it. It is also possible that new...
thunderstorm cells closely link to the existing ones. The subsequent emergence of new cells leads to the fact that the multi-cell thunderstorm can last several hours.

The grouping of lightning discharges, simulating the allocation of thunderstorm cells, is performed using cluster analysis. The grouping is made with the help of well-known algorithms with a ready-made software implementation, for example, k-means, DBSCAN, grid algorithms, etc. [1-4]. On the global scale, works on the clustering of lightning discharges registered by WWLLN (Word Wide Lightning Location Network) August 7-8, 2017 in the south of Western Siberia (68–90° E, 48–58° N). When choosing the time, weather conditions were taken into account that contributed to the formation of thunderstorms in the area of the study [7].

2. Materials and methods

2.1 Clustering algorithm

One of the clustering algorithms that allow to select clusters of arbitrary shape and determine their hierarchical structure is the FRiS-Tax algorithm [6], based on the use of the Function of Rival Similarity FRiS-function [8].

The grouping of data on lightning discharges is based on the following idea. If a discharge arises “close” from an already existing cluster, then it belongs to this cluster. Otherwise, it is the first discharge of the new cluster. The "proximity" of lightning discharges is usually determined by means of threshold values.

Let the results of the registration of lightning discharges be represented as the vector \( \mathbf{o}=(x,y,t) \), where \( x \) is the north latitude, \( y \) is the east longitude, and \( t \) is the time of registration of the discharge.

It is believed that lightning discharges \( \mathbf{o}_1=(x_1,y_1,t_1) \) and \( \mathbf{o}_2=(x_2,y_2,t_2) \) are “close” if the distance between them \( \rho(\mathbf{o}_1, \mathbf{o}_2) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2} \) is less than some given value \( \rho^* \) and the difference in recording time \( \tau(\mathbf{o}_1, \mathbf{o}_2) = |t_1-t_2| \) does not exceed some value \( \tau^* \). The values of \( \rho^* \) and \( \tau^* \) determine the clustering parameters and set the thunderstorm-neighborhood \( \epsilon_{\rho\tau} \). A thunderstorm is “close” to cluster \( C \), if in its neighborhood \( \epsilon_{\rho\tau} \) there are discharges of this cluster and the discharge is “close” to the center of the cluster, called the stolp (the concentration center of the thunderstorm).

Let the objects of the two already “existing” clusters \( C_1 \) and \( C_2 \) be in the neighborhood \( \epsilon_{\rho\tau} \) of the thunderstorm discharge. For clusters \( C_1 \) and \( C_2 \), the stolps are defined, respectively \( s_1 \) and \( s_2 \). Discharge \( \mathbf{o} \) is close to both \( C_1 \) cluster objects and \( C_2 \) cluster objects. Clusters \( C_1 \) and \( C_2 \) rival for the \( \mathbf{o} \) object. The decision on whether this object belongs to one of the rivaling clusters is made on the basis of the values of the FRiS function:

\[
F_{1/2}(\mathbf{o}) = \frac{\rho(\mathbf{o}, s_1) - \rho(\mathbf{o}, s_2)}{\rho(\mathbf{o}, s_1) + \rho(\mathbf{o}, s_2)}
\] (1)

The values of the function \( F_{1/2}(\mathbf{o}) \) vary from -1 to +1. If \( F_{1/2}(\mathbf{o}) \) close to +1, then \( \mathbf{o} \) belongs to cluster \( C_1 \). If the value is close to -1, then \( \mathbf{o} \) falls into the rival cluster \( C_2 \). If \( \rho(\mathbf{o}, s_1) = \rho(\mathbf{o}, s_2) \) and respectively \( F_{1/2}(\mathbf{o}) = F_{2/1}(\mathbf{o}) \), then the object is located on the border between the clusters.

