Modeling the Distribution of Lightning Strike Distances Outside a Preexisting Lightning Area

THESIS

Dawn L. Sanderson, Captain, USAF

AFIT-ENC-MS-19-M-003

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENC-MS-19-M-003

MODELING THE DISTRIBUTION OF LIGHTNING STRIKE DISTANCES OUTSIDE A PREEXISTING LIGHTNING AREA

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Applied Mathematics

Dawn L. Sanderson, MS

Captain, USAF

March 2019

AFIT-ENC-MS-19-M-003

MODELING THE DISTRIBUTION OF LIGHTNING STRIKE DISTANCES OUTSIDE A PREEXISTING LIGHTNING AREA

THESIS

Dawn L. Sanderson, MS
Captain, USAF

Committee Membership:

Dr. Edward White
Chairman

Lt Col Andrew J. Geyer, Ph.D.
Member

Dr. Alex Gutman
Member

Abstract

Air Force Instruction 91-203 (AFI 91-203) directs that a lightning warning be issued when lightning is occurring or imminent within a 5 nautical mile (NM) radius of a predetermined location or activity. The 45 Weather Squadron (WS), located on the central eastern coast of Florida, balances the safety of personnel and space launch vehicles with lost productivity of taking shelter from lightning. The primary objective of this study investigates if this 5 NM safety radius can be reduced while maintaining a desired level of safety. The research uses processed Lightning Detection and Ranging (LDAR) data to map the movement of preexisting lightning storms using ellipses, which are updated with every lightning flash. A systematic recording ensues for the distance from the ellipse boundary of each flash occurring outside the ellipse. All of those exterior flash distances are then used to find the best-fit distribution from which the stand-off distance for the desired level of safety can be calculated. The distances from the edge of the ellipse are fit to a Weibull distribution and a new warning distance of 4 NM is selected as the most appropriate distance to balance safety and increase productivity. The 4 NM radius is tested with a resulting failure rate of .277%, with a savings of 22.5 8-hour man days a year for the months of May through September.

Table of Contents

Page

List of Figures

List of Tables

MODELING THE DISTRIBUTION OF LIGHTNING STRIKE DISTANCES OUTSIDE A PREEXISTING LIGHTNING AREA

I. Introduction

1.1 Background

Lightning is one of the most powerful and frequent natural phenomena that poses a risk to everyday life. Historical data reports that on average there are 50 lightning strikes on earth in a single second, leading to more than eight million strikes per day (Kalair et al., 2013). Particularly concerning to the safety of human life, equipment, and machines is the appearance of cloud-to-ground (CG) lightning which represents the greatest threat to life and property (Shivalli, 2016). Due to the severe danger lightning presents for both personnel and equipment, the Air Force (AF) and its civilian counterparts continually look to find the ideal balance between safety and productivity.

Located on the central eastern coast of Florida, the 45th Weather Squadron (45 WS) provides weather services to Cape Canaveral Air Force Station (CCAFS), the Kennedy Space Center (KSC), and Patrick Air Force Base (PAFB). The forecasts, weather warnings, watches, and advisories delivered by the 45 WS provide weather safety for over 25,000 personnel and over $20 Billion of resources to include facilities, boosters, and payloads (Roeder et al., 2017). While an ideal location for mission requirements, the severity and frequency of the weather patterns in this region have earned it the reputation as the thunderstorm capital of the United States (US) (Roeder et al., 2017). The more than 2,500 lightning watches and warnings per year

delivered by the 45 WS result in false alarms 40% of the time leading to many delayed or canceled launches (NASA Facts, 2006). These numbers, coupled with noted discrepancies between the process of issuing lightning watches and warnings and techniques applied in previous studies, suggest that the current 5 nautical miles (NM) safety standard set by the Air Force Instruction 91-203 (AFI 91-203) may be reduced to a distance that would incur fewer losses in man-hours while still maintaining the necessary level of safety (Department of the Air Force, 2012).

1.2 Problem Statement

The two primary lightning advisories issued by the 45 WS are lightning watches and lightning warnings. Lightning watches are issued when lightning is expected within the lightning warning circle(s) with a desired lead-time of 30 minutes and the upgrade to lightning warning occurs when lightning is imminent or occurring within the lightning warning circle(s) (Roeder et al., 2017). Currently 10 lightning warning circles exist to protect the personnel and equipment responsible for the various missions associated with the 45 WS. The radii of the warning circles vary between 5 NM for the protection of single small facilities, and 6 NM for larger facilities or a grouping of several closely located smaller facilities (Roeder et al., 2017).

Prior research on the appropriate distance of these safety buffers has been conducted in an attempt to optimize the balance between safety and operational impact. Parsons (2000) noted as motivation for her thesis nearly two decades ago that the lightning safety standards, the same 5 NM distance, resulted as an arbitrary response to lightning incidents which induced an increase in the warning distance until an appropriate balance between threat and impact was achieved. Thus, earlier research served to fill a considerable void by providing support to the 5 NM criterion for lightning warning that was previously unsubstantiated. However, weather experts at the

45 WS have found that a discrepancy still exists with the method of issuance of their lightning warnings and how past research determined the optimal stand-off distance.

Past studies that helped to solidify the AFI 91-203 safety standards of a 5 NM warning buffer utilized techniques that determined CG lightning strike distances originating from a storm's center; yet, the lightning warning process employed by the 45 WS is based on the edge of the lightning area (Roeder et al., 2017). Because of this difference, the 45 WS lightning warnings incorporate not only the required 5 NM radii, but also includes the radius of the lightning area which ranges anywhere from 3-7 NM (Roeder, 2008). While this extra distance does not impinge on the safety concerns of lightning warnings as it allows for a greater safety buffer, it does have a negative effect on productivity. The 45 WS acknowledges that because the "operational impact of the lightning warning circles is proportional to the area of the circle and thus scales as the square of the radius, even a relatively small reduction in the size of the warning circle can yield a large reduction in lost work time" (Roeder et al., 2017:8). In order to reduce the area of the warning circles, the 45 WS determined that research is required that examines the total lightning strike distances beyond the edge of a preexisting lightning area.

The primary objective of this thesis is to determine the appropriate stand-off distance from the edge of a preexisting lightning area for the desired level of safety. The overall process through which this is accomplished relies on using processed Lightning Detection and Ranging (LDAR) data to map the movement of preexisting lightning areas using ellipses which are updated with every lightning flash. A systematic recording ensues for the distance of each flash occurring outside the ellipse. All of those exterior flash distances are then used to find the best-fit distribution, potentially an extreme value distribution, from which the stand-off distance for the desired level of safety can be calculated. Though the primary consideration for this thesis is that of a

preexisting area of lightning approaching the 45 WS warning circles, an investigation of locally developing lightning, before a lightning area has formed, also presents as a topic of interest.

Because of rapid storm development in central Florida, often enough time does not always exist in order to generate a lightning warning with 30 minutes of lead time before the formation of a storm in the immediate local area. Therefore, a secondary purpose of this research, accomplished in conjunction with the previously outlined process, lies in documenting the displacement and distribution of the early flashes in a developing storm. The goal being to verify the current notion that early flashes tend to occur close to the storm's origin. Because the earlier outlined best-fit distribution is determined based on potentially greater distances from the origin of the lightning area than those found in a developing storm, the findings from the primary objective will satisfy the safety requirements necessary for those locally developing lightning storms as well. Corroborating the concept that initial flashes occur close to the storm origin will substantiate the current processes that dictate the issuance of lightning watches when preexisting lightning areas are absent.

