Journal of Intelligent & Fuzzy Systems 33 (2017) 79–92
DOI:10.3233/JIFS-161152
IOS Press

Lightning prediction using satellite atmospheric sounding data and feed-forward artificial neural network

Elton Rafael Alves[a,*], Carlos Tavares da Costa Jr[a], Márcio Nirlando Gomes Lopes[b], Brígida Ramati Pereira da Rocha[a,b] and José Alberto Silva de Sá[c]

[a]*Graduate Program in Electrical Engineering, Federal University of Pará, Rua Augusto Corrêa, Guamá, Belém, Pará, Brazil*

[b]*Operations and Management Center of the Amazonian Protection System, Avenida Júlio Cesar, Val-de-Cans, Belém, Pará, Brazil*

[c]*Center of Natural Sciences and Technology, Pará State University, Travessa Doutor Enéas Pinheiro, Marco, Belém, Pará, Brazil*

Abstract. Atmospheric discharges offer great risks to the population and activities that involve different systems such as telecommunications, energy distribution and transportation. Lightning prediction can contribute to minimize the risks of this natural phenomenon. Therefore the present paper presents a model for lightning prediction based on satellite atmospheric sounding data, calibrated and validated with lightning data in an Amazon region particular area through an investigation that considered five period cases for validation of lightning prediction: case 1 (one hour), case 2 (two hours), case 3 (three hours), case 4 (four hours) and case 5 (five hours). The machine learning technique used to predict lightning was the Artificial Neural Network (ANN) trained with Levenberg-Marquardt backpropagation algorithm to classify modeling related to lightning prediction. This classification relied on the possibility of lightning prediction from the vertical profile of air temperature obtained from satellite NOAA-19. Results show that ANN was capable of identifying adequately the class to which a new event belongs to in relation to categories of occurrence and absence of lightning with better performance than traditional methodologies.

Keywords: Classifiers, artificial neural network, prediction of atmospheric discharges, satellite atmospheric sounding

1. Introduction

Lightning is a natural phenomenon of complex origin, characterized by the flow of an impulsive current of high intensity and short duration that runs for a few kilometers in an ionized channel in the atmosphere of the earth [41]. Lightning can be classified as: cloud-to-ionosphere, intracloud, cloud-to-cloud or cloud-to-ground [30]. The current flow of an atmospheric discharge produces intense heating, with temperatures superior to 30,000K, luminous effect (flash) and quick expansion of air near the sound wave channel (thunder). A cloud-to-ground lightning is characterized by origin in Cumulonimbus clouds [41]. Typically, lightning occurs during storms [6, 14].

Cloud-to-ground lightning is one of the most interesting types for research purposes due to effects directly related to human loss and material damage [40]. One of the most negative effects of lightning to society is related to the number of deaths and injuries

[*] Corresponding author. Elton R. Alves, Graduate Program in Electrical Engineering, Federal University of Pará, Rua Augusto corrêa, Guamá, CEP 66075-110, Belém, Pará, Brazil. Tel.: +55 91 998041834; E-mail: eltonrafaelalves@gmail.com.

caused by this phenomenon which is strongly associated to weather conditions [33]. Global studies of recent decades estimate the occurrence of 6,000 to 24,000 casualties due to lightning [37]. In Brazil in 2009, the rate of deaths caused by lightning was estimated in 0.8 per million [36]. Furthermore, there are economic and social damages to telecommunication systems, power distribution and flying generated by direct and indirect lightning which lead to monetary loss [14, 26, 42]. Cloud-to-ground lightning is one of the main causes of interruption of electric power services [25]. In Brazil, 50% to 70% of interruptions in power distribution were specifically related to effects of cloud-to-ground lightning [16].

It is evident that lightning can become a potential threat to many human activities with great negative impacts. Currently, technology is not capable of avoiding lightning [26]. However, the adoption of effective strategies to predict lightning may help reducing and eliminating the impacts of lightning. In this sense, scientific investigations have been conducted to predict lightning incidence. The majority of these investigations defend a relation between favorable atmospheric conditions and lightning [16, 20, 21]. These conditions are determined by thermodynamic indexes and parameters mainly from conventional radiosounding [34], for instance Severe Weather Threat Index-SWEAT [35], Convective Available Potential Energy-CAPE [31] and Convective Inhibition Energy-CINE [13], Showalter Index [4], Lifted Index [22], K Index [23], Totals Totals Index [35], Precipitable Water [35]. Some of these studies attempted to relate some of the thermodynamic indexes and parameters using techniques of computational intelligence to develop models for lightning prediction. Weng et al. [26] used K Index and Lifetd Index to predict lightning in Malaysia by training a backpropagation algorithm neural network. Meanwhile Johari et al. [9] used meteorological data such as wind, air temperature, dew point temperature, relative humidity in air, among others as predictive parameters for lightning in a backpropagation neural network. Wang et al. [20] used CAPE as well as K Index, Index Jefferson and SWEAT index as input for a backpropagation neural network. Others like Zepka et al. [15], Zepka et al. [16] and Sá et al. [21] were also based on the conditions for atmospheric instability to predict lightning. These thermodynamic indexes and parameters are obtained through conventional radiossode data. Frankel et al. [8] developed an architecture of neural network that enables space-time mapping to predict lightning at the Cape Canaveral Air Force Station (CCAFS) and the Kennedy Space Centre (KSC). Lu et al. [19] have conducted statistical studies in the province of Huna, China, summarizing typical situations of weather conditions as favorable to lightning occurrence. Other researchers such as Juntian et al. [14] and Zeng et al. [33] have based their models of lightning prediction on atmospheric electrostatic field formed by storms. In spite of all the advances in models for predicting atmospheric discharges, there is plenty to discuss, analyze and discover.

Atmospheric convection is a fundamental physical process in the formation of and support to storm clouds, in particular those of the Cumulonimbus (Cb) [28]. These clouds generate great cores of electric charge and its process of formation depends on thermodynamic forcing (humidity and temperature) and/or dynamic forcing (meteorological systems, wind). The geographic location of Amazon as in other tropical regions presents favorable climate conditions to the formation of Cb clouds. Consequently the region presents elevated keraunic level, that is, a big number of thunderstorm days [17].

The Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM) is an institution linked to Ministério da Defesa in Brazil, which has as its mission the promotion of environmental protection for Legal Amazon that includes environmental monitoring by means of orbital sensors. CENSIPAM owns three stations located in the city of Porto Velho – Rondônia, Manaus – Amazonas and Belém – Pará, which receive data from the polar orbit satellites including NOAA-18 and NOAA-19. These satellites have atmospheric sounding devices capable of obtaining vertical profiles of temperature contributing to important analysis of regional atmospheric behavior. Satellite atmospheric sounding is a relevant tool for determining conditions of atmospheric stability once the density of the stations that operate conventional radiosonde data is extremely low in Amazon.

Storms are characterized as non-linear chaotic phenomena that involve a complex dynamics from its formation to its dissipation, which makes predictions harder [20]. Artificial Neural Networks (ANN) are a good option to attempt the modeling of a highly non-linear atmospheric phenomenon without the physical domain of the formation process of a storm cloud, just by non-linear mapping of input and output data for patterns of atmospheric weather [1]. ANN are non-linear classifying algorithms which attempts to simulate in a computer parallel processing of neurons in the human brain [26]. They are a powerful tool frequently used in applications of engineering problems of complex and non-linear nature. They have great learning skills through training, generalization, pattern recognition and prediction tasks. Li [11], Zhang et al. [43] and Yu and Chen [12] show that feedforward ANN might be applied to chaotic systems.