The stolp of cluster \( C \) is determined based on a modified FRiS-function calculated for all objects of cluster \( C \) and the so-called virtual cluster, all objects of which are at a fixed distance \( \rho^* \) from the objects of cluster \( C \). Each object of cluster \( C \) is sequentially considered to be the stolp \( s \). For the actual object \( s \) and all other objects \( \mathbf{o} \), the sum of the values of the FRiS-function is calculated:

\[
F^*_{\rho}(\mathbf{o}) = \frac{\rho(\mathbf{o}, s) - \rho^*}{\rho(\mathbf{o}, s) + \rho^*}
\] (2)
The object for which the sum of the FRiS functions is maximal is defined as a stolp. The distance between object \( o \) and cluster \( C \) is equal to the distance from object \( o \) to pillar \( s \) of this cluster.

Also, if the objects of two already “existing” clusters \( C_1 \) and \( C_2 \) are in the neighborhood \( \varepsilon_{\text{pr}} \) of the thunderstorm discharge \( o \), then the clusters can be included in one class. Clusters are combined into a class by determining the area of their rivalry and estimating the distance between the clusters. For clusters \( C_1 \) and \( C_2 \), there is a rival zone in which the absolute value of the FRiS function of objects is less than a certain threshold \( |F_{ij}(o)| < \rho^* \). The distance between clusters \( C_1 \) and \( C_2 \) is estimated as follows. In the competition zone, two objects \( a \) are selected from \( C_1 \) and \( b \) from \( C_2 \), with the minimum distance. This distance is designated as \( D_{12} \). From each of the objects \( a \) from \( C_1 \) and \( b \) from \( C_2 \), the distances \( D_a \) and \( D_b \) to the nearest neighboring object from the corresponding cluster are calculated. Clusters \( C_1 \) and \( C_2 \) are considered to belong to the same class if the value of the three quantities \( D_{12} \), \( D_a \) and \( D_b \) charges, which implements the construction of clusters using the FRiS function and the procedure for combining clusters into classes, consists of the following steps.

Thus, the algorithm for grouping lightning discharges, which implements the construction of clusters and their combination into classes using the FRiS function, consists of the following steps.

1. A priori parameters \( \rho^* \) and \( \tau^* \) are specified. The sample of discharges is ordered by time.
2. The first discharge is the stolp of the first cluster.
3. For each of the following discharge conditions are checked:
   a. if the neighborhood \( \varepsilon_{\text{pr}} \) is empty, then the discharge is the stolp of the new cluster;
   b. if in the neighborhood \( \varepsilon_{\text{pr}} \) there are discharges that belong to the same cluster, then the discharge is added to this cluster; a new stolp is selected from the cluster objects;
   c. if the neighborhood \( \varepsilon_{\text{pr}} \) contains discharges that belong to different rival clusters, then the two closest stolps of these clusters are determined and the decision on whether the discharge belongs to one of the rival clusters is made on the basis of the value of the FRiS function calculated by equation (1); in the cluster to which a new discharge was added, a new stolp is selected by equation (2); the condition is checked to combine rivaling clusters into a class.

A separate cluster of lightning discharges selected as a result of the algorithm execution is a thunderstorm cell model. Combined set of clusters (a class) may be associated with a thunderstorm focus or a multi-cell thunderstorm (thunderstorm focus). The class may consist of one cluster, in this case we can talk about a single-cell thunderstorm.

2.2 Data on lightning discharges
WWLLN is a global multi-point lightning detection system. At present, WWLLN consists of several dozen stations (over 70) located in different countries of the world [9]. One of the receiving points of this network is located in Gorno-Altaisk (Altai Republic) and has been operating since 2015. A lightning discharge is detected by the system if its signal is detected by five or more station.

For the south of Western Siberia, it has been established that the cold front and the front of occlusion are most conducive to thunderstorm formation [7]. The occurrence of multi-cell thunderstorms is most often associated with the passage of fronts [5]. To test cluster analysis in order to identify thunderstorm cells and their groups determining multi-cell thunderstorms, August 7-8, 2017 were selected, when a network of weather stations in the south of Western Siberia predicted and observed baric formations and associated thunderstorms. On the indicated days from 7 am (local time) on August 7 and until 7 am on August 8 the WWLLN network registered more than 3,000 lightning discharges, more than 80% of which originated in the territory of Altai Republic and in the northeast Kazakhstan. At the same time, the difference in time between two successive lightning discharges of the entire sample did not exceed 12 minutes.

