Estimation of –CG lightning distances using single-station E-field measurements and machine learning techniques

Adonis. F. R. Leal,
UFPA
Belém, Brazil
adonisleal1@gmail.com

V. A. Rakov
University of Florida
Gainesville, USA

Elton Rafael Alves
UNIFESSPA
Marabá, Brazil

Márcio N. G. Lopes
CENSIPAM
Belém, Brazil

***Abstract*— Machine learning (ML) techniques have been used around the world to solve different problems. In this work, we applied ML techniques to estimate lightning distance for return strokes (RS) in negative cloud-to-ground (-CG) flashes. The approach uses E-field records from a single-station system. A lightning electric field waveform dataset containing more than 1500 waveforms of negative RS recorded at LOG, Florida, US was used to train and validate the ML classifiers. The dataset was split into day time and night time records. For day-time records, the quadratic Support Vector Machine (SVM) classifier was the one with the best accuracy (80%) and for night-time, the best one was the linear SVM with an accuracy of 88%. The ML classifiers were applied to estimate lightning distance in thunderstorms in the Amazon region of Brazil, and the results were compared against GOES-16 images and STARNET lightning location data. The main application of such methodology is for regions with no lightning location systems or no communication links to obtain lightning location data.**

Keywords*— *Machine learning; lightning distance; E-field

I. INTRODUCTION

Lightning can be defined as a transient high-current (typically tens of kiloamperes) discharge that crosses the air. Frequently, those high-current discharges touch the ground and can represent a risk to human life or equipment. Modern lightning location systems (LLSs) use electric and magnetic fields radiated by lightning to geolocate lightning discharges in real-time, hence one can use this information for thunderstorm now-casting and forecasting, lightning warning, lightning risk assessment, lightning research, and other application.

In the absence of LLSs, or in absence of communication links to obtain LLSs data, other solutions for lightning warning or for lightning research are needed. In some parts of the Amazon region of Brazil, establishing communication links is a hard task, mostly because of the rainforest and the lack of roads. Sometimes the only communication link is made via satellites, so that, it is very expensive or impracticable. Therefore, single-station systems without the need of communication link are needed for such locations.

In the present work, an approach to estimate the lightning distance for negative cloud-to-ground (-CG) flashes using measurements of lightning electric fields at a single station and machine learning techniques, is developed and tested.

A. Negative return-stroke electric field waveform characteristics

Lightning electric field waveforms have been extensively examined in the past decades and consequently, many features have been revealed. The lightning electric field waveforms in the time domain changes with the lightning type, lightning intensity (peak current) and the distance between the measuring point and the lightning location.

Lin et al. [1], for instance, examined in detail wideband electric field waveforms in the time domain produced by the negative cloud to ground (-CG) return strokes (RS) in Florida at distances ranging from 1 to 200 km. Figure 1 shows the characteristic waveforms of first and subsequent negative CG stroke for distances ranging from 10 to 200 km.

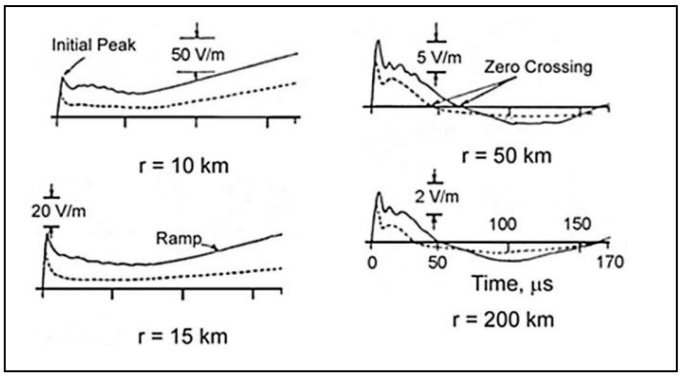

Fig. 1. Typical vertical electric field intensity waveforms for first (solid line) and subsequent (dashed line) return strokes at distances of 10, 15, 50, and 200 km. Adapted from Lin et al. [1].

For close distances, within a few kilometers, the electric fields of RS in CG flashes are dominated by the electrostatic component of the total electric field, which appears as a ramp after the initial peak. The initial field peak is the dominant feature of the electric field waveforms beyond about 10 km. The initial field peak is primarily due to the radiation component of the total field. The features of the field waveforms from relatively close lightning are determined primarily by the characteristics of the lightning, as opposed to propagation effects.

Lightning electric field waveforms observed at ranges from a few hundred kilometers to tens of thousands of kilometers are significantly influenced by the propagation of field in the Earth-ionosphere cavity (atmospheric waveguide) by way of multiple reflections from the ionosphere and the Earth's surface.