The present study aims at presenting a new approach to predicting lightning for the Amazon region through the applicability of atmospheric sounding data by satellite and artificial neural networks. As a consequence, the study helps in making decisions regarding the necessary preventive measures to minimize damages provoked by lightning in regions where radiossonde data is unavailable.

2. Background: Data used in the study

2.1. Atmospheric sounding

Atmospheric soundings are dynamic and thermodynamic measures of the atmosphere of the earth to obtain information related to various atmospheric parameters, such as atmospheric pressure, air temperature, dew point temperature, relative humidity in air at several levels of atmospheric pressure. A radiosonde is an instrument transported by a weather balloon that measures atmospheric parameters transmitted to a receiving station [39]. As the balloon ascends, the radiosonde registers and transmits measures at different altitudes above the surface of the Earth. These data are sent to a station on the ground that processes the signals. Radiosondes are launched twice a day at 00:00 UTC (Universal Time Coordinate) and 12:00 UTC from a weather station, reaching on occasion horizontal distance of 300 km. Its vertical limit reach is given by the bursting of the balloon.

A different form of atmospheric sounding data collection is through sensors installed in orbital platforms. As sensors are installed in polar orbit satellites with trajectories that last about 100 minutes, the same area can be visited twice a day, in average, at different times. Satellite NOAA-19 presents an orbit that crosses the Equator line near the Amazon region during the day at about 17 UTC.

The ATOVS sensor integrates the data collection platform of satellite NOAA-19 which allows for vertical sounding. ATOVS is constituted by the following sensors: Advanced Microware Sounding (AMSU), High Resolution Infrared Radiation Sounder Version 4 (HIRS/4) and Microwave Humidity Sounder (MHS). AMSU is composed by an AMSU-A instrument which is formed by another two subinstruments (AMSU-A1 and AMSU-A2) that consist of a radiometer of multichannel microwaves (15 channels) used to measure vertical profiles of global temperature and providing information on atmospheric water in all its forms (except small ice particles, which are invisible to microwave frequencies). HIRS/4 is a cross-track line scanning device projected to measure scene radiance in 20 spectral bands allowing the calculation of vertical temperature profile from earth surface up to 40 km of altitude. Multispectral data consists of a visible channel (0.69 µm), seven shortwave channels (2,188 to 2,657 µm), and twelve long wave channels (669 to 1,529 µm). MHS sensors are shortwave radiometers of high calibration with 5 channels of frequency from 89 to 190 GHz; it is responsible for tracing humidity vertical profile, detecting clouds, and precipitation, to name a few of its functions.

Satellite atmospheric sounding data allows the obtainment of atmospheric vertical profile for any region at a given moment. Figure 1 shows a Skew-T Log-P diagram from radiosounding from a station at the airport of Belém (SBBE), Pará, Brazil. The diagram shows variation of air and dew point temperatures at different levels of atmospheric pressure.

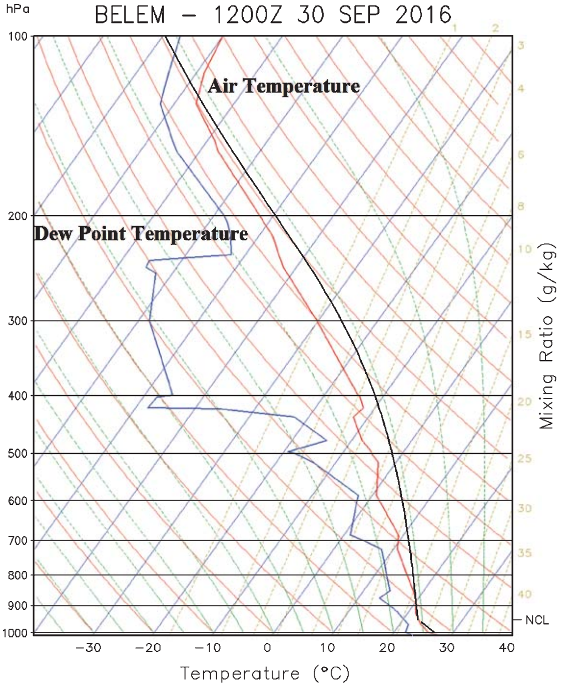

Fig. 1. Vertical profile from radiosounding.

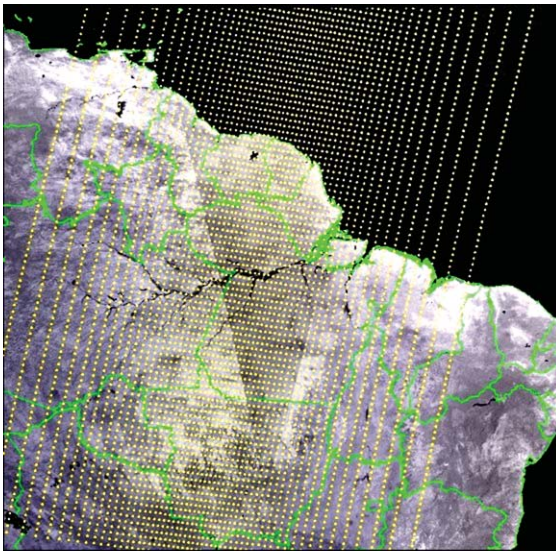

Fig. 2. Graphical interface of the Terascan® software.

The vertical sounding obtained from satellite presents advantageous characteristics when compared to conventional radiosoundings, such as operational use, due to its low cost and especially in Amazon, to the density of points to be covered. Figure 2 presents the graphical interface of Terascan viewer of the Terascan® software used by CENSIPAM to visualize scanning lanes, and their respective sounding coordinates obtained for the Amazon region. During each satellite scan, sounding points are generated as shown in Fig. 2. Each point corresponds to a measure of local sounding.

The evaluation of the quality of satellite sounding data has achieved global scale to the detriment of conventional radiosounding [18].

Knowing the vertical profile of temperature through data collected from atmospheric sounding (radiosonde and satellite) allows for the analysis of atmospheric stability of a region. A condition for instability is capable of producing intense convective activity that leads to formation of Cb clouds [3]. These clouds are normally correlated to the process of formation of atmospheric discharges, due to great centers of positive and negative electric charges in them.

2.2. *Atmospheric instability indexes*

Atmospheric instability indexes indicate the potential for formation of the Cumulonimbus clouds (Cb) that are responsible for atmospheric discharges [10]. In order to predict rain and severe storms, instability indexes were developed based on vertical profiles of temperature, humidity and wind. These indexes are used to evaluate thermodynamic conditions of the atmosphere and then verify the possibility of Cb clouds. Examples of indexes that can be used to predict convective storms: Total Totals Index [35], K Index [23] and Convective Available Potential Energy (CAPE) [31], to name a few.

2.2.1. *Total Totals Index*

An index used to assess storm strength. This index is a combination of Vertical Totals Index (VTI) and Cross Totals Index (CTI). VTI is given by Equation (1) while CTI is given by Equation (2) below:

$$VTI = T_{850} - T_{500} \tag{1}$$

Where T_{850} is the temperature (°C) for pressure level 850 mb and T_{500} is the temperature (°C) for pressure level 500 mb.

$$CTI = T_{d850} - T_{500} \tag{2}$$

Where T_{d850} is the dew point temperature (°C) for pressure level 850 mb and T_{500} is the temperature (°C) for pressure level 500 mb.