Thus, we frame the primary research question: *what is the shortest approximate distance from the edge of a preexisting lightning area that incorporates both the necessary safety requirements and risk of being struck*? A secondary question being addressed considers: *what is the mean distance from the center of the initial flashes of lightning in a developing thunderstorm*? As a byproduct of answering these two questions, underlying questions will also be undertaken that consider the distributions regarding the number of flashes in a storm and the size of the preexisting lightning area as to aid in the justification of any changes that are made to the lightning warning circles.

1.3 Thesis Organization

The overall organization of the remaining chapters of this thesis have the following layout. Chapter II provides a review of the literature pertaining to the background of relevant lightning research as well as the potential statistical methods applied to the primary research problem. This includes an extensive look into best ellipse fitting routines, prospective extreme value distributions that may be applicable, and Goodness-of-Fit (GoF) testing techniques. In Chapter III those statistical methods introduced in Chapter II are developed and outlined with respect to the research questions. Chapter IV consists of a formal presentation of the results and analysis. Lastly, Chapter V gives a conclusion to the research with recommendations on further areas of study. All relevant acronyms introduced throughout this paper can be found in Appendix A.

II. Literature Review

2.1 Overview

This section covers a broad range of topics to include the general scientific description of a lightning strike, current lightning safety standards, previous research on the topic, current prediction methods, ellipse fitting routines, potential distributions for the distance lightning strikes outside a preexisting lightning area, and GoF testing.

2.2 The Lightning Flash

Scientists have studied the phenomenon of lightning in great detail throughout history. While advances in technology have significantly aided such endeavors, much about the true nature of lightning remains unknown. For the purpose of this study, this section contains a brief review of the more widely accepted scientific findings regarding lightning. All relevant information in this section originates from Dwyer and Uman (2014).

When the necessary physical properties to produce lightning exist, the idealized primary charge structure of an isolated, mature thundercloud consists of many tens of Coulombs of positive charge in its upper portions and a more or less equal negative charge in its lower levels. Often a smaller portion of the cloud, underneath the negatively charged section, exists as another positively charged section. A depiction of this cloud structure can be seen in Figure 1. Also seen in Figure 1 are the various types of lightning discharges that occur.

Generally the lightning events separate naturally into two categories, those flashes that connect with the ground, and those that do not. A lightning discharge that strikes the ground is known as CG lightning and will be the primary focus of this research. Lightning discharges that do not strike the ground are considered in-cloud

lightning, and based on the dispatched location, divide into intracloud, intercloud, and cloud-to-air lightning as depicted in Figure 1. The terms *lightning event*, *lightning discharge*, *lightning flash*, *lightning strike*, *lightning stroke* and *lightning* will be used interchangeably throughout the remainder of the thesis to describe both CG lightning and in-cloud lightning.

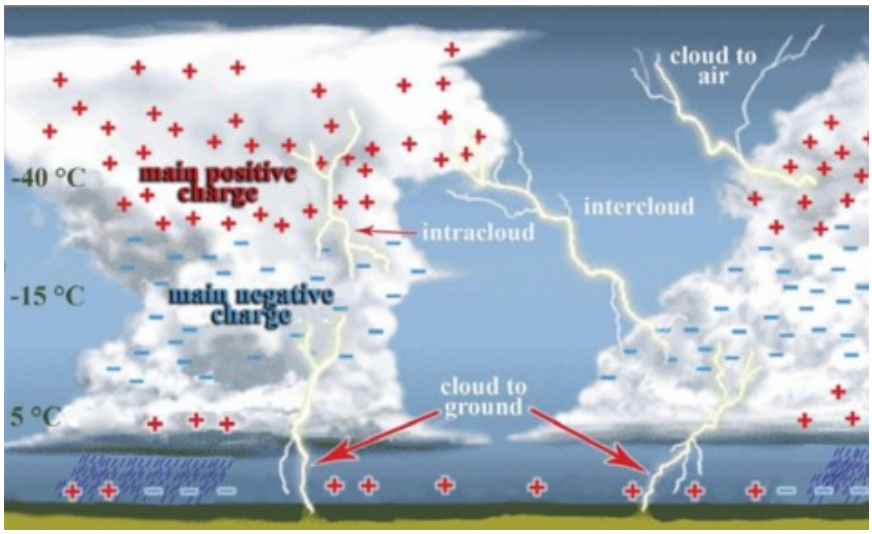


Figure 1: Charge Structure of Two Thunderclouds

Within CG lightning, there are four types that can be seen in Figure 2. The primary differences between the four forms lie in the sign of the electrical charge carried in the initial "leader" and by the direction of propagation of that leader. Negative CG lightning flashes portrayed in Figure 2a result from the lowering of a negative charge from the negatively charged portion of the cloud to the ground. About 90% of CG flashes are negative downward strokes; whereas close to 10% of CG lightning flashes originate from the positively charged region of the cloud resulting in a positively charged downward stroke as seen in Figure 2c.

The other portion of CG lightning in Figure 2b and 2d, technically ground-to-cloud lightning, are relatively uncommon and are upward initiated from mountaintops, tall man-made towers, or other tall objects, towards the cloud charge regions.