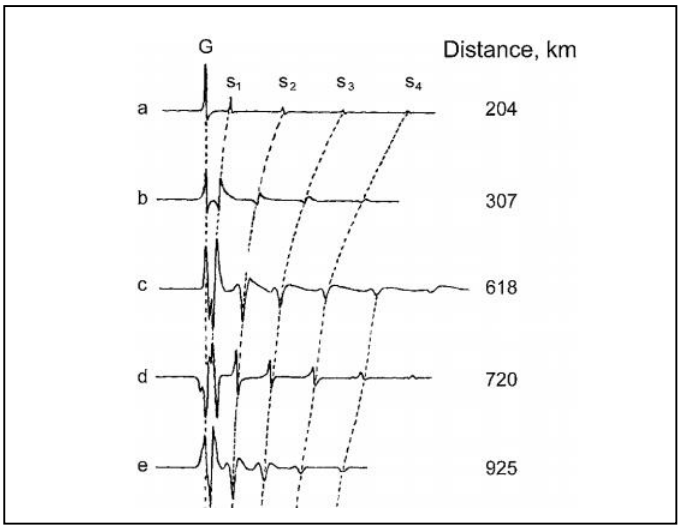

Fig. 2. Typical vertical electric field waveforms for return strokes recorded at far distances ranging from 204 to 925 km. G indicates the position of the ground wave and S1, S2, S3, and S4 indicate the skywaves or reflections. Adapted from Schonland [2].

The multiple reflections from the ionosphere and the Earth's surface can be identified in the electric field waveforms by a sequence of skywaves (individual waveforms) after the ground wave (see Fig. 2). As the distance increases, the skywaves appear to become closer to the groundwave and eventually appear to overlap. The same behavior is observed in both day-time and night-time records, however, the distance between the peaks (individual waveforms) are different, and the peaks are sharper for night-time records. These differences are mainly due to the ionosphere properties at day and night times.

In summary, the same lightning event (RS in negative CG flash) can have different representative electric field waveforms depending on the distance at which the field was recorded and whether the field was recorded at day or night time conditions.

B. Single-station lightning location techniques

There are different methods for locating lightning from their radiated electromagnetic fields. The most common multiple-station techniques include the magnetic direction finding (MDF), time of arrival (TOA), and interferometry [3]. In all of these techniques, more than one station is needed in addition to communication links between the stations and a central processor unit.

There are also a few approaches to locating lightning (or thunderstorm) based on single-station measurements of electromagnetic fields. Single-station lightning location systems are used mostly for research [4] and usually combine MDF with a technique to estimate the distance of the source.

One approach to estimate the lightning distance based on single-station electric field measurement is described in [5] and [6]. Their approach is based on the analyses of the time interval between the multiple peaks in the lightning electric field waveform due to multiple reflections between the ionosphere and the ground. Most of these approaches are reliable only for far lightning when one can record the skywaves.

According to [7], it is possible to locate close lightning with single-station measurements based on the spatial and frequency dependences of the ratio of their electric and magnetic fields.

Other approaches to locating lightning with a single station are described by [8]–[13].

C. Classification algorithms (Supervised Machine Learning)

Supervised learning, in the context of machine learning, is a type of algorithm in which both input (observations or examples) and known responses to the data (i.e., labels or classes) are provided. Input and responses data are labeled and used to train a model that generates predictions for the response to new data.

Considering a dataset with lightning electric field waveforms in which each waveform is labeled with a distance between the measuring point and the lightning channel. The aim of machine learning algorithms in this work is to find a decision border that separates the lightning electric field waveforms in different classes. Each class corresponds to a distance between the measuring point and the lightning channel. The algorithms will use the discreet waveforms in the time domain as predictors. Each set of predictors is labeled with a distance that corresponds to the response.

Different machine learning algorithms can find different decision borders. In the present work, three machine learning models/techniques were applied: a distance-based method (k-NN); a search-based method (decision tree) and an optimization method (support vector machine - SVM).

k-nearest neighbors (k-NN)

The machine learning algorithm k-NN employs a supervised method for data classification using as the main criterion the shorter distance between neighbors in the feature space [14]. The number of nearest neighbors is the key to the performance of the classifier [15]. The k-NN goal is forming a generalization based on the training set in order to elevating the accuracy of classification for the new data. This algorithm assumes that the training set is composed of descriptive variables and their classifications. It uses such variables to classify a new object.

Assuming a training set formed by n observations and their classes. The distance can be calculated between the new observation and each point in the training set. Based on the calculated distance, the nearest k neighbors are identified. Next, based on the chosen k, a new observation is attributed to the class with the greater number of observations in k.

Decision tree

Decision trees are employed on a dataset to predict a feature target based on input features [16]. Induction process of decision trees aims at recursively portioning a training set until each subset contains cases of a single class.

In a classification tree, either the root node or the treetop is the starting point formed by the whole learning set, that is, the input features. A node is a subset of the set of attributes and it might be terminal or non-terminal. A non-terminal node breaks down into other nodes named children nodes. That division is determined by a condition on the value of a single attribute according to which the examples are divided into different nodes.

A non-divisible node is a terminal node to which a class is attributed. Each example in the set falls onto a terminal node.