In the computational experiment, we adopted two conditions for the parameterization of a cluster of lightning discharges: \( \rho^*_1=50 \text{ km} \), \( \tau^*_1=30 \text{ min} \) [3] and \( \rho^*_2=15 \text{ km} \), \( \tau^*_2=20 \text{ min} \) [4]. For each cluster, the start time (the time of the first lightning discharge assigned to this cluster) and the end time (the registration
time of the last discharge included in the cluster), and the duration of the cluster’s existence (the difference between the start and end times) were calculated. In addition, for each cluster, the number of discharges and the area were determined (based on the method of convex hulls, built on the thunderstorm discharges of the cluster).

For each combination of two or more clusters united in a class, the time of existence and area were also calculated. The duration of the class was the difference between the time of the last and the first lightning discharge. The class area was determined similarly to the cluster area, taking into account all discharges belonging to the class.

3. Results and discussion

In an experimental sample of 3984 lightning discharges, the number of clusters with 2 or more discharges was: 249 for the parameters $\rho_1^* = 50$ km, $\tau_1^* = 30$ min; 1068 for the parameters $\rho_2^* = 15$ km, $\tau_2^* = 20$ min. At the same time, for the parametrization conditions $\rho_1^* = 50$ km, $\tau_1^* = 30$ min, 65 discharges were defined as “noise” (a cluster consisting of one discharge), which accounted for 2% of the total experimental sample. For parameterization $\rho_2^* = 15$ km, $\tau_2^* = 20$ min, the “noise” share turned out to be much higher - 500 discharges (13%).

Table 1 presents the statistical characteristics of the duration of existence and the area of clusters of lightning discharges for the parametrization conditions used.

With parameterization $\rho_1^* = 50$ km, $\tau_1^* = 30$ min in half of the cases, the lifetime of the clusters varies from one to two hours. Thus, the increase in the linear dimensions of a thunderstorm cell, adopted in the work [3] increase in linear dimensions of a thunderstorm cell to 50 km, due to its possible displacement during 30 minutes of existence by 20-30 km, leads to the selection using cluster analysis of longer time and larger areas of thunderstorm activity. These areas of lightning activity can be interpreted as areas of possible thunderstorm cell activity within one to two hours.

With parametrization $\rho_2^* = 15$ km, $\tau_2^* = 20$ min, the cluster lifetime does not exceed 40 minutes in 90% of cases. The median duration of one cluster is 12 minutes, the lower and upper quartiles are 5 and 24 minutes (Table 1). The area of clusters is in good agreement with the linear dimensions of the thunderstorm cell, in 88% of cases it does not exceed 400 square kilometers. In general, there is a better consistency of the clusters allocated in this case to the characteristics of thunderstorm cells [5]. It is important to note that the sizes of the main areas of discharges in a thundercloud are smaller than the size of the cloud itself.

Table 1. Statistical characteristics of clusters with 2 or more lightning discharges

| Parameter | Threshold distance and time between discharges |
|-----------|---------------------------------------------|
|           | 50 (km) and 30 (minutes) | 15 (km) and 20 (minutes) |
| Duration (minutes) | | |
| Mean | 62.6 | 17.7 |
| Median | 45.0 | 12.7 |
| minimum – maximum | 0.0-256.9 | 0.0-112.8 |
| lower quartile (25%) - upper quartile (75%) | 19.9-97.9 | 5.3-24.1 |
| Area (square km) | | |
| Mean | 2871 | 177 |
| Median | 1224 | 61 |
| minimum - maximum | 2-34559 | 0-2687 |
| lower quartile (25%) - upper quartile (75%) | 374-4338 | 20-186 |

However, it is worth considering that the error in determining the coordinates of lightning discharges varies widely, and on the average it is 5 km. This causes the expansion of the parametrization boundary conditions and reduces the reliability of the grouping of lightning discharges. The minimum time interval between discharges equal to 20 minutes coincides with the time of the existence of a thunderstorm cell in a mature state. However, lightning discharges can occur
during its formation and decay. At low threshold linear dimensions and the duration of the existence of clusters of lightning discharges, a significant noise level was established. A separate thunderstorm discharge in this case may represent a thunderstorm cell.