The Vertical Totals Index is given by the combination of Equations (1) and (2):

$$TTI = VTI + CTI \tag{3}$$

The higher the values for TTI the more instable the atmosphere is. For TTI values above (48°C) there is the strong possibility of storm and lightning for the Amazon region.

2.2.2. *K index*

An index used to assess the convective potential for storms. It is given by Equation (4) below:

$$VTI = (T_{850} - T_{500}) + T_{d850} - (T_{700} - T_{d700}) \tag{4}$$

Where T_{850} is the air temperature (°C) for pressure level 850 mb, T_{500} is the air temperature (°C) for pressure level 500 mb, T_{d850} is the dew point temperature (°C) for pressure level 850 mb, T_{700} is the air temperature (°C) for pressure level 700 mb and T_{d700} is the dew point temperature (°C) for pressure level 700 mb.

The more positive the index, the more is the chance of storms. Values for K above 20 indicate 20% to 40% probability of storm formation and k values above 40 indicate probability close to 100% for heavy storms. For the Amazon region K value was considered more than or equal to 31 as reference value for storm and lightning occurrences.

2.2.3. *Convective Available Potencial Energy (CAPE)*

CAPE corresponds to the positive area of a sounding in the thermodynamic diagram (Skew-T Log-P diagram) that indicates the amount of energy available for convection, a fundamental process in the formation of storm clouds. It is obtained by Equation (5) below:

$$CAPE = \int_{NE}^{NCE} \left[\frac{\Theta_e(z) - \Theta_{es}(z)}{\Theta_{es}(z)} \right] dz \qquad (5)$$

Where NCE is the level of spontaneous convection that corresponds to the moment in which part of the air starts to rise spontaneously. NCE is the limit inferior of the integral that determines the CAPE area. NE is the balance level that corresponds to equality between environment temperature and temperature of a part of the rising air. NE is the limit superior of the integral that represents the top of the cloud.

Values for CAPE obtained with Equation (5) and the stability conditions related to these values are shown in Table 1.

Table 1
Values for CAPE

CAPE (J/Kg)	Related conditions
$500 \leq CAPE \leq 1000$	Weak instability
$1000 \leq CAPE \leq 500$	Moderate instability
$CAPE \geq 2500$	Strong instability

2.3. *Lightning Localization System (LLS)*

Atmospheric discharges irradiate electromagnetic energy in a broad lane of frequencies that can be captured by sensors of electromagnetic frequency installed in radio antennas in LLS stations. In Brazil, Sferics and Timing Ranging Network-STARNET is a long-range network that can detect and locate lightning occurrences. It uses radio antennas to detect frequencies of electromagnetic waves that operate in VLF (Very Low Frequency) [7]. These antennas measure vertical electromagnetic fields in the frequency band of 7-15 GHz. Determining the position of the atmospheric discharge with STARNET is accomplished through the technique of arrival time differences (ATD) of the electromagnetic field [2].

STARTNET works with VLF radio antennas operating in Cape Verde, Guadalupe (Caribbean), Chile, Argentina, Manaus (Brazil), Belém (Brazil), Fortaleza (Brazil), Bahia (Brazil), São Paulo (Brazil), São Martinho da Serra (Brazil), Campo Grande (Brazil) and Brasilia (Brazil).

3. Materials and methods

In this section, the area of investigation, the data used, their pre-processing and the model of training for the neural network as well as its asessment will be discussed.

3.1. *Area of investigation and pre-processing*

The region of investigation comprehends eight areas located in the Northeast of the state of Pará, in Brazil, as shown in Fig. 3, totalizing areas between latitudes 01°S and 02°S, and longitudes of 46.5°W and 48.5°W. The study encompasses a total area of 24.961,861 km^2.

Among all eight sites selected for application of the methodology destined to prediction of lightning, just one point (Belém) situated in site 01, presents regular conventional sounding, which characterizes the other areas as uncovered regions. In addition, in this site of investigation there are elevated indexes of atmospheric discharges [17].

3.1.1. *Data description*

The geographic coordinates for the centroid of all eight sites shown in Fig. 3 are shown in Table 2. These geographical locations served as references for sounding data collection from satellite. Therefore, for the centroid of each site, sounding measures were obtained. Consequently, for each day 8 sounding measures were obtained for the geographic coordinates in Table 2.

In this study, sounding data from satellite NOAA-19 were used, manipulated through Terascan®. Air temperature and dew point temperature were the atmospheric parameters obtained for 30 levels of atmospheric pressure in each area centroid.

Sounding data collected from satellite NOAA-19 concerned the months between June and December of 2014. Thus, 608 examples were extracted for the set of data. The reference schedule to obtain satellite data was 17 UTC. In this sense, prediction was done an hour ahead of 18 UTC, considering five cases of validity time for prediction models as shown in Fig. 4. The cases considered for prediction were: case 1 (one hour), case 2 (two hours), case 3 (three hours),

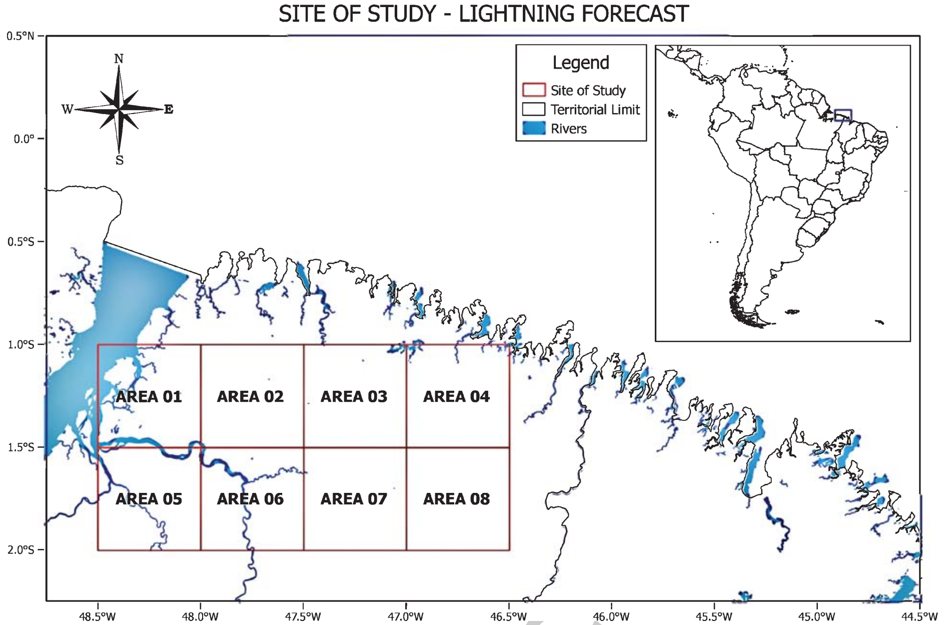

Fig. 3. Area of investigation.

Table 2
Geographic coordinates for the centroid of each area

Area	Coordinates	
	Latitude	Longitude
A1	1.25°S	48.25°W
A2	1.25°S	47.75°W
A3	1.25°S	47.25°W
A4	1.25°S	46.75°W
A5	1.75°S	48.25°W
A6	1.75°S	47.75°W
A7	1.75°S	47.25°W
A8	1.75°S	46.75°W

case 4 (four hours), and case 5 (five hours). For each case, models based in ANN were developed through several performance tests for all five validity times.