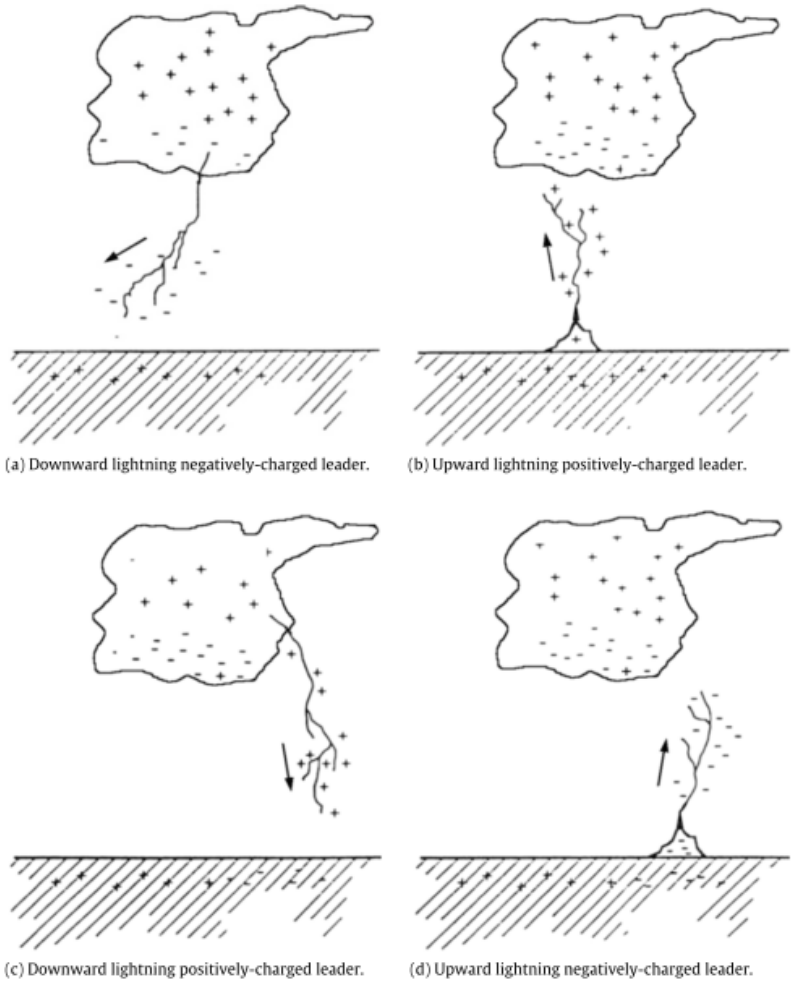


Figure 2: Four Types of CG Lightning Flashes

Given the varying types of CG lightning, the development of a lightning strike is introduced in terms of the negative CG flash as this is the most frequent manifestation; illustrations of the entire process can found in Appendix B. The negative charge of a CG lightning strike exhibits as an electrical discharge known as a *step leader* which moves downward from cloud to ground in discrete segments. After approaching the ground, the large negative charge attracts a positive charge from the Earth's surface, and, when the electric field intensity near the ground from these charges becomes large enough, upward-going, positively-charged electrical discharges from the ground or from grounded objects will be initiated. When the two charges connect, the downward

step leader and the upward charge, the lightning strike point is determined as well as the primary channel of strike between cloud and ground. This initial process, known as the first *return stroke*, may conclude the lightning flash if sufficient negative charge does not exist in the upper portion of the previous stroke channel within about 100 ms from the cessation of the current of the previous stroke. Nearly 80% of negative CG flashes in temperate regions will have more than one stroke, typically three to five.

If adequate negative residual charge does exist, another charge leader, known as the *dart leader*, travels down the return stroke channel with a negative charge, providing the necessary electricity for any subsequent return strokes. Because of the formation of new paths to the ground achieved by the dart leaders, a large portion of all CG lightning strikes hit the Earth in different locations. With this process in mind, we next consider the safety standards set forth to protect life and property from the damaging effects of lightning.

2.2.1 Safe Distance Criterion

In January of 1998 during the American Meteorological Society's (AMS) Annual Meeting, a Lightning Safety Group (LSG) met to discuss the inconsistent lightning safety sanctions as well as new developments in lightning knowledge (Holle et al., 1999). Referencing advances in the understanding of thunderstorm behavior discovered by López and Holle (1999), the meeting resulted principally in the creation of the *30-30 rule* (Holle et al., 1999). Now the standard recommendation in lightning safety adhered to by agencies such as the Center for Disease Control (CDC), National Aeronautics and Space Administration (NASA), National Weather Service (NWS), and the Occupational Safety and Health Administration (OSHA), the 30-30 rule provides safety guidance for both the onset and cessation of lightning.

The first 30 of the 30-30 rule refers to the number of seconds between a lightning flash and the subsequent sound of thunder. Every five seconds of time lapse between the lightning flash and the following thunder clap indicates a mile of distance from the actual strike. Based on their findings of the distance subsequent flashes of lightning can travel, López and Holle (1999) suggested 6-8 statute miles (M) was a safer distance for resigning outdoor activities from the previous 2-3 M, which lead to the creation of the first half of the 30-30 rule by the LSG (Holle et al., 1999). Thus, standard guidance suggests suspending outdoor activities when lightning occurs within 6 M (5.21386 NM), that is, when the time measurement between flash and thunder reaches 30 seconds or less. The second 30 of the rule references the number of minutes to wait from the last lightning flash or thunder to give a clearance to resume outdoor activities. Although the creation of the AF lightning guidance does not have as defined a lineage, the practices follow a similar pattern.

During the same time frame, the AF also sought to update lightning safety standards based on research studies. The lightning strike and subsequent death of an airman in 1996 at Hurlburt Field prompted the beginning of the investigation and resulted in several studies, which are outlined in the next section, on the distance CG lightning can travel (Cox, 1999). After these studies, which used different techniques, provided similar results on lightning strike distance, the AF safety standard settled on 5 NM as the radius of optimum safety.

As previously mentioned, the 45 WS bases their weather watches and warnings on the broader AF safety regulation, AFI 91-203 which states that for any predesignated locations or activities, a lightning watch is in effect 30 minutes prior to a thunderstorm being within a 5 NM radius (Department of the Air Force, 2012). A lightning watch does not constitute a halt in activities but paves the way for the official lightning warning. A lightning warning goes into effect once lightning occurs within a 5

NM radius of predetermined locations and activities and mandates that personnel in affected locations or engaged in affected activities disengage from outdoor activity and seek shelter (Department of the Air Force, 2012).

Cessation of AF lightning warnings follows similar protocols to the second half of the 30-30 rule with the exception that a cancellation of a lightning warning does not occur if the potential exists for follow-on lightning activity to transpire within 30 minutes (Department of the Air Force, 2012). The current standards have had the desired effect in regards to maintaining the safety of AF life and property. However, the methods used in the studies that helped to form the regulations, described in the following section, do not concur with the issuance processes that result in actual watches and warnings, leading to the primary purpose of this research.

2.3 Previous Research

In past research, several different techniques have been used to determine the horizontal strike distance of CG lightning. These methods include the Weather Surveillance Radar - 88 Delta (WSR-88D) storm centroid, Distance Between Successive Flash (DBSF), and a LDAR and National Lightning Detection Network (NLDN) data pairing method.

2.3.1 WSR-88D Storm Centroid Method

The first of these techniques, the WSR-88D storm centroid, uses the National Severe Storms Laboratory (NSSL) Storm Cell Identification and Tracking (SCIT) algorithm and the WSR-88D storm series algorithm (Parsons, 2000). Once a storm centroid has been identified via the algorithm, the second step of the process superimposes the lightning strike data and calculates the horizontal distance between lightning strike and storm centroid. Used in several studies, this method provided

an initial answer to the question of how far horizontally a CG lightning strike travels but neglected the concept of storm origin as no true positive identification could be made in reference to which thunderstorm the lightning strike actually originated.

Cox (1999) employed the WSR-88D storm centroid method using Build 10.0 of WSR-88D Algorithm Testing and Display System (WATADS) and compared it to the DBSF method for two months worth of data, April and July of 1996. Results for the WSR-88D portion of Cox's analysis returned, for the months of April and July respectively, 39% and 32% of lightning strikes occurring at a distance greater than 5 NM. Conducted a year earlier in 1998 by Renner, a similar study using Build 9.0 of WATADS for the same time frame, showed 75% of all lightning flashes were within 10 NM for April, and 85% to 90% were within 10 NM for July. As well as presenting a difficulty with identifying storm origin for each lightning strike, Cox (1999) noted that the WSR-88D method required an extreme amount of time and disk space to process the data.

2.3.2 DBSF Method

The second method, DBSF, relies on NLDN data and groups strikes as clusters based on spatial and temporal criteria. Different studies determined varying time and spatial constraints on the grouping process. Once grouped, a lightning centroid calculation ensues by taking the mean distance of each flash in the cluster. Subsequently, the distance of each strike results from the distance an individual strike occurs from that centroid. One issue discovered with this method lay in the number of isolated flashes that appeared; an isolated flash refers to one that has no assigned cluster as no other flashes met the spatial and temporal criteria to constitute a grouping. Another concern, the same as the primary concern with the WSR-88D method, is that the true origin of the flash could not be identified by this method.