Support vector machine – SVM

The idea of SVM algorithms are to create a line or a hyper plane which separates the data into classes. It is possible to employ SVMs to obtain linear frontiers in the classification of the linearly separable dataset that has near-linear distribution [17]. However, there are cases in which the use of a non-linear border is more adequate to separate classes. SVMs deal with non-linear problems by mapping the training set from its original finite-dimensional space, referred to as input, onto a much higher-dimensional space called feature space.

The SVM perform such mapping by using kernel functions. Some of the most frequently used kernels are polynomials, radial base and sigmoid. Choosing an SVM then involves choosing the kernel function.

II. DATA AND METHODOLOGY

We have used two datasets in this work, one was used to train the classification models (dataset-1), and the other was composed of new data feed into the models to estimate unknown distances of RS in negative CG flashes (dataset-2).

The electric field waveforms in both datasets were obtained using the Lightning Detection and Waveform Storage System (LDWSS) [18], [19]. The LDWSS has a bandwidth from 160 Hz to 500 kHz, the decay time constant of 1 ms, the time resolution (sampling interval) of 1 μs, and the digitizing system has an RTC (Real Time Clock) which is synchronized with a GPS module.

Machine learning algorithms need standard data as input to generate a prediction model. In order to standardize the lightning waveforms in the datasets, we choose two parameters: time window and normalized ground wave peak. Here, each RS electric field waveform, in the datasets, is displayed in a time window of 1.1 ms, with its peak at 100 us, and all the waveforms were normalized relative to the peak amplitude of the ground-wave. Thus, all events have a ground-wave peak equal to 1. The reference to field polarity of –CGs in this paper is based on the atmospheric electricity sign convention, according to which a downward-directed electric field (or electric field change) vector is assumed to be positive. Thus, -CG waveforms have positive E-field polarity.

The dataset-1 is composed by RS electric field waveforms recorded at the Lightning Observatory in Gainesville (LOG) [20], Florida, on August 2016. The RS electric field waveforms in dataset-1 were grouped in distance ranges centered at 20, 50, 100, 150, 200, 250, 300, 350, 400, 450 and 490 km with a bin size of ± 5 km relative to the center distance. The distances from LOG to the lightning channels were determined using NLDN (U.S. National Lightning Detection Network) data. See a few examples of RS electric field waveforms from dataset-1 in Fig. 3.

NLDN has undergone many upgrades since its beginning, 1994-1995 [21], 2003-2004 [22], and August 2013 [23]. After the most recent upgrade, the median location error is estimated to be approximately 200-300 m inside the network [23]. These errors are acceptable for the ± 5 km bin size.

In dataset-1, besides the distance ranges, the data were grouped into Day-time and Night-time categories. The summary of dataset-1 is given in Table 1.

TABLE I. NUMBER OF EVENTS IN DIFFERENT DISTANCE RANGES IN DATASET-1

Day-Time records												
Center of distance range, km	20	50	100	150	200	250	300	350	400	450	490	**20 - 490**
Sample size	36	284	230	231	115	61	109	91	53	26	19	**1255**
Night-Time records												
Center of distance range, km	20	50	100	150	200	250	300	350	400	450	490	**20 - 490**
Sample size	25	54	81	67	30	13	10	-	15	5	10	**310**

Each distance range is ± 5 km of its center.

Note that for night-time records we do not have any records in the 350 ± 5 km range and the sample size for night-time is smaller than for day-time.

The dataset-2 is composed by negative RS electric field waveforms recorded in the Amazon region of Brazil. Currently, two LDWSS stations are installed in the Amazon region, one in Belem at the Federal University of Para (UFPA), and the other in Maraba at the Federal University of the Southern and Southeastern Para (UNIFESSPA), Brazil. The stations are 440 km from each other (see Fig. 4).