This study mainly seeks to evaluate whether the conditions for atmospheric stability in terms of vertical profile of environment temperature (air and dew point temperatures) obtained by means of sounding from satellite NOAA-19 are favorable to thunderstorms forecast for all five cases analyzed in Fig. 4.

Historic data on lightning used in this study comes from the STARNET database. Therefore, the amount of lightning was counted in terms of period of soundings data obtained from satellite NOAA-19. The absence of lightning was denoted “0” and incidence of lightning “1”.

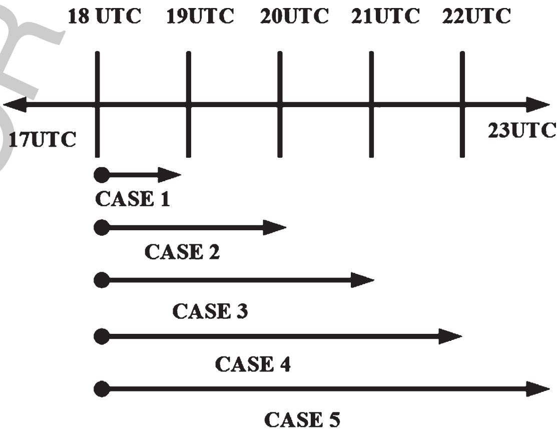

Fig. 4. Lightning prediction cases.

3.2. Data selection

As there is a monotonic relation between pressure and altitude in each vertical column of the atmosphere, meteorology usually uses pressure as a vertical coordinate to simplify equations for solving thermodynamic problem. For a point situated at a certain height, the higher the value of the coordinate of pressure, the closer to the surface the point will be. In such cases, the unity that normally applies to atmospheric pressure is the HectoPascal (hPa).

Atmospheric parameters obtained from sounding, air temperature and dew point temperature refer to the following levels of atmospheric pressure: 10hPa, 15hPa, 20hPa, 25hPa, 30hPa, 50hPa, 60hPa, 70hPa, 85hPa, 100hPa, 115hPa, 135hPa, 150hPa, 200hPa, 250hPa, 300hPa, 350hPa, 400hPa, 430hPa, 475hPa, 500hPa, 570hPa, 620hPa, 670hPa, 700hPa, 780hPa, 850hPa, 920hPa, 950hPa, 1000hPa. These levels of pressure went through a process of reduction of variables through the Principal Component Analysis (PCA).

Characteristics of distribution of data variability for air temperature and dew point at all 30 levels of atmospheric pressure are shown in Figs. 5 and 6. The layer from 850hPa and 150hPa presented low dispersion concerning the data on air temperature while dew point temperature presented lower dispersion for higher levels of atmosphere, between 135hPa and 10hPa.

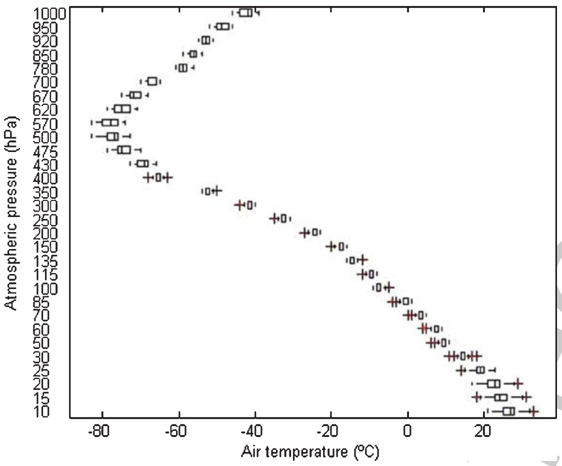

Fig. 5. Variability for air temperature.

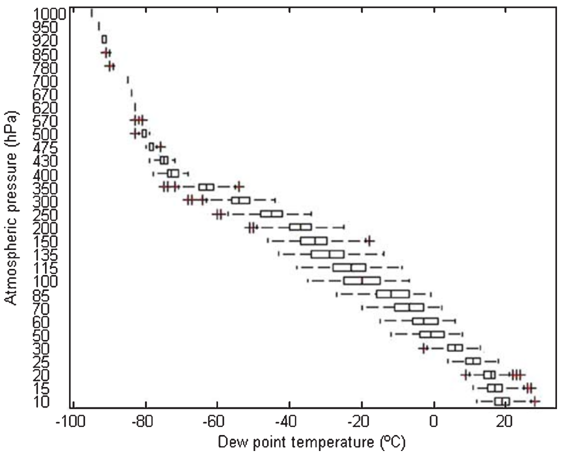

Fig. 6. Variability for dew point temperature.

It was verified that of all 30 components relative to air temperature variables, the first three components represented approximately 78% of total explained variance. Thus, the fifteen different levels of pressure for which air temperature presented the best coefficients in the linear combination of their respective principal components were: 10hPa, 15hPa, 20hPa, 25hPa, 30hPa, 50hPa, 60hPa, 70hPa, 85hPa, 100hPa, 115hPa, 850hPa, 920hPa, 950hPa e 1000hPa. At such pressure levels there was greater variability of air temperature, thus being defined as input variable for ANN training.

From the 30 principal components of variables for dew point temperatures, the first component represented approximately 81% of total explained variance. Thus, seventeen variables related to dew point temperature presented considerable numeric values for their respective coefficients in the linear combination of these principal components: 150hPa, 200hPa, 250hPa, 300hPa, 350hPa, 400hPa, 430hPa, 475hPa, 500hPa, 570hPa, 620hPa, 670hPa, 700hPa, 780hPa, 850hPa, 920hPa e 1000hPa that corresponds to variables of greater variability. Therefore, these levels composed the input variable for dew point temperature used for ANN training.

Variables explained by the principal components of air temperature until the ninth component, and dew point temperature until the third principal component is shown in Figs. 7 and 8.

3.3. Data normalization

After data selection, normalization was executed to reduce discrepancies between values of input variables. Equation (6) describes the method of normalization used:

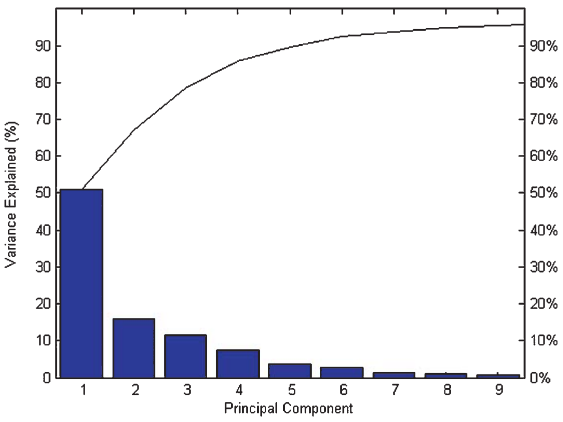

Fig. 7. Principal components: air temperature.

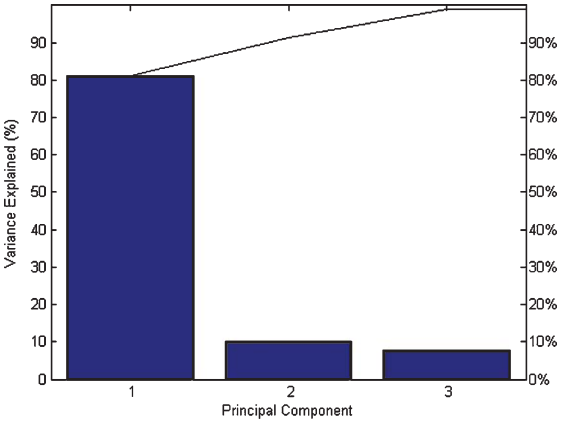

Fig. 8. Principal components: dew point temperature.

$$value_{normalized} = \frac{value_{original} - minA}{maxA - minA} \quad (6)$$

This normalization transformed original values of input variables into values of the interval [0, 1].