Earlier referenced as the study that drove the 30-30 rule, López and Holle (1999) employed the DBSF algorithm grouping flashes with the constraint that successive flashes occur within five minutes and no more than 15 kilometers (KM) apart. As previously mentioned, this study resulted in an updated assessment that to ensure adequate personal safety, outdoor activity should desist at a distance of 6 M from the last lightning strike. Cox (1999) applied the DBSF method in conjunction with the WSR-88D method with the constraint of grouping clusters based on flashes falling within either 6 minutes of time from the first flash of the group and within a distance of 15 KM. Providing similar results to the WSR-88D method, Cox (1999) found that 34.41% of strikes for the April data fell beyond the 5 NM safety distance and 18.04% of July's strike data had the same result.

By far the most comprehensive study to date conducted using the DBSF method, Parsons (2000) focused on reducing the number of isolated flashes by increasing the clustering criteria to 15 minutes and 17 KM apart. The data used spanned from 1995 to 1999 and covered nearly the entire continental United States; defined for the study, six regions separated this area and analysis ensued for each region. Specifically, Region Five included the central eastern coast of Florida. Besides increasing the amount of time and distance criteria that determined flash clustering, this study included the isolated flashes in the analysis and found that in Region Five particularly, "the probability of an isolated flash striking an exact location does not drop below 5% until 19 KM for the isolated flashes" (Parsons, 2000:45). Each region produced similar results in regards to the clustered strikes as well; reports for Region Five found that lightning occurred at distances greater than 5 NM between 21% and 32% of the time.

2.3.3 LDAR and NLDN Pairing Method

To correct for the issue of storm origin discussed in the previous two methods, McNamara (2002) utilized a method of pairing flashified LDAR data with the ground strike data provided by NLDN. The process begins by taking LDAR data, which is described in further detail in a later section, and using a flash-grouping program to group the LDAR data points into flashes. To include data points in a flash, the program detects if a flash falls within a 3 second window from the time of the first data point in the flash. Another requirement of a data point being included in a flash is that it must fall within 0.5 seconds of the previous data point in the flash. A distance limit also factors into the determination of whether or not a data point belongs to a specific lightning flash. Once flash grouping concludes, the procedure of branch grouping begins. Branch data points fall within .03 seconds of each other and also rely on similar spatial constraints as those set for the flash grouping.

After all LDAR data has been processed, the pairing of LDAR flashes to NLDN strike locations begins. To pair origin and strike data, the process uses an Interactive Data Language (IDL) program that compares NLDN strike locations and the timing and location of the flashified LDAR data. Again, time and spatial constraints played a role in the pairing as a flash is only be considered as the origin of the ground stroke if the LDAR data point occurs within 1 second and 50 KM of the NLDN detected ground strike. After identifying the flash that produced the ground strike, the program proceeded to calculate the horizontal distance of the strike from the origin.

The data were analyzed both as a whole and seasonally. The seasonal analysis included a breakdown of the data by months, creating the four seasons, Summer, Fall, Winter, and Spring. Binning the data based on distances, a frequency distribution was created to display the number of all the flashes occurring at each distance. McNamara

(2002) included in the study over one and a half million lightning strokes during the period of March 1997 to December 2000. Evaluating the data in its entirety produced 28.4% of strikes extending beyond 5 NM (McNamara, 2002). The seasonal data produced anywhere from 18.7% to 38.5% of strikes occurring outside the 5 NM distance. Thus, along with the other studies by Cox (1999) and Parsons (2000) who sought specifically to verify the AF's 5 NM lightning standards, McNamara (2002) concluded that a high percentage of strikes occur at distances greater than 5 NM.

While limitations also existed with McNamara's (2002) research as some assumptions made could potentially have lead to very conservative estimates of the true horizontal distance CG lightning travels, his research filled a considerable hole in regards to determining the origin of a lightning strike in conjunction with its strike location. However, all of the studies outlined looked at determining the distance lightning strikes travel from storm origin when actual forecasting methods execute watches and warnings based on the edge of the storm or preexisting lightning area.

2.3.4 Current Methods of Lightning Strike Prediction

Several methods of lightning strike prediction exist in the meteorological community. The majority rely on radar and local sensors, when available, to detect the physical characteristics associated with active or developing thunderstorms. Historical data plays an important role as factors observed during previous lightning events have been studied to detect the exact environment or conditions that generate lightning capable clouds. Private companies in particular strive to be on the cutting edge of lightning prediction techniques, primarily due to the extreme danger lightning presents to life and property and the monetary opportunities that danger imposes. However, government and research based organizations also continuously look to improve upon current methods.

Many existing detection systems commercially available attempt to determine the path of lightning by tracking previous strikes with NLDN data or by measuring electrical activity in the atmosphere in a specific location. One company, WeatherData, Inc., patented a lightning detection system that uses "an algorithm which analyzes radar data to locate areas where cloud tops extend above a predetermined temperature threshold and have sufficient radar reflectivity" (WeatherData, Inc., 2018:1). This leads to detecting cloud particles with higher electrical charge, indicating the likelihood of lightning. Other studies have looked to create a set of forecast equations designed to predict lightning strike by a statistical process that uses Principal Component Analysis (PCA) with logistic regression (Bothwell and Richardson, 2014). Regardless of the studies that exist, very few weather services or local weather forecasters provide lightning forecasts or warnings.

Researchers at the University of Oklahoma's Cooperative Institute for Mesoscale Meteorological Studies and the National Oceanic and Atmospheric Administration's (NOAA) NSSL perceived this absence and have developed a system that attempts to bridge this gap (Calhoun et al., 2018). Essentially this model considers storm conditions in the local area and uses historical data from past storms in a machine learning random forest algorithm that generates a single probability of lightning. Trials have been conducted with NWS forecasters, emergency managers, and broadcast meteorologists in an effort to test the usability of the software; currently, the Graphical User Interface (GUI) still requires further simplification to be considered an effective lightning prediction tool.

Although the civilian market offers various techniques for lightning detection and prediction, the military also boasts an array of cutting edge technology. Because so many sensitive and expensive systems exist within the Department of Defense (DoD), the military, and other government agencies, have consistently seen the need

for lightning detection systems to protect both personnel and equipment. The 45 WS in particular hosts one of the most advanced lightning detection systems in terms of the shear number of sensors available for data collection. While the primary tool for forecasting locally developing thunderstorms lies in traditional radar, the 45 WS has also historically housed and currently maintains a series of detection tools to assist in forecasting.

These tools include the LDAR system and the Cloud-to-Ground Lightning Surveillance System (CGLSS) both utilized prior to 2008. Then during the period of 2008-2016, forecasters relied on the Four Dimensional Lightning Surveillance System (4DLSS), which consisted of upgraded and combinatorial use of the LDAR network and the CGLSS. Finally, most currently in use since 2016 is the Mesoscale Eastern Range Lightning Information Network (MERLIN). The Launch Pad Lightning Warning System (LPLWS) and the NLDN also assist in lightning detection at the 45 WS. These various technologies, combined with the two weather radar systems and satellite imagery, allow forecasters to make the most informed decisions using forecaster techniques known as *The Pinder Principles* (Roeder and Pinder, 1998). Based on The Pinder Principles, several radar rules determine the issuance of lightning watches and warnings. Phenomena concentrating on cloud and lightning type in conjunction with temperature and elevation levels play a pivotal role as well as local convergence and consequent convection (Roeder and Pinder, 1998).