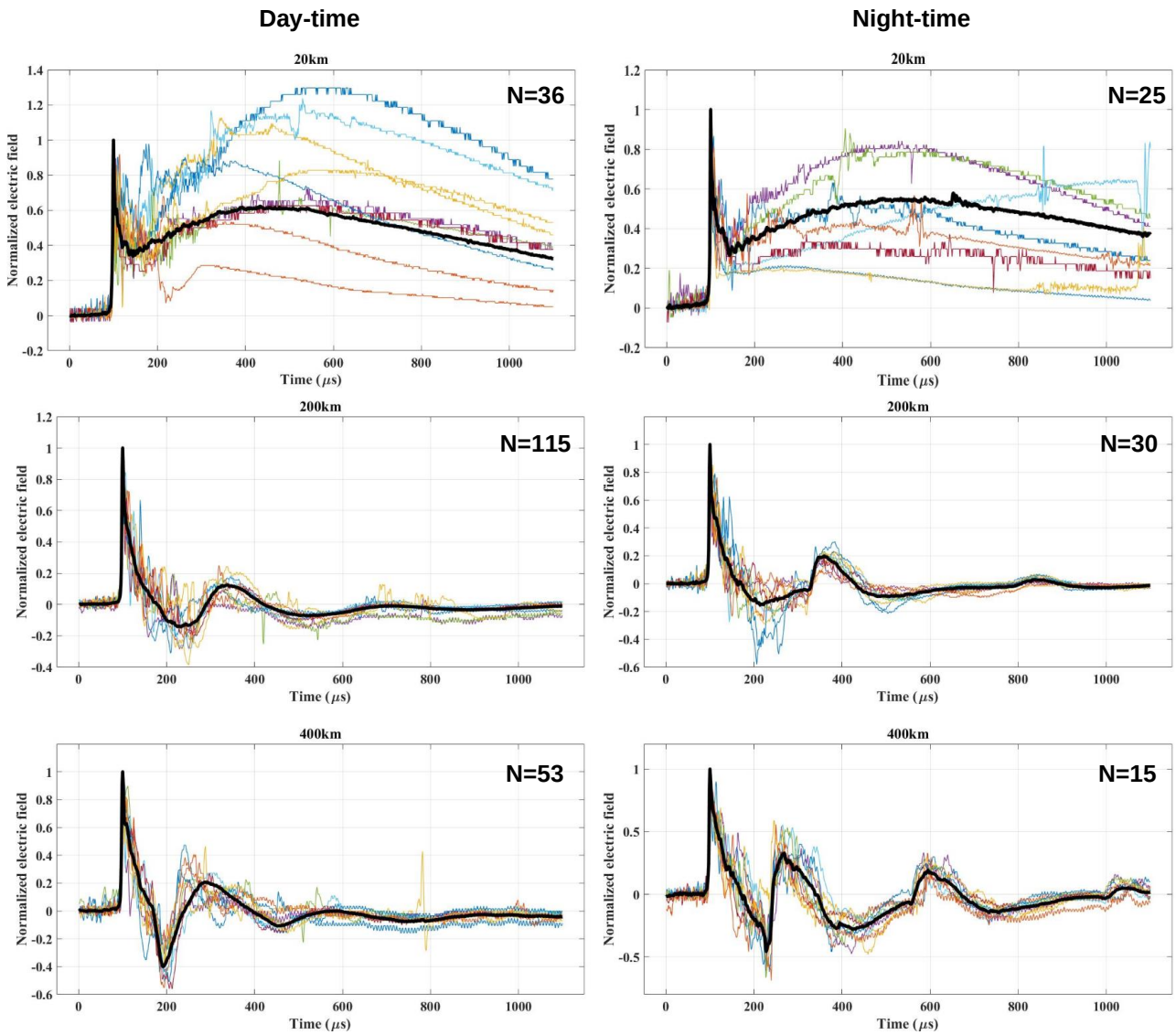

Fig. 3. Examples of electric field waveforms from dataset-1 for ranges of 20, 200 and 400 km for day time and night time conditions. The fine colored lines are the individual waveforms. The thick black line corresponds to the mean waveform that was obtained by averaging individual waveforms. N is the sample size.

Fig. 4. Currently locations of the LDWSS stations in the Amazon region.

We choose data from both stations to build dataset-2. From UNIFESSPA-station, we had 28 events being 11 recorded at daytime and 17 at nighttime on December 04, 2018 and December 05, 2018. From UFPA-station, we choose 18 events from a close thunderstorm that occurred at daytime on December 11, 2017. The same thunderstorm was also discussed in [24].

Because of the differences between lightning electric field waveforms recorded at day-time and night-time, we applied the same methodology for both (day-time dataset-1) and (night-time dataset-1) in order to obtain two distinct models for estimate the distances.

The machine learning algorithms training were performed with Matlab toolbox “Classification learner”. In the classification, learner toolbox Decision Tree technique is divided in Fine, Medium and Coarse. The KNN is divided in Fine, Medium, Coarse, Cosine, Cubic, and Weighted. The SVM is divided among Linear, Quadratic, Cubic, Fine

Gaussian, Medium Gaussian, and Coarse Gaussian. We applied the dataset to all of these 15 machine learning techniques in order to find more accurate classifier.

The input data is composed of predictors and responses. The 1100 samples of each waveform (1.1 ms time window) were the predictors, and the distance range labels (20, 50, 100, 150, 200, 250, 300, 350, 400, 450 and 490 km) were the responses.

We held out 30% of the data for use in the validation, that is, 70% of the dataset was used to train the algorithms, and 30% was used for testing the resultant classification model.

The summary of the methodology is presented in the Fig.5.