3.4. Design of neural network

Two types of data were used in this proposition, meteorological data selected as input (air temperature and dew point temperature) and historical data for lightning as target output, as shown in Fig. 9. Each application of neural network used in the three case studies, received input data denoted by *D* and *I*, and target output as *t*.

Characteristic for the input matrix of the ANN is given by:

$$P = [D, I, t]$$

Where *D* is the data for air temperature (°C), *I* is the data for dew point temperature (°C) and *t* is the historic vector of lightning. Matrix *D* is given by:

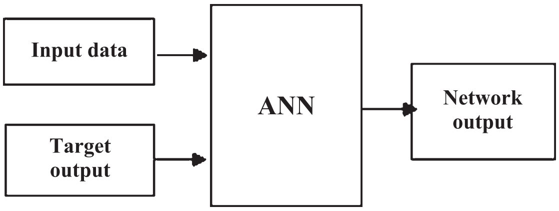

Fig. 9. Model of ANN training.

$$D = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

Where n is the number of input patterns and $[m \times n]$ is the size of the matrix and m is the the number of input variables for air temperature (°C). Matrix *I* is given by:

$$I = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ y_{m1} & y_{m2} & \dots & y_{mn} \end{bmatrix}$$

Where n is the number of input patterns and $[m \times n]$ is the size of the matrix and m is the number of input variables for dew point variables (°C). The target output, denoted by *t*, is given by the following vector:

$$t = \begin{bmatrix} c_1 \\ c_2 \\ . \\ . \\ . \\ c_n \end{bmatrix}$$

Where C_n is given by "0" or "1".

The ANN application used in this study serves to identify simultaneously whether the conditions for air temperature and dew point temperature obtained during the passage of the satellite can evolve for thunderstorms. Classification techniques might be used as predictive modeling, when a classifier is used to identify to which class a new example belongs [32]. In this sense, an ANN was used to recognize patterns to classify predictive attributes of training for five ANN in a set of two categories "0" for not occurrence of lightning and "1" for incidence of lightning. The schematics for the ANN model used to predict meteorological events for this study is shown in Fig. 10.

The adopted model consisted of a three-layer feedforward ANN: the first layer is the input (vector *P* of input); the second layer is a hidden layer (given by hidden neurons) and the third layer is the output that

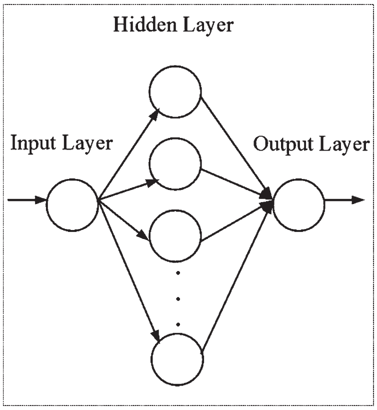

Fig. 10. ANN model.

represents the output of lightning predictor (0 or 1). A sigmoid activation function in the hidden layer was used as shown in Equation (7). The non-linear nature of the sigmoid function is essential for the development of the ANN that must represent the non-linear characteristics in the set of input data.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

The activation function for each output layer was the softmax function, exhibited in Equation (8). This activation function is implemented in the output layer of the ANN to classify predictive patterns.

$$f(z) = \frac{e^{z_i}}{\sum_{m=1}^{k} e^{z_m}} \tag{8}$$

Where k represents the number of classes of the output layer.

ANN training was done with backpropagation Leve-nberg-Marquardt algorithm which allows a faster training for ANN in comparison to other training methods [29].

3.5. Evaluation method

One way to assess ANNs was the employment of the rate for adequate classification obtained through a confusion matrix between the predicted class (output class) and the true class (real class). In the matrix, each column represents a true result, while each line represents a predicted result. The confusion matrix was used to assess models of classification used in this study is exposed in Table 3 [38].

Table 3
Confusion matrix for the evaluation of classification models

	True Class	
Predicted Class	TP	FP
	FN	TN
	P	N

When a positive example was classified as positive by the classifier, it was computed into the matrix as true positive, however if it was classified as negative, then was denominated false negative. When a negative example was classified as negative by the classifier then it was computed as true negative, however if it was classified as positive, then it was classified as false positive. From these crossings we obtained: TP (True Positives), FP (False Positives), FN (False Negatives) and TN (True Negatives). Finally, N represents the number of total negative events and P the number of total positive events, which are analyzed by the classifier.

From Table 3 possible metric analyses of the classifier performance are also exposed: true positives rate (*tp_rate*), false positives rate (*fp_rate*), false negatives rate (*fn_rate*), true negatives rate (*tn_rate*) and classification accuracy.

$$tp\_rate = \frac{TP}{P} \times 100 \tag{9}$$

$$fp\_rate = \frac{FP}{N} \times 100 \tag{10}$$

$$fn\_rate = \frac{FN}{P} \times 100 \tag{11}$$

$$tn\_rate = \frac{TN}{N} \times 100 \tag{12}$$

$$accuracy = \frac{TP + TN}{P + N} \times 100 \tag{13}$$

Result analyses were also conducted through the Receiver Operating Characteristic Curves (ROC). ROC graph allows the visualization of a classifier's performance [38]. In this ROC graph the true positives rate of classifiers is registered on the Y axis, while the false positive rate is registered on the X axis. An ideal classifier is represented by the point $(0, 1)$. Therefore in comparative terms between classifiers, the further to Northwest (*tp_rate* is higher and/or *fp_rate* is shorter) of the horizontal line of the ROC space, the better the development of the classifier [5].

Another form of evaluating the quality of a classifier is through the calculation of the area under the curve (AUC) of the bi-dimensional ROC space [5, 27, 38], which constitutes an acceptable form of general comparison between classifier performances [24].

An ideal classifier has AUC=1, that happens when the classifier is able to separate classes perfectly.

In Looy et al. [27] the calculation of AUC is evaluated as a measure of more sophisticated quality than the general accuracy obtained by the classifier. Therefore, AUC was obtained for each study case to confirm the accuracy of the results.

Attempting to avoid over fitting of the ANN, that is, a bad generalization of the ANN through the incorporation of noise present in the input data, three sets of data were used: training, validation and test to adjust the best prediction models. Then, the stopping criterion of the ANN consisted in preventing the ANN from excessively adjusting the set of training and validation, by using a set of test data.

4. Results and discussions

MATLAB® Neural Network Toolbox was used to configure ANN as shown in Fig. 10. All 608 examples of sounding data were randomly divided in: training data for modification of ANN weights; validation data to estimate the capacity of generalization of the ANN; and test data to test the generalization of ANN. Therefore, results of test data were used as global accuracies for the models.

As the ANNs were trained with 426 patterns (≈ 70%), 91 patters were used for validation (≈ 15%) and 91 patterns for test (≈ 15%). In each prediction case analyzed, the number of neurons in the hidden layer varied during the training of ANN, in order to evaluate ANN performance for validity times for the analyzed predictions analyzed. Properties of these ANN developments used in the present study are summarized in Table 4.

Results for prediction tests with the best configuration of HNN (number of hidden neurons on the hidden layer) obtained for each ANN in all five cases are shown in Table 5.