Regardless of the technology or methods being applied to lightning forecasting, an alignment of prediction techniques and the distance CG lightning travels proves vital. All previous methods of determining the horizontal distance CG lightning travels concentrated on the distance from the center of the storm. However, the systems described in this section generate watches and warnings based on the edge of the storm or preexisting lightning area. Therefore, this research seeks to resolve this

discrepancy by applying a new technique that considers the distance lightning travels from the boundary of the preexisting lightning area.

2.4 Lightning Data Source

We now turn to a brief review of the system from which the data for this study is procured. Several references to the LDAR network have already been made. The following section outlines the origin of the LDAR-I system and discusses the updates made to that system to create the LDAR-II network that supplies the data for this research.

2.4.1 Lightning Detection and Ranging (LDAR) System

Developed in 1971 by NASA engineer Carl Lennon and considered one of the first operational systems for detecting in-cloud lightning, the LDAR system has served the operational and research needs of KSC and CCAFS for over four decades (Starr et al., 1998). The original LDAR-I system employed seven Very High Frequency (VHF) radio receivers with 66-MHz central frequency and 6-MHz bandwidth, six of the receiver stations surround one central site at distances ranging from 6 to 10 KM (Starr et al., 1998).

When upgrades were made to the system to create the LDAR-II network in 2008, the older receivers were replaced with nine LDAR-II sensors (Roeder, 2010). A depiction of the site locations for LDAR-II can be seen in Figure 3. Unless otherwise cited, the remainder of the information about the LDAR data set in this section originates from Roeder (2010) and any use of the term LDAR going forward refers to the LDAR-II system.

The LDAR system senses the radio pulses generated by lightning step leaders as well as other in-cloud lightning mechanisms and operates by identifying the electro-

magnetic pulses emitted by those lightning events. The difference in time-of-arrival (TOA) of these radio pulses at pairs of LDAR-II sensors is used to calculate a hyperbolic volume. Then, the intersection of four different hyperbolae locates the step leader in three dimensions. As the nine LDAR sensors provide numerous locations for a single event, in order to produce a single 3-dimensional (3D) location and time of each lightning strike, the best location of the step leader is procured using a statistical Chi-Squared minimization technique.

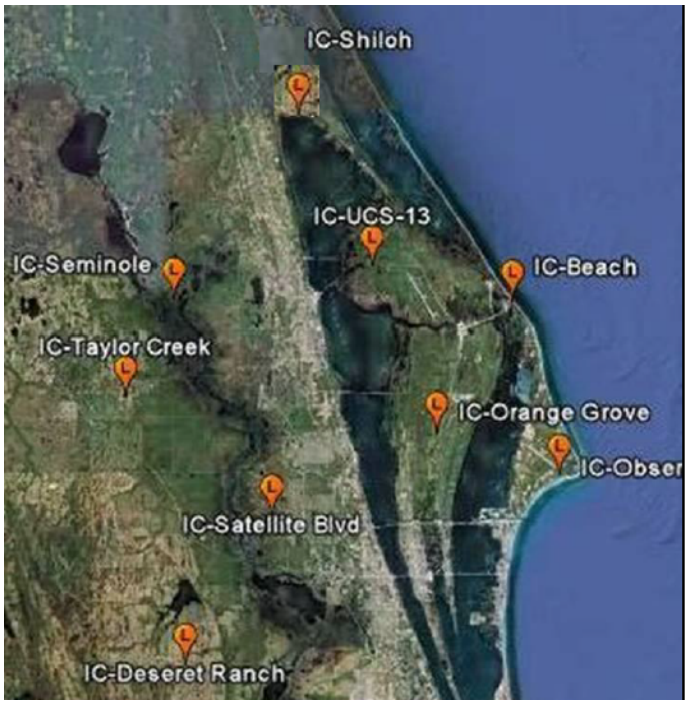
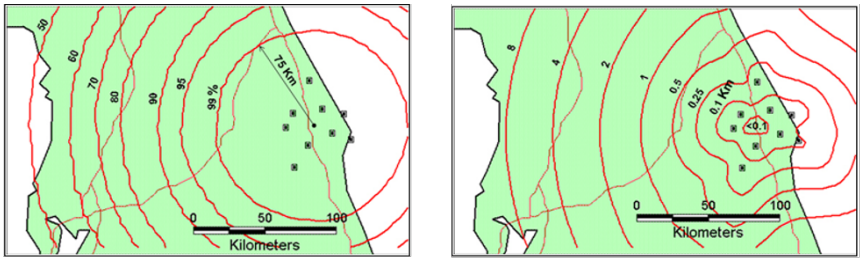


Figure 3: LDAR-II Site Locations

An accuracy test conducted in 1995 of the legacy LDAR-I system found that when fully operational, flash detection efficiency neared 100% and false alarm rates fell to less than 1% (Mata and Wilson, 2012). In his 2010 review of the upgraded system, Roeder confirms that the updated LDAR network's performance exceeds its earlier edition with a 140% increase in detection. As well as having excellent efficiency, the LDAR system offers an indication of lightning activity 10 to 20 minutes earlier than other systems. The data from LDAR also provides deeper knowledge into the origin

of a lightning flash as it presents a 3D view of numerous data points (Britt et al., 1998). These advantages to LDAR data as opposed to other systems serve as the primary justification for using LDAR data in this study.

Several limitations to the LDAR system are important to note here. Primarily, the LDAR system does not relay CG lightning strikes. The 4DLSS previously mentioned as a tool for lightning detection by the 45 WS relies on the updated CGLSS portion to provide actual CG strike locations. The second item of concern is the limiting factor of location accuracy and detection rate as the radius of inclusion expands beyond the central cluster of sensors. These decreases in detection rate and accuracy as the radius expands can be seen in Figure 4 under the assumption that all nine sensors are used in the solution. The first of these limitations will inform certain assumptions made during the research process, while the second will help to inform the choices of data reduction for this study.



a) LDAR Detection Rate

b) LDAR Location Accuracy (50% confidence)

Figure 4: Performance of LDAR across East Central Florida

2.5 Ellipse Fitting Methods

One aspect of this research includes defining a boundary around a preexisting lightning area in order to characterize the edge of the storm from which a lightning watch or warning is issued. This process is achieved using ellipses as the boundary. In this section we outline the various parameterizations of ellipses and discuss three

different fitting methods to include least squares best fit ellipses, PCA confidence ellipses, and minimum area ellipses.

2.5.1 Ellipse Parameterizations

An ellipse is defined as the set of all points (x, y) in a plane, the sum of whose distances from two distinct fixed points (foci) is constant. An ellipse has five parameters including the center coordinates, two semiaxes, and the tilt. In Figure 5 the left image depicts the location of the foci of an ellipse, while the right image shows the major and minor axes along with the center and two sets of vertices. The center is the midpoint of the major axis and the minor axis is perpendicular to the major axis at the center of an ellipse. Though not depicted in Figure 5, the tilt of an ellipse refers to the rotation of an ellipse with respect to the X and Y axes of the coordinate plane. The various algorithms used to fit ellipses often rely on different parameterizations of an ellipse.