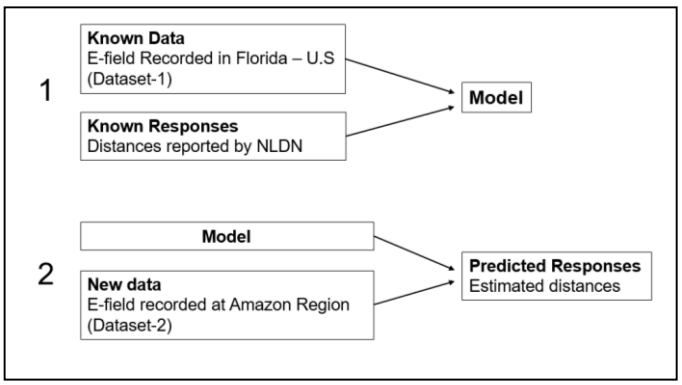

Fig. 5. Methodology flow chart. (1) Procedure to train the machine-learning model. (2) Procedure to estimate the –CG lightning distances.

III. RESULTS

After applying the dataset-1 to the 15 machine learning techniques, and validating each model with 30% validation data, we compare their accuracy. The 15 machine learning techniques were used for both day and night-time data. The summary of the results of the best classifier for each machine model technique is shown in Table 2.

TABLE II. BEST CLASSIFIER FOR EACH MACHINE-LEARNING MODEL TECHNIQUE

	Day-time		Night-time	
	Best Classifier	Accuracy	Best Classifier	Accuracy
Decision Tree	Fine Tree	58%	Fine Tree	69%
	Medium Tree		Medium Tree	
KNN	Weigthed KNN	68%	Fine KNN	82%
SVM	Quadratic SVM	80%	Linear SVM	88%

According to Table 2, SVM algorithms presented a better performance than k-NN and decision tree algorithms. SVM can handle high-dimensional data due to its convex optimization of the problem [25]. In addition, it can map the data using different kernel functions. Thus, it was noted that the optimization-based method (SVM) was more efficient than methods based on distance (k-NN) and search methods (decision tree).

In order to understand how the more accurate classifier performed in each class, we used the confusion matrix plot. The confusion matrix helps us to identify the areas where the classifier has performed poorly.

In the confusion matrix plot, the rows show the true class, and the columns show the predicted class. The confusion matrix was calculated using the 30% validation data. The diagonal cells show where the true class and the predicted class match. If these cells are green, the classifier has performed well. The number of observations is shown in each cell.

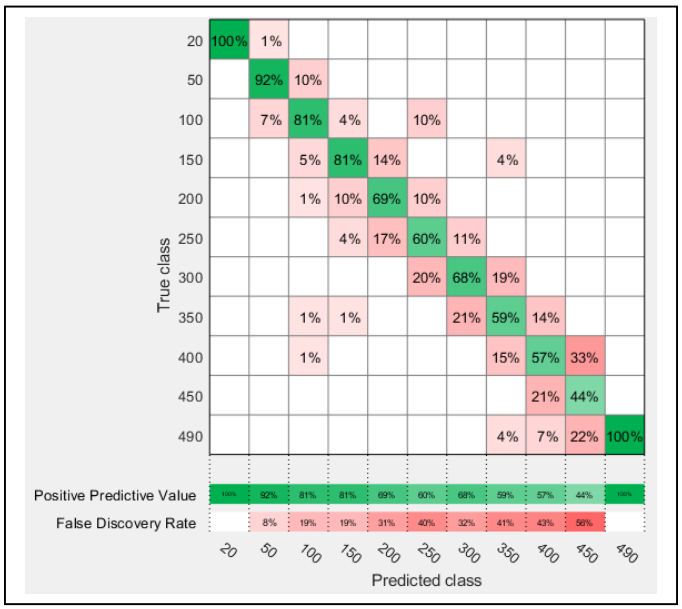

Fig. 6. Confusion matrix plot for Quadratic SVM classifier for **Day-time** records in sub-dataset-1. The sample sizes of validation data (30%) for each class are: 20 km, N= 9; 50 km, N=86; 100 km, N=73; 150 km, N=73; 200 km, N=36; 250 km, N=10; 300 km, N=38; 350 km, N=27; 400 km, N=14; 450 km, N=9; 490 km, N=1.

Fig. 7. Confusion matrix plot for Linear SVM classifier for **Night-time** records in sub-dataset-1. The sample sizes of validation data (30%) for each class are: 20 km, N= 6; 50 km, N=15; 100 km, N=28; 150 km, N=21; 200 km, N=9; 250 km, N=4; 300 km, N=2; 400 km, N=5; 450 km, N=1; 490 km, N=2.