Table 4
Properties of development of ANN

ANN Properties	Properties
Training technique	Levenberg-Marquardt
Transfer function of the hidden layer	Sigmoid
Transfer function of the hidden output	Softmax
Neurons in the hidden layer	1-30
Momentum constant	0.001
Training patterns	426
Validation patterns	91
Test patterns	91

Table 5
Results for cases 1, 2, 3, 4 and 5

	HNN	Accuracy(%)
Case 1	28	72.5
Case 2	1	75.8
Case 3	11	74.7
Case 4	16	86.8
Case 5	21	95.6

Table 6
Confusion and accuracy matrix

Case 1		Case 2		Case 3	
52.7%	16.5%	46.2%	14.3%	54.9%	17.6%
11%	19.8%	9.9%	29.7%	7.7%	19.8%
Accuracy (72.5%)		Accuracy (75.8%)		Accuracy (74.7%)	
Case 4		Case 5			
79.1%	6.6%	92.3%	3.3%		
6.6%	7.7%	1.1%	3.3%		
Accuracy (86.8%)		Accuracy (95.6%)			

The best accuracy rates of test (%) were, respectively, for case 1, case 2, case 3, case 4, and case 5: 72.5% with 28 hidden neurons, 75.8% with 1 hidden neuron, 74.7% with 11 hidden neurons, 86.8% with 16 hidden neurons and 95.6% with 21 hidden neurons. It is noted that in cases 1, 2, and 3 results are close together, while in cases 4 and 5 there was greater increase in accuracy. According to accuracy results, it is noted that ANN improved progressively with increase in time for the prediction window, obtaining the best result in case 5.

The confusion matrix obtained in this study from the results of Table 5 is shown in Table 6. The best rate of true positives and false positives were obtained in case 5, with total accuracy of 95.6%. This demonstrates that the ANN was able to correctly predict for *tp_rate* $= 92.3\%$ situations in which atmospheric discharges will truly occur or not occur and error with *fp_rate* $= 3.3\%$.

Results obtained throught algorithm Levenberg-Marquardt backpropagation (trainlm), shown in Table 5, were compared to another two training algorithms already tested: Scaled Conjugate Gradient backpropagation (traincsg) and Gradient Descent backpropagation (traingd). Results for training algorithms are comparatively shown in Table 7.

It is noted that trainlm obtained the best performance for all prediction cases analyzed while presented the worst performance in cases 1 to 4, with 17 and 4 hidden neurons (HNN), respectively. In cases 3 and 5, traingd and traincsg presented similar accuracies of 73.6% in case 3 and 93.4% in case 5. In case 3, trainscg needed more HNN (HNN = 21) to achieve accuracy of 73.7% in relation to traingd; and for case 5, it needed less HNN (HNN = 17) to achieve accuracy of 93.4% in relation to traingd. In case 2,

Table 7
Comparison of prediction accuracy

	Case 1		Case 2	
Algorithm	HNN	Accuracy (%)	HNN	Accuracy (%)
trainlm	28	72.5	1	75.8
traincsg	4	71.8	3	73.6
traingd	17	69.2	1	75.4
	Case 3		Case 4	
Algorithm	HNN	Accuracy (%)	HNN	Accuracy (%)
trainlm	11	74.7	16	86.8
traincsg	21	73.7	16	85.7
traingd	3	73.7	4	84.6
	Case 5			
Algorithm	HNN	Accuracy (%)		
trainlm	21	95.6		
traincsg	17	93.4		
traingd	26	93.4		

traingd presented higher accuracy than traincsg. Therefore it is evident why algorithm trainlm was used in this study of lightning prediction as it presented the best results.

This methodology was validated by comparison to a traditional methodology (TM) for lightning prediction based on the obtainment of indexes of atmospheric instability. Such validation was conducted by considering study site 01 (Fig. 3). CAPE and Total Totals Index (TTI) was obtained for this site corresponding to data from satellite atmospheric sounding. Therefore, the comparative validation on the same date and time between the proposed methodology (PM) and the traditional methodology was achieved. Figures 11, 12 and 13 show behavior of values for the indexes. The red line in each figure corresponds to minimum necessary values for thunderstorms for the Amazon region. This minimum value of reference for thunderstorm was denoted “1” as the lightning occurrence. The non-occurrence of lightning was denoted “0” as well as values below the reference value of the index.

It was assessed whether the index values led to the occurrence or not of thunderstorms in all five cases of prediction. Such comparison was conducted through the STARNET lightning database [7]. Afterwards, the accuracy (%) of K index (KI), Total Totals Index (TTI) and CAPE were assessed. Figures 14, 15, 16, 17 and 18 exhibit prediction accuracy for indexes CAPE, KI, TTI and PM for cases 1, 2, 3, 4, and 5, respectively. It is clear that PM greatly improves the efficiency of prediction in all cases when compared to traditional methodologies.

The comparative ROC graph evaluates the best results obtained through an ANN is shown in Fig. 19 for the task of predicting atmospheric discharges in

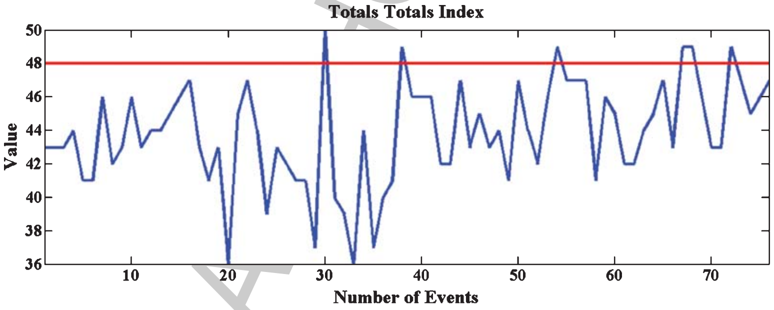

Fig. 11. Index TTI.

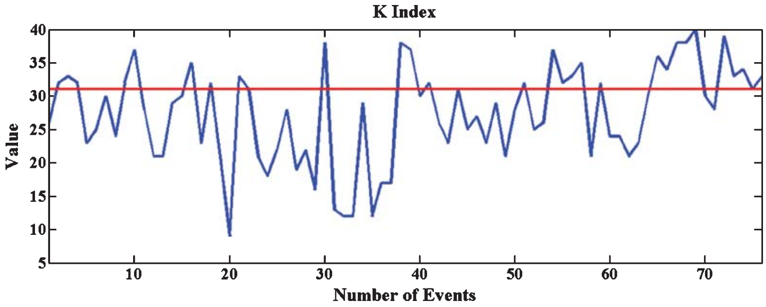

Fig. 12. K Index.

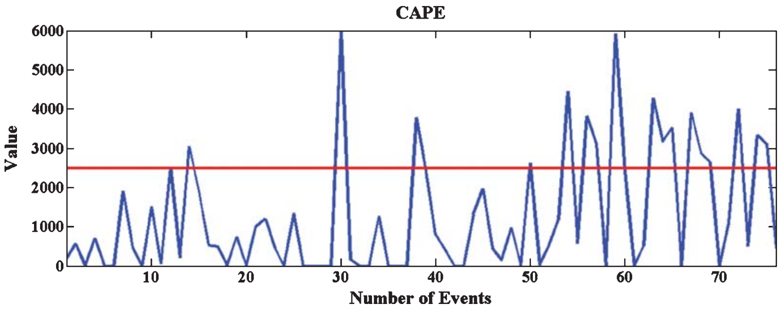

Fig. 13. CAPE.