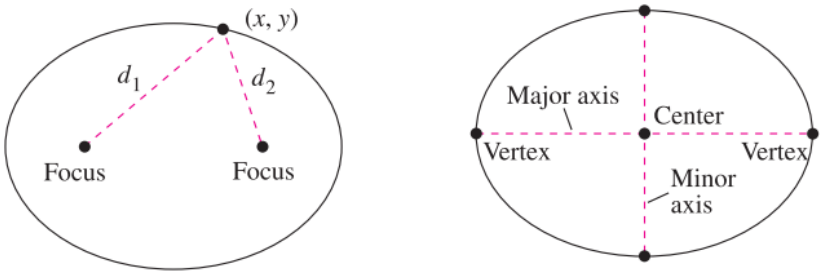


Figure 5: Components of an Ellipse

Van Loan (2006) describes the following parameterizations of an ellipse as well as conversions from one form to the other. First, the conic representation of an ellipse is defined as the set of points (x, y) that satisfy

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \tag{1}$$

if $B^2 - 4AC < 0$. To avoid degenerate ellipses, it is also required that $\frac{D^2}{4A} + \frac{E^2}{4C} - F > 0$. While the conic representation is useful, we primarily define ellipses using the parametric form.

In its most basic form, the parametric representation consists of four of the five parameters of an ellipse with center (h, k) and semiaxes a and b:

$$\left(\frac{x-h}{a}\right)^2 + \left(\frac{y-k}{b}\right)^2 = 1 \tag{2}$$

Equation 2 can also be written as the set of points $(x(t), y(t))$ where

$$\begin{aligned} x(t) &= h + a\cos(t) \\ y(t) &= k + b\sin(t) \end{aligned} \tag{3}$$

and $0 \leq t \leq 2\pi$. To include the tilt of the ellipse in Equation 3, a counter-clockwise rotation by τ radians takes the form:

$$\begin{aligned} x(t) &= h + \cos(\tau)[a\cos(t)] - \sin(\tau)[b\sin(t)] \\ y(t) &= k + \sin(\tau)[a\cos(t)] + \cos(\tau)[b\sin(t)] \end{aligned} \tag{4}$$

Thus, Equation 4 gives the parametric representation of an ellipse with all five parameters.

The last representation is the foci/string form. With two foci points $F_1 = (x_1, y_1)$ and $F_2 = (x_2, y_2)$, let s be a positive number greater than the distance between them. Then the set of points (x, y) that satisfy

$$\sqrt{(x-x_1)^2 + (y-y_1)^2} + \sqrt{(x-x_2)^2 + (y-y_2)^2} = s \tag{5}$$

defines an ellipse. Appendix C includes a table of these parameterizations as well as

the various methods of conversion from one form to another. Now that the various parameterizations of an ellipse have been defined, we turn to the different methods of fitting ellipses.

2.5.2 Least Squares Best Fit Ellipses

Because of the various applications of pattern recognition to which the shape of an ellipse can be applied, the process of finding a best fit ellipse has been widely studied. The different approaches may be divided into two broad categories, often known as the geometric and algebraic approaches. The geometric methods include the use of such techniques as the Hough transform and fuzzy clustering; these methods are robust to outliers but are slow and expensive computationally requiring large memory while producing lower levels of accuracy (Halíř and Flusser, 1998). In contrast, the algebraic methods are based on "optimization of an objective function which characterizes a goodness of a particular ellipse with respect to the given set of data points"(Halíř and Flusser, 1998:1). These algebraic methods provide greater speed and accuracy but have a higher sensitivity to outliers than the geometric methods. The least squares approach falls into the category of algebraic methods.

The premier work on the least squares approach from which more refined techniques have developed was done by Fitzgibbon et al. (1996). For the purposes of outlining the least squares best fit approach, we concentrate this section on the fitting routine that became known as Fitzgibbon's technique and reference the 1996 paper on the results of those efforts. Methods used prior to this technique for general conic fitting, even for ellipse specific data, would often result in hyperbolas and parabolas. Thus, the new method sought to apply a constraint to the general conic fitting least squares approach that would force the resulting conic to be an ellipse.

The general conic least squares approach to fitting an ellipse begins by taking

Equation 1 subject to the ellipse specific constraint

$$B^2 - 4AC < 0 \tag{6}$$

The polynomial form can be rewritten with vectors

$$\begin{aligned} \boldsymbol{a} &= [A, B, C, D, E, F]^T \\ \boldsymbol{x} &= [x^2, xy, y^2, x, y, 1] \end{aligned} \tag{7}$$

in the vector form

$$F_{\boldsymbol{a}}(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{a} = 0 \tag{8}$$

For a general conic, fitting a set of points (x_i, y_i) for $i = 1, ..., N$ to the given data may then be achieved by minimizing the sum of squared algebraic distances of the points to the conic which is represented by coefficients $\boldsymbol{a}$:

$$\begin{aligned} \min_{\boldsymbol{a}} \sum_{i=1}^{N} F(x_i, y_i)^2 &= \min_{\boldsymbol{a}} \sum_{i=1}^{N} \left(F_{\boldsymbol{a}}(\boldsymbol{x_i})\right)^2 \\ &= \min_{\boldsymbol{a}} \sum_{i=1}^{N} (\boldsymbol{x_i} \cdot \boldsymbol{a})^2 \end{aligned} \tag{9}$$

Equation 9 can be solved by the standard least squares approach but the results may not necessarily produce an ellipse. Therefore, a specific constraint is introduced to ensure an ellipse-specific solution; that is, the inequality of Equation 6 becomes an equality constraint

$$4AC - B^2 = 1 \tag{10}$$

Now the ellipse-specific fitting problem can be redefined as

$$\min_{\boldsymbol{a}} ||\boldsymbol{D}\boldsymbol{a}||^2 \quad \text{subject to} \quad \boldsymbol{a}^T \boldsymbol{C} \boldsymbol{a} = 1 \tag{11}$$

The minimization problem from Equation 11 may now be solved by a quadratically constrained least squares minimization. Applying the Lagrange multipliers gives the system

$$\begin{aligned} \boldsymbol{S}\boldsymbol{a} &= \lambda \boldsymbol{C}\boldsymbol{a} \\ \boldsymbol{a}^T \boldsymbol{C}\boldsymbol{a} &= 1 \end{aligned} \tag{12}$$

where $\boldsymbol{S} = \boldsymbol{D}^T\boldsymbol{D}$. This system from Equation 12 can then be solved using generalized eigenvectors where there exist up to six solutions. However, the overall solution is chosen to yield the minimal positive eigenvalue. The design matrix $\mathbf{D}$, constraint matrix $\mathbf{C}$, and scatter matrix $\mathbf{S}$ from this process can be seen in Appendix D. Figure 6 shows the results of employing Fitzgibbon's method to fit an ellipse to a series of lightning flashes. From Figure 6 we observe the fit of an ellipse to data points, where by the least squares method, the distance from the edge of the ellipse to the points is minimized. For all plots moving forward, including Figure 6, the X and Y axes are the distance (KM) from the legacy central LDAR site located at (0,0). Another relevant method that exists for fitting ellipses is the application of confidence ellipses using PCA.