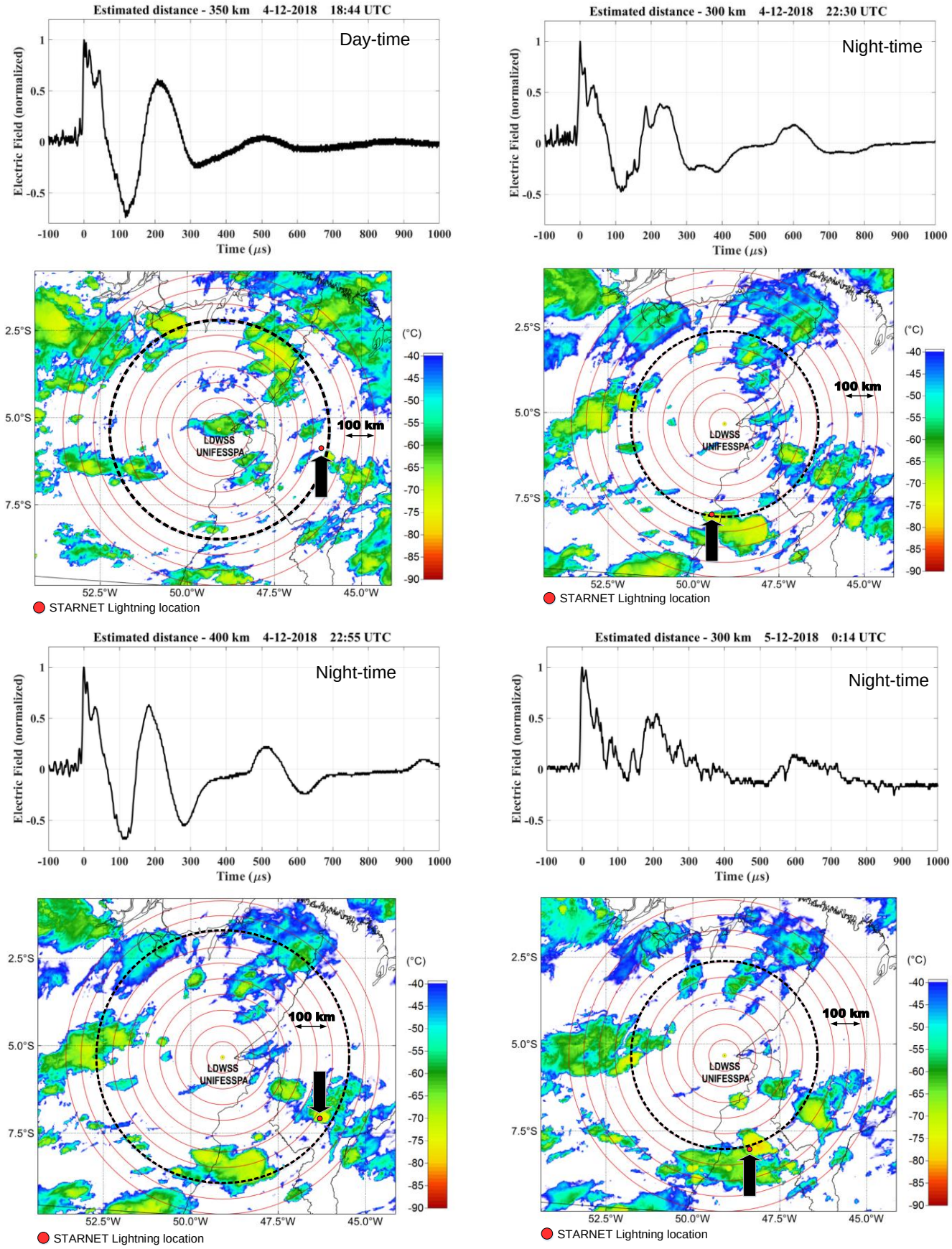

Fig. 8. Examples of negative RS E-field waveforms recorded at the LDWSS-UNIFESSPA station. The estimated distance is shown at the top of each waveform panel. Shown below each waveform is the infrared images from the GOES-16 satellite and STARNET lightning location. The black dashed-line circle is centered at LDWSS location and has radius equal to the estimated distance.