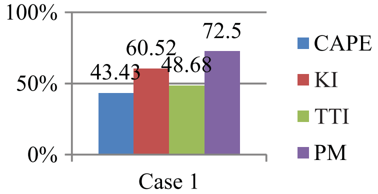

Fig. 14. Case 1: Prediction accuracy comparison between PM (Proposed Methodology) and TM (Traditional Methodologies).

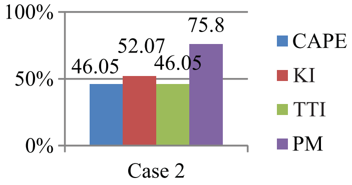

Fig. 15. Case 2: Prediction accuracy comparison between PM (Proposed Methodology) and TM (Traditional Methodologies).

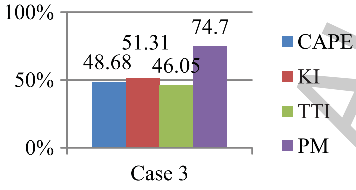

Fig. 16. Case 3: Prediction accuracy comparison between PM (Proposed Methodology) and TM (Traditional Methodologies).

cases 1, 2, 3, 4, and 5 according to the visualized in Table 6. It is noted that case 5 presented better rates for true positives and smaller rates of false positives, being close to axis Y (left side) indicating that it had the best results in all other cases, according to Table 5. Cases 1, 2, and 3 presented closer results. While case 4 presented better accuracy compared to case 1, case 2, and case 3.

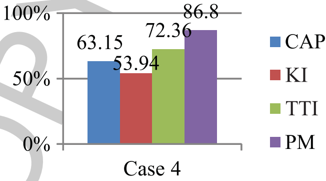

Fig. 17. Case 4: Prediction accuracy comparison between PM (Proposed Methodology) and TM (Traditional Methodologies).

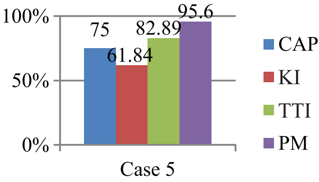

Fig. 18. Case 5: Prediction accuracy comparison between PM (Proposed Methodology) and TM (Traditional Methodologies).

Punctual values for AUC obtained in this study are shown in Table 8. All cases had UAC values much higher than 0.5. The value of AUC = 0.5 corresponds to the incapacity of the classifier in predicting or not electric discharges. Therefore, in all cases the results obtained for AUC were considered satisfactory. Case 5 presented higher value of AUC = 0.871 thus confirming that it presented the best performance among all cases analyzed. That is, of all five cases applied for this study, this was the model of prediction that best adjusted to the local atmospheric conditions. That does not exclude the possibility of using the other models; however they are less reliable in their results. Results are consistent with the importance of sounding data from satellite in the prediction of atmospheric discharges in the Amazon region. It is possible to affirm greater efficiency for predicting lightning during a validity time of until 5 hours, while worse efficiency would be of 1 and 3 hours.

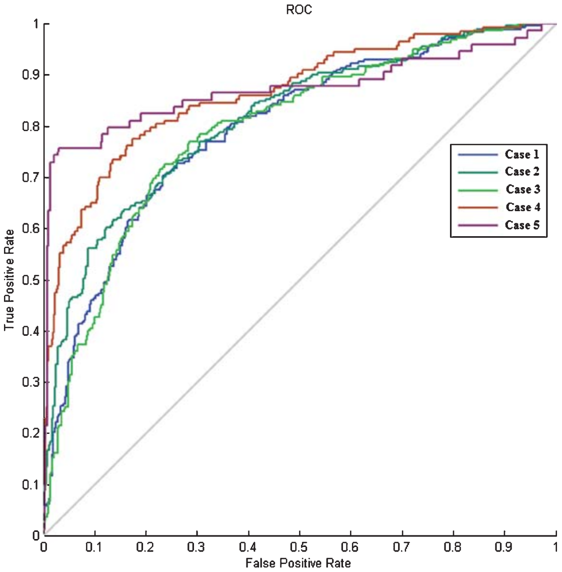

Fig. 19. ROC graph for the best results of cases 1, 2, 3, 4 and 5.

Table 8
Results for AUC obtained from ROC

	Case1	Case 2	Case 3	Case 4	Case 5
AUC	0.794	0.813	0.793	0.864	0.871

5. Conclusion

A new approach for lightning prediction based on data from satellite atmospheric sounding was presented by the application of ANN to classify predictive models. Best accuracies for prediction models in this study show that ANN is an extremely powerful tool in pattern recognition tasks. It managed to demonstrate the occurrence of high rated of accuracy of events considered true positive.

This technique obtained for case 5 accuracy of 95.6% and AUC = 0.871, being able to predict the incidence of lightning up to five hours in advance for the Northwest of Pará. To demonstrate the improvement of the proposed model for prediction tasks for atmospheric discharges, traditional methods based on calculations of atmospheric instability indexes (KI, TTI and CAPE) were obtained. Accuracy values for the proposed model were superior to traditional methodologies demonstrating a more useful tool and more effective in the prediction of atmospheric discharges in the Amazon region.

The prediction model was based on the use of just two meteorological variables obtained from sounding from satellite: air temperature and dew point temperature. However, these two parameters were sufficient for the ANN to recognize patterns strongly correlated to historic data of lightning. As these two measures are the basis for the derivation of some thermodynamic indexes and parameters of conventional soundings, indirectly, the models were able to incorporate conditions for instability that could lead to the formation of lightning storm clouds.

Results suggest the potential for expanding of lightning predictions to reach all Legal Amazon. Applications of this information are innumerous, comprehending protection of human lives and management of different systems as telecommunications, power distribution, aerial operations, railway and highways, and many others.

A large database will certainly contribute to improvements in the technique and achieve more promising results in the future. Similarly, better coverage of a lightning detection network is fundamental to reduce noise and improve the development of learning algorithms to model the phenomenon.

Acknowledgments

We would like to thank the Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM) for providing NOAA-19 atmospheric sounding data, manipulated through Terascan® software.

References

[1] A. Manzato, Sounding-derived indices for neural network based short-term thunderstorm and rainfall forecast, *Atmospheric Research* **83** (2007), 349–365.

[2] A.C.L. Lee, An experimental study of the remote location of lightning flashes using a VLF arrival time difference technique, Quart, *J R Met Soc* **112** (1986), 203–229.

[3] A.F. Ali, D. Johari, N.F.N. Ismail, I. Musirin and N. Hashim, Thunderstorm forecasting by using artificial neural network, *Proceedings of 5th International Power Engineering and Optimization Conference*, Shah Alam, Selangor, Malaysia, 2011, pp. 369–374.

[4] A.K. Showalter, A stability index for forecasting thunderstorms, *Bull Amer Meteor Soc* **34** (1947), 250–252.

[5] A.R.V. Erkel and P.M.T. Pattynama, Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology, *European Journal of Radiology* **27**(2) (1998), 88–94.

[6] C. Liu and S. Heckman, Using total lightning data in severe storm prediction: Global case study analysis from North America, Brazil and Australia, *Proceedings of 11th International Symposium on Lightning Protection*, Fortaleza, Brazil, 2011.

[7] C.A Morales, J.R. Neves and E. Anselmo, Sferics Timing and Ranging Network-Starnet: Evaluation over South

America, *Proceedings of 11th International Symposium on Lightning Protection*, Fortaleza, Brazil, 2011.