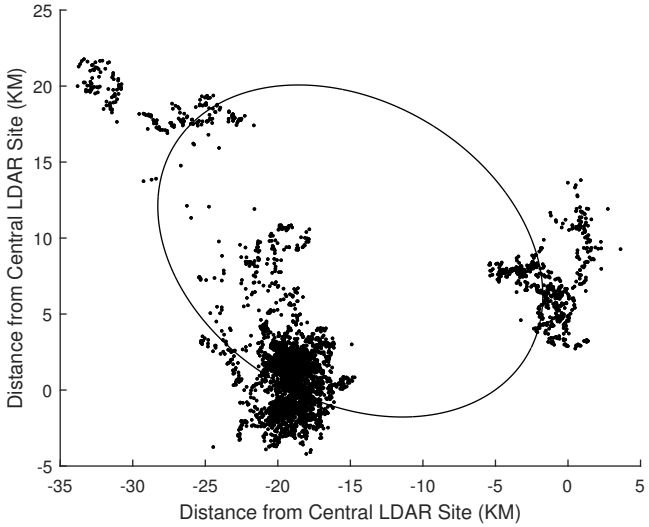


Figure 6: Example of Least Squares Fitzgibbon Ellipse Specific Method

2.5.3 PCA Confidence Ellipses

A commonly applied multivariate statistical tool, PCA seeks to determine the principal components of a data set that explain the greatest amount of the total variation in the data. The first principal component extracted accounts for the largest amount of the total variation in the data and "gives the direction in which the variance of the data is maximal"(Wijewickrema and Papliński, 2004:1). The second principal component, orthogonal to the first, "accounts for the maximum amount of the remaining total variation not already accounted for by the first principal component"(Dillon and Goldstein, 1984:25). Employing the concepts of PCA with the assumption of an underlying multivariate normal distribution allows for the construction of confidence ellipses around a given set of data points. The following outline of the application of PCA to create confidence ellipses originates from the 2004 work of Wijewickrema and Papliński.

We start with a set of data, X, which consists of n m-dimensional vectors

$$X = \begin{bmatrix} x(1) \\ \vdots \\ x(n) \end{bmatrix} \tag{13}$$

where $x(k) = [x_1(k)...x_m(k)]$. Because we are only working with 2D data, $m = 2$, thus the matrix X is $n \times 2$ and consists of the coordinates of all the points for which we are trying to draw an ellipse. The covariance matrix S of the 2D vectors provides a measure of the strength of the correlation between components and can be calculated by

$$S = \frac{1}{n-1} X^T \cdot X \tag{14}$$

We also note that S is a symmetrical positive-definite matrix and X is not a zero matrix.

Once the covariance matrix is obtained, we next find the eigenvectors v and corresponding eigenvalues λ such that

$$S \cdot v = \lambda \cdot v \tag{15}$$

Since S is a positive definite matrix, S will only have real eigenvalues. Therefore, we can use matrix decomposition on Equation 15 as follows

$$S \cdot V = V \cdot \Lambda, \quad \text{or} \quad S = V \cdot \Lambda \cdot V^{-1} \tag{16}$$

where Λ is a diagonal matrix of eigenvalues and V is a matrix whose columns are the corresponding eigenvectors. We also know that because the eigenvectors are orthogonal to each other and are of unit length, the matrix of eigenvectors is an orthogonal matrix, meaning we can rewrite Equation 16

$$S = V \cdot \Lambda \cdot V^{T} \tag{17}$$

From this process we construct the components of the confidence ellipse. The center of the ellipse can be calculated as the mean of all the data points. We then take the direction of eigenvector one, the first principal component, to be the major axis of the ellipse and the direction of eigenvector two, the second principal component, to be the minor axis of the ellipse. The lengths of the major and minor axes are then determined by using the eigenvalues that correspond with the principal components (eigenvectors), λ_1 and λ_2 respectively. The size of the ellipse will depend upon the level of confidence desired where higher levels of confidence result in larger ellipses. To calculate the lengths of the axes, a_i, we use the following

$$a_i = \sqrt{z \cdot \lambda_i} \quad i = 1, 2 \tag{18}$$

where z is a z-score for a specified α level with m and $n - m$ degrees of freedom. Figure 7 gives a depiction of this method at varying levels of confidence on a series of

lightning flashes. In Figure 7, each PCA ellipse is actually being drawn based on the convex hull of the data points, the reasoning for this is discussed further in Chapter III.

While PCA Ellipses may be effective given certain objectives, there often exists the need to bound a set of points by the smallest possible ellipse. Although PCA Ellipses may accomplish this bounding objective incidentally given a small enough α value, a separate method exists dedicated entirely to this objective. Therefore, in the next section we discuss the minimum area ellipse fitting method.

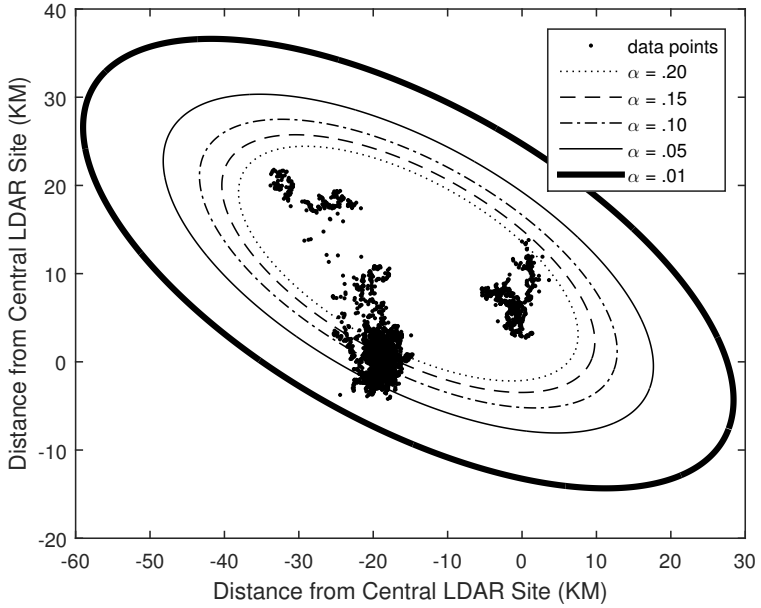


Figure 7: PCA Confidence Ellipse Results for Varying Levels of α

2.5.4 Minimum Area Ellipses

Applicable to fields such as optimal design, computational geometry, convex optimization, computer graphics, pattern recognition, and statistics, the problem of finding a minimum volume enclosing ellipsoid (MVEE) has been studied for decades resulting in a variety of algorithms (Todd and Yildirim, 2007). These algorithms fall into several categories including first-order algorithms, second-order interior-point al-

gorithms, or combinations of these two types of algorithms. For the purposes of this review, we focus our attention on the overall optimization problem of defining a MVEE and an example of one specific algorithmic implementation, the Khachiyan algorithm. The following outline originates from an overview of the Khachiyan algorithm by Todd and Yildirim (2007).

First, we introduce the notation and formulation of the problem. A full-dimensional ellipsoid $\mathscr{E}_{Q,c}$ in $\mathbb{R}^d$ is specified by a $d \times d$ symmetric positive definite matrix Q and a center $c \in \mathbb{R}^d$ represented by

$$\mathscr{E}_{Q,c} = \{x \in \mathbb{R}^d : (x-c)^T Q(x-c) \leq 1\} \tag{19}$$

The volume of the ellipsoid $\mathscr{E}_{Q,c}$, denoted by $\text{Vol}(\mathscr{E}_{Q,c})$, is given by $\text{Vol}(\mathscr{E}_{Q,c}) = \rho \det Q^{-1/2}$, where ρ is the volume of the unit ball in $\mathbb{R}^d$. We then define the scaled volume by

$$\text{Vol}(\mathscr{E}_{Q,c}) := \det Q^{-1/2} \tag{20}$$

Now, let $\mathscr{A} := \{a^1, ..., a^m\} \subset \mathbb{R}^d$ be a finite set of vectors whose affine hull is $\mathbb{R}^d$. We will assume $\mathscr{A}$ is centrally symmetric for the sake of simplicity but note that there is a lifting operation defined by Todd and Yildirim (2007) to induce central symmetry when $\mathscr{A}$ is not centrally symmetric.