It is important to note that the considered lightning discharges on August 7-8 occurred in the conditions of passage of various frontal systems in which multi-cell thunderstorms are most often formed. Therefore, the assessment of the ability of the FRIS-Tax algorithm to allocate classes of clusters of thunderstorm discharges corresponding to multi-cell thunderstorm is of particular importance.

For the clustering parameters under consideration, a different number of classes was allocated: 119 classes for $\rho_1^*=50$ km, $\tau_1^*=30$ min and 469 classes for $\rho_2^*=15$ km, $\tau_2^*=20$ min. For these variants of the parametrization conditions, about 80% of classes consist of one cluster of lightning discharges (figure 1). In general, the reason for the weak integration of clusters of lightning discharges into classes is the low percentage of registration of lightning discharges by the WWLLN network. Those discharges that could be "connecting" for clusters, the network does not detect.

Table 2 presents the statistical characteristics of the duration and area of classes consisting of two or more clusters. They are in good agreement with the time of existence and the linear dimensions of the multi-cell thundercloud, multi-cluster classes, which were distinguished by parameterization $\rho_1^*=50$ km, $\tau_1^*=30$ min. Under these conditions, the duration of classes is mainly several hours, the area varies within thousands and tens of thousands of square kilometers. These indicators are significantly lower for classes obtained at $\rho_2^*=15$ km, $\tau_2^*=20$ min.

Thus, the parametrization conditions $\rho_2^*=15$ km, $\tau_2^*=20$ min are not suitable for distinguishing classes that approximately simulate a multi-cell thunderstorm. To merge two neighboring clusters into a class, the same minimum distance and time parameters are used as for cluster allocation. Consequently, clusters will be included in one class in which the distance from the objects belonging to their intersection zone to other objects of the class does not exceed 15 km and the time does not exceed 20 minutes. Therefore, for such conditions of parametrization, an additional refinement of the clustering algorithm is required to change the threshold distance in order to combine clusters into classes at least up to 40 km. In this regard, for the formation of classes, the parametrization conditions $\rho_1^*=50$ km, $\tau_1^*=30$ min take precedence. Since in this case the possible distance between the
thunderstorm cells is already taken into account equal to 20-30 km and the possible bearing error of the coordinates of the lightning discharge.

Table 2. Parameters of classes consisting of 2 or more clusters of lightning discharge

| Parameter                        | Threshold distance and time between discharges |
|----------------------------------|-----------------------------------------------|
|                                  | 50 (km) and 30 (minutes) | 15 (km) and 20 (minutes) |
| Number of classes                | 16                                             | 71                                             |
| Duration (minutes)               | Mean: 226,2                                    | 47,2                                           |
|                                  | Median: 170,1                                  | 39,6                                           |
|                                  | Minimum – maximum: 32,2-545,7                   | 4,0-132,6                                      |
|                                  | Lower quartile (25%) - upper quartile (75%): 120,5-354,0 | 22,2-65,1                                      |
| Total area (square km)           | Mean: 28374                                    | 922                                           |
|                                  | Median: 16240                                  | 488                                           |
|                                  | Minimum – maximum: 1602-82096                  | 32-6011                                       |
|                                  | Lower quartile (25%) - upper quartile (75%): 4594-40856 | 243-1142                                      |

In general, the sequence of the occurrence of clusters within the selected classes $\rho_1 *=50$ km, $\tau_1 *=30$ min agrees well with the behavior of long-lasting thunderstorm foci, which description was made for Yakutia [11, 2]. In the group of clusters, combined by FRIS-Tax algorithm into one class, the difference in the start time of the clusters varies from 5 to 40 minutes. Predominantly clusters start 20-40 minutes before the end of the previous clusters. Almost simultaneous start of separate clusters located at a distance of more than 50 km from each other is noted. There are the longest clusters that last from 2 to 4 hours. Clusters of smaller duration and intensity appear on the periphery of these long-lasting and intense lightning activity clusters (thunderstorm foci) during their life cycle.