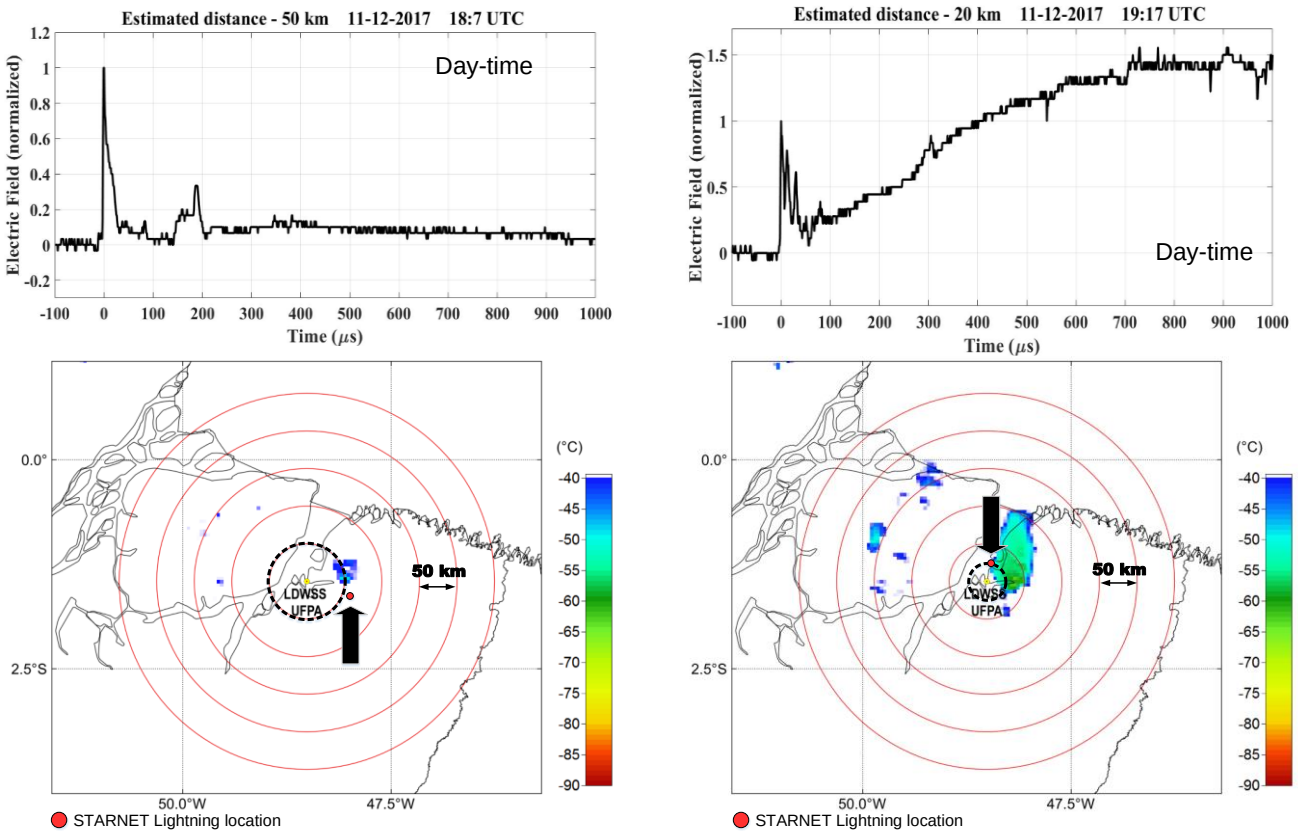

Fig. 9. Examples of negative RS E-field waveforms recorded at the LDWSS-UFPA station. The estimated distance is shown at the top of each waveform panel. Shown below each waveform is the infrared images from the GOES-13 satellite and STARNET lightning location. The black dashed-line circle is centered at LDWSS location and has radius equal to the estimated distance.

Both quadratic and linear SVM models for Day-time and Night-time records were used to estimate –CG lightning distances in dataset-2. In order to check how the estimation was performing, we verified the cloud condition at the time where the electric field was recorded. The cloud conditions were verified with the infrared channel images from the GOES-16 satellite.

In Figures 8 and 9, it is shown the electric field waveforms recorded in the Amazon region (Marabá and Belém), the respective estimated lightning distance using the machine learning model (Day-time and Night-time), the infrared channel images from GOES-16 satellite at the same time of the recorded electric fields and STARNET lightning location.

As noted in Figures 8 and 9 the estimated distances are very close to the STARNET location.

IV. Summary

A methodology for estimating –CG lightning distance from single-station E-field measurements and machine learning techniques was presented. For day-time records, the quadratic SVM classifier was the one with the best accuracy (80%) and for night-time, the best one was the linear SVM with an accuracy of 88%. Both classifiers were used to estimate –CG lightning distances from electric fields recorded in the Amazon region of Brazil. The results showed good agreement with the local thunderstorms condition retrieved from the infrared cloud images from the GOES-16 satellite. The main application of such methodology is for regions with no LLS or no communication systems to get global LLS data or any other source of lightning data. In addition, the system is low-cost and is able to work stand-alone.

Acknowledgment

The authors would like to thank R. L. Holle and W. A. Brooks of Vaisala Inc. for providing the NLDN data, and the STORM-T / IAG / USP Laboratory and USP for the availability and processing of STARNET data.

References

[1] Y. T. Lin et al., “Characterization of lightning return stroke electric and magnetic fields from simultaneous two-station measurements,” J. Geophys. Res., vol. 84, no. C10, p. 6307, 1979.

[2] B. F. J. Schonland, J. S. Elder, D. B. Hodges, W. E. Phillips, and J. W. van Wyk, “The Wave Form of Atmospherics at Night,” Proc. R. Soc. A Math. Phys. Eng. Sci., vol. 176, no. 965, pp. 180–202, Oct. 1940.

[3] V. A. Rakov, Fundamentals of Lightning. New York: Cambridge University Press, 2016.