As we are assuming $\mathscr{A}$ is centrally symmetric, then MVEE($\mathscr{A}$) is centered at the origin. If $\mathscr{A} \subset \mathscr{E}_{Q,c}$ where $c \neq 0$, then $\mathscr{A} \subset \mathscr{E}_{(1/\theta)Q,0}$, where $\theta := 1 - c^T Q c < 1$, which implies that the latter ellipsoid has a smaller volume by Equation 20. Thus, the problem of computing MVEE($\mathscr{A}$) can be defined as the following convex optimization problem

$$\begin{aligned} \big(\mathbf{P}(\mathscr{A})\big) \min_M \quad & -\log \det M \\ \text{s.t.} \quad & (q^i)^T M q^i \leq 1, \quad i = 1, ..., m, \\ & M \in \mathbb{R}^{n \times n} \text{ is symmetric and positive definite} \end{aligned} \tag{21}$$

where $M \in \mathbb{R}^{n\times n}$ is the decision variable. The Khachiyan algorithm focuses on the Lagrangian dual of $\big(\mathbf{P}(\mathscr{A})\big)$ which is equivalent to

$$\begin{aligned} \big(\mathbf{D}(\mathscr{A})\big) \ \max_p \quad & \phi(p) := \log\ \det\ \Lambda(p) \\ & \text{s.t.} \quad e^T p = 1, \\ & \qquad p \geq 0 \end{aligned} \tag{22}$$

where $p \in \mathbb{R}^m$ is the decision variable and $\Lambda : \mathbb{R}^m \to \mathbb{R}^{n\times n}$ is the linear operator given by

$$\Lambda(p) := \sum_{i=1}^{m} p_i q^i (q^i)^T \tag{23}$$

The Khachiyan algorithm seeks a minimum-volume ellipsoid containing $\mathscr{L} := \text{conv}\{\pm q^1, ..., \pm q^m\}$. However, since it is a dual algorithm, it constructs a sequence of ellipsoids $\mathscr{E}_k := \{y \in \mathbb{R}^n : y^T \Lambda(p^k)^{-1} y \leq 1\}$ satisfying $\mathscr{E}_k \subseteq \mathscr{L}$ and stops when $\mathscr{L} \subseteq \sqrt{(1+\epsilon)n}\mathscr{E}_k$, that is when a pre-specified tolerance is reached. An example of an execution of the Khachiyan algorithm with a tolerance of .01 can be seen in Figure 8.

From Figure 8 we see the implementation of a 2-Dimensional (2D) MVEE where the ellipse is the smallest ellipse enclosing a given series of lightning flashes. A careful examination of Figure 8 yields the discovery that several data points fall outside the MVEE; this is due to the specific tolerance of .01 used when running the algorithm. A smaller tolerance would increase the accuracy of the algorithm at the cost of increasing the computational speed of the algorithm. Further justification for the selection of a .01 tolerance is given in Chapter III.

Because of certain limitations of the first two methods discussed, we choose to incorporate the latter method. Although we are only working with 2D data, minimum area ellipses will continue to be referred to throughout the rest of this text as MVEE

as the generalized concept for multidimensional use, whether 2D or greater, is commonly referenced in literature as MVEE. Further details on the MVEE method as the selection of choice are given in Chapter III. We now consider the various distributions that may be applied to the resulting array of distances collected that occur beyond the boundaries of the ellipses.

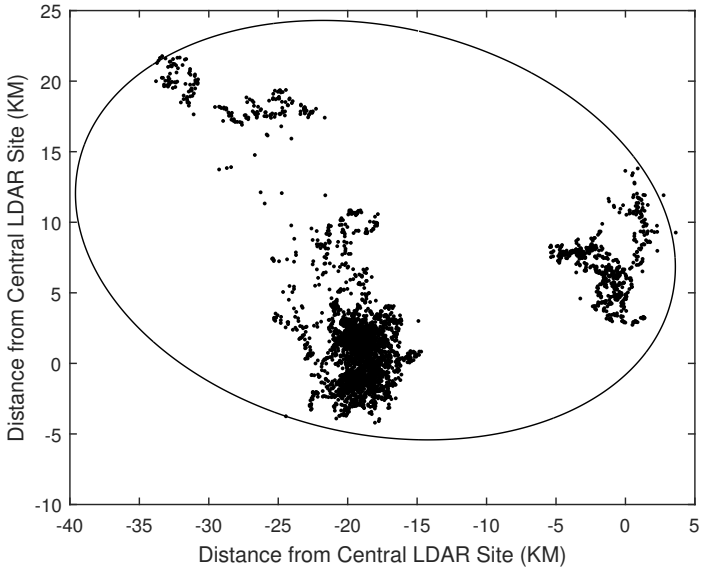


Figure 8: Example of Minimum Area Ellipse using the Khachiyan Algorithm

2.6 Potential Distributions

Due to the nature of the phenomenon being studied, a different set of distributions must be considered when hypothesizing the form of the distribution of lightning strikes outside a preexisting lightning area. This section overviews the Generalized Extreme Value (GEV) distribution and the three GEV distribution sub-types.

2.6.1 Extreme Value Distributions (EVD)

The GEV continuous probability distribution originates from the Extreme Value Theorem (EVT) which states that a given function $f(x)$ that is continuous on the

closed interval $[a, b]$ has both a maximum and minimum value on the interval $[a, b]$. Any extreme value distribution is considered a limiting model, which is a model that determines how large or small the data might be. The information in this section originates from Kotz and Nadarajah (2000).

Given a large set of independent, identically distributed (iid) random values, the GEV distribution may be used to model the smallest or largest values. The GEV distribution consists of three parameters and unites three underlying distributions. The three parameters are location - μ, scale - σ, and shape - ξ, leading to the following Cumulative Distribution Function (CDF) of the GEV distribution:

$$F(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{24}$$

Equation 24 is valid when:

$$\sigma > 0$$
$$1 + \xi\tfrac{x-\mu}{\sigma} > 0$$
$$\xi, \mu \in \mathbb{R}$$

The three underlying models represent different shapes of a respective parent distribution; thus, it is customary to select one of the three sub-models based on the behavior in the tail of the parent distribution. However, if the behavior of the parent distribution is unknown, then choosing the proper sub-model proves difficult, leading to the use of the GEV distribution accounting for all three forms. The three sub-models are the Gumbel Distribution (Type I), the Fréchet Distribution (Type II), and the Weibull Distribution (Type III).

The most commonly known Extreme Value Distribution (EVD), the Gumbel distribution, has two forms, one for the minimum and one for the maximum. It is unbounded and defined on the entire range of real numbers. The parameter of the GEV distribution related to the Gumbel distribution is the location parameter μ.

Increasing μ shifts the distribution to the left, while decreasing μ shifts it to the right. The Gumbel distribution is found when the shape parameter $\xi = 0$ in the GEV distribution. The CDF for the Gumbel distribution follows