4. Conclusion

On the basis of traditional weather observations data, the time interval with the most striking manifestation of thunderstorms was selected for the territory of the southeastern part of Western Siberia and northeastern Kazakhstan. During the study of thunderstorm activity in a limited space and at a certain time, two sets of parameters for the clustering of lightning discharges were tested. The features of the reflection of the space-time distribution of lightning activity were established for each set of parameters used.

The clusters obtained by parametrization $\rho_2 *=15$ km, $\tau_2 *=20$ min can be considered a lower estimate of the size and lifetime of thunderstorm cells, but real thunderstorm cells can be large. To merge such cells into classes, it is necessary to additionally set a threshold value for the cluster rival area of 30–40 km. In our opinion, in this case it is possible to obtain classes that are better consistent with the multicell thunderstorm model. To reflect a more complete picture of the development and nature of the behavior of thunderstorm cells, it is further necessary to take into account in the cast of the classes also individual lightning discharges, distinguished by the clustering algorithm as noise.

Parameterization $\rho_1 *=50$ km, $\tau_1 *=30$ min leads to the pointing of longer in time and larger in areas thunderstorm activity fields in comparison to the description of thunderstorm cells. These areas of lightning activity can be interpreted as areas of possible thunderstorm cell activity within 1-2 hours. A good agreement with the nature of the emergence and development of thunderstorm foci was obtained for the classes of the clusters. This confirms the ability of the clustering method to separate clusters and their classes that well describe the phenomenon under study.

The cluster analysis of lightning discharges once again confirms the substantial dependence of the results on the parametrization conditions - the maximum distance and time between two discharges for the cluster. A comparative analysis of two different methods of parameterization allowed us to
determine the direction of further research. To obtain a more adequate thunderstorm cell model when clustering WWLLN data, it is necessary to adjust FRIS-Tax algorithm and to conduct a series of additional computational experiments with other parameterization conditions.

References

[1] Adzhieva A A and Shapovalov V A 2016 Cluster analysis in automatic detection and maintenance of thunderstorm foci according to the data of the thunderstorm-direction network Engineering journal of Don 2 pp 164-71
[2] Kononov I I and Yusupov I E 2004 Cluster analysis of storm activity Journal of communications technology and electronics vol 49 3 pp 283-91
[3] Shabaganova S N, Karimov R R, Kozlov V I and Mullayarov V A 2012 Characteristics of storm cells from observations in Yakutia Russian Meteorology and Hydrology 37 pp 746-51
[4] Hutchins M L, Holzworth R H and Brundell J B 2014 Diurnal variation of the global electric circuit from clustered thunderstorms. Journal of Geophysical Research: Space Physics vol 119 1 pp 620-9
[5] Matveev L T 1984 General meteorology course. Atmospheric physics (Leningrad: Hydrometeoizdat) p 750
[6] Borisova I A 2007 FRiS-Tax algorithm of taxonomy Scientific Bulletin of NSTU 3 (Novosibirsk: NSTU Publ) pp 3-12
[7] Kocheeva N A, Chankibaeva M H, Miniaev A I, Sukhova M G and Modorov A A 2017 Peculiarities of Thunderstorms’ Occurrences on the Border of the Western Siberia Plains and Altai Mountains Journal of Environmental Management & Tourism vol 8 5(21) pp 1061-8
[8] Zagoruiiko N G 2008 Intellectual data analysis based on a rival similarity function Optoelectronics, Instrumentation and Data Processing vol 44 pp 211-217
[9] World Wide Lightning Location Network (wwlln.net)
[10] Kozlov V I, Mullayarov V A and Karimov R R 2011 The spatial distribution of the lightning in the east of Russia on remote sensing data Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa vol 8 3 pp 257-62
[11] Kozlov V I, Mullayarov V A and Vasil’ev A E 2003 Characteristics of thunderstorm cores in Yakutia in 1993-2001 from instrumental observations Russian Meteorology and Hydrology 2 pp 27-31

Acknowledgments

The authors wish to thank the World Wide Lightning Location Network (http://wwlln.net), a collaboration among over 70 universities and institutions, for providing the lightning location data used in this paper.