[4] V. A. Rakov and M. A. Uman, Lightning: Physics and Effects. New York: Cambridge University Press, 2003.
[5] V. Ramachandran, J. N. Prakash, A. Deo, and S. Kumar, "Lightning stroke distance estimation from single station observation and validation with WWLLN data," in Annales Geophysicae, 2007, vol. 25, no. 7, pp. 1509–1517.
[6] I. Nagano, S. Yagitani, M. Ozaki, Y. Nakamura, and K. Miyamura, "Estimation of lightning location from single station observations of sferics," Electron. Commun. Japan (Part I Commun., vol. 90, no. 1, pp. 25–34, Jan. 2007.
[7] M. Chen, T. Lu, and Y. Du, "Experimental study of single-station lightning locating technique," in 2011 7th Asia-Pacific International Conference on Lightning, 2011, pp. 59–64.
[8] E. T. Pierce, "Some techniques for locating thunderstorms from a single observing station," Vistas Astron., vol. 2, pp. 850–855, Jan. 1956.
[9] L. H. Ruhnke, "Distance to Lightning Strokes as Determined from Electrostatic Field Strength Measurements," J. Appl. Meteorol., vol. 1, no. 4, pp. 544–547, Dec. 1962.
[10] G. Heydt and H. Volland, "A new method for locating thunderstorms and counting their lightning discharges from a single observing station," J. Atmos. Terr. Phys., vol. 26, no. 7, pp. 780–783, Jul. 1964.
[11] D. T. Kemp, "The global location of large lightning discharges from single station observations of ELF disturbances in the Earth-ionosphere cavity," J. Atmos. Terr. Phys., vol. 33, no. 6, pp. 919–927, Jun. 1971.
[12] W. Harth and J. Pelz, "Eastern thunderstorms located by VLF atmospherics parameters," Radio Sci., vol. 8, no. 2, pp. 117–122, Feb. 1973.
[13] A. V. Panyukov, "Estimation of the location of an arbitrarily oriented dipole under single-point direction finding," J. Geophys. Res. Atmos., vol. 101, no. D10, pp. 14977–14982, Jun. 1996.
[14] Y. Udovychenko, A. Popov, and I. Chaikovsky, "Ischemic heart disease recognition by k-NN classification of current density distribution maps," in 2015 IEEE 35th International Conference on Electronics and Nanotechnology (ELNANO), 2015, pp. 402–405.
[15] X. Han, L. Quan, X. Xiong, and B. Wu, "Facing the classification of binary problems with a hybrid system based on quantum-inspired binary gravitational search algorithm and K-NN method," Eng. Appl. Artif. Intell., vol. 26, no. 10, pp. 2424–2430, Nov. 2013.
[16] M. Klaučo, M. Kalúz, and M. Kvasnica, "Machine learning-based warm starting of active set methods in embedded model predictive control," Eng. Appl. Artif. Intell., vol. 77, pp. 1–8, Jan. 2019.
[17] G.-R. Ji, P. Han, and Y.-J. Zhai, "Wind Speed Forecasting Based on Support Vector Machine with Forecasting Error Estimation," in 2007 International Conference on Machine Learning and Cybernetics, 2007, pp. 2735–2739.
[18] A. F. R. Leal, V. A. Rakov, J. Pissolato Filho, B. R. P. Rocha, and M. D. Tran, "A Low-Cost System for Measuring Lightning Electric Field Waveforms, its Calibration and Application to Remote Measurements of Currents," IEEE Trans. Electromagn. Compat., vol. 60, no. 2, pp. 414–422, Apr. 2018.
[19] A. F. R. Leal, V. A. Rakov, and B. R. P. da Rocha, "Upgrading the Low-Cost Lightning Detection and Waveform Storage System," IEEE Trans. Electromagn. Compat., 2018.
[20] V. A. Rakov, S. Mallick, A. Nag, and V. B. Somu, "Lightning Observatory in Gainesville (LOG), Florida: A review of recent results," Electr. Power Syst. Res., vol. 113, pp. 95–103, 2014.
[21] K. L. Cummins, M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer, "A Combined TOA/MDF Technology Upgrade of the U.S. National Lightning Detection Network," J. Geophys. Res. Atmos., vol. 103, no. D8, pp. 9035–9044, Apr. 1998.
[22] K. L. Cummins and M. J. Murphy, "An Overview of Lightning Locating Systems: History, Techniques, and Data Uses, With an In-Depth Look at the U.S. NLDN," IEEE Trans. Electromagn. Compat., vol. 51, no. 3, pp. 499–518, Aug. 2009.
[23] A. Nag, M. J. Murphy, K. L. Cummins, A. E. Pifer, and J. A. Cramer, "Recent Evolution of the U . S . National Lightning Detection Network," in 23rd International Lightning Detection Conference & 5th International Lightning Meteorology Conference, 2014.
[24] A. F. R. Leal, R. Shinkai, M. N. G. Lopes, B. R. P. Rocha, V. A. Rakov, and J. Lapierre, "First Lightning Electric Field Waveform Recorder Permanently Operating In the Eastern Amazon: Preliminary Results," in Ground'2018 & 8th LPE, 2018, pp. 55–60.
[25] Richhariya, B., and Muhammad Tanveer. "EEG signal classification using universum support vector machine." Expert Systems with Applications 106 (2018): 169-182.