[8] D. Frankel, I. Schiller, J.S. Draper and A.A. Barnes, Use de neural network to prediction lightning at kennedy space center, *Seattle International Joint Conference Neural Network* **1** (1991), 319–324.

[9] D. Johari, T.K.A. Rahman and I. Musirin, Artificial neural network technique for lightning prediction, *Proceedings 5th Student Conference on Research and Development*, Malaysia, 2007.

[10] D.R. Macgorman and W.D. Rust, The electrical nature of storms, New York: Oxford University Press, 1998, p. 422.

[11] D.-M. Li, Identification of chaotic systems with large noise on regularized feedforward neural network, *International Conference on Machine Learning and Cybernetics* **7** (2005), pp. 4060–4063.

[12] E.S. Yu and C.Y.R. Chen, Traffic prediction using neural network, *Global Telecommunications Conference, Including a Communications Theory Mini-Conference, Technical Program Conference a Record, IEEE in Houston* **2** (1993), pp. 991–995.

[13] F.P. Colby Jr, Convective inhibition as a predictor of convection during AVE-SESAME II, *Mon Wea Rev* **1984**(112), 2239–2252.

[14] G. Juntian, G. ShanQiang and F. Wanxing, A lightning motion prediction technology based on spatial clustering method, *Proceedings 7th Asia-Pacific International Conference on Lightning*, Chengdu, China, 2011.

[15] G.S. Zepka and A.C.V. Saraiva, Forecast using WRF model over EDP distribution companies areas, *Proceedings of 12th International Symposium on Lightning Protection*, Belo Horizonte, Brazil, 2013.

[16] G.S. Zepka, O. Pinto Jr. and A.C.V Saraiva, Lightning forecasting in southeastern Brazil using the WRF model, *Atmospheric Research* **135-136** (2014), 344–362.

[17] I.R.C.A. Pinto and O. Pinto Jr., Cloud-to-ground lightning distribution in Brazil, *Journal of Atmospheric and Solar-Terrestrial Physics* **65** (2003), 733–737.

[18] J. Li, W.W. Wolf, W.P. Menzel, W. Zhang, H.L. Huang and H.T. Achtor, Global sounding of the atmosphere from atovs measurements: The algorithm and validation, *Journal of Applied Meteorology* **39** (2000), 1248–1268.

[19] J. Lu, H. Zhang, L. Yang, B. Li, Z. Fang and X. Xu, Forecast method of lightning activity based on the weather conditions, *Proceedings of 7th Asia-Pacific International Conference on Lightning*, Chengdu, China, 2011.

[20] J. Wang, B. Zhou and S. Zhou, Lightning potential forecast over Nanjing with denoised sounding-derived indices based on SSA and CS-BP, *Atmospheric Research* **137** (1994), 245–256.

[21] J.A.S. Sá, B.R.P. Rocha, A.C. Almeida and J.R.S. Souza, Recurrent self-organizing map for severe weather patterns recognition, In *Recurrent Neural Networks and Soft Computing*, ISBN 979-953-307-546-3, 2011.

[22] J.G. Galway, The lifted index as a predictor of latent instability, *Bull Amer Meteor Soc* (1956), 528–529.

[23] J.J. George, Weather Forecasting for Aeronautics, Academic Press, 1960, p. 673.

[24] K. Woods and K.W. Bowyer, Generating ROC curves for artificial neural networks, *IEEE Transactions on Medical Imaging* **16**(3) (1997), pp. 329–337.

[25] L. Pu, J. Deng, S. Qian, S. Gu and X. Cao, A new method to study on distribution characteristics of cloud-to-ground lightning, *Proceedings International Conference on High Voltage Engineering and Application*, Shanghai, China, 2012, pp. 17–20.

[26] L.Y. Weng, J.B. Omar, Y.K. Siah, S.K. Ahmed and I.B.Z. Abidin, Lightning forecast using ANN-BP e radiosonde, *Proceedings International Conference on Intelligent Computing and Cognitive Informatics*, Kuala Lumpur, Malaysia, 2010, pp. 152–155.

[27] Looy, et al., Prediction of dose escalation for rheumatoid arthritis patientes under infliximab treatment, *Engineering Applications of Artificial Intelligence* **19** (2006), 819–828.

[28] M.A. Uman, Natural lightning, *IEEE Transactions on Industry Applications* **30**(3) (1994), 785–790.

[29] M.H. Fun and M.T. Hagan, Levenberg-Marquardt training for modular networks, *Proceedings of International Conference on Neural Network* **1** (1996), 468–473.

[30] M.L. Akyinyemi, A.O. Boyo, M.E. Emetere, M.R. Usikalu and F.O. Olawole, Lightning a fundamental of atmospheric electricity, *Proceedings International Conference on Environment Systems Science and Engineering* **9** (2014), pp. 47–52.

[31] M.W. Moncrieff and M.J. Miller, The dynamics and simulation of tropical cumulonimbus and squall lines, *Quart J Roy Meteor Soc* **1976** (102), 373–394.

[32] P.N. Tan, M. Steinbach and V. Kumar, *Introduction to data mining*, Boston: Addison-Wesley Longman: Boston, USA, 2005.

[33] Q. Zeng, Z. Wang, F. Guo, M. Feng, S. Zou, H. Wang and D. Xu, The application of lightning forecasting based on surface electrostatic field observations and radar data, *Atmospheric Research* **71** (2013), 6–13.

[34] R. Pippier, A review of static stability indices and related thermodynamic parameters, *Ilinois State Water Survey Climate and Metereology Section*, USA, 1988, p. 94.

[35] R.C. Miller, Notes on analysis and severe storm forecasting procedures of the Air Force GlobalWeather Central, Tech Rept 200(R), Headquarters, Air Weather Service, USAF, 1972, p. 190.

[36] R.J. Holle, Some Aspects of Global Lightning Impacts, *Proceedings International Conference on Lightning Protection (ICLP)*, Shanghai, China, 2014, pp. 1390–1395.

[37] R.L. Holle, Annual rates of lightning fatalities by country, *Preprints, International Lightning Detection Conference, Tucson, Arizona*, Vaisala, 2008, p. 14.

[38] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters* **27**(86) (2006), 861–874.

[39] T.A. Musa, S. Amir, R. Othman, S. Ses, K. Omar, K. Abdullah, S. Lim and C. Rizos, GPS meteorology in a lowlatitude region: Remote sensing of atmospheric water vapor over the Malaysian Peninsula, *Journal Atmosp Solar-Terrest Phys* **73** (2011), 2410–2422.

[40] V. Cooray, C. Cooray and C. Andrews, Lightning caused injuries in humans, *Journal of Electrostatics* **65**(5-6) (2007), 386–394.

[41] V.A. Rakov and M.A. Uman, Lightning: Physics and Effects, New York: Cambridge Universty Press, 2003.

[42] V.A. Rakov, M.A. Uman, M.I. Fernandez, C.T. Mata, K.J. Rambo, M.V. Stapleton and R.R Sutil, Direct lightning strikes to the lightning protective system of a residential building: Triggered-lightning experiments, *IEEE Transaction on Power Delivery* **17**(2) (2002), 575–586.

[43] W. Zhang, J. Liang, J. Wang and J. Che, Chaotic time series forecasting based on fuzzy adaptive PSO for feedforward neural network training, *The 9th International Conference for Young Computer Scientists*, 2008, pp. 3022–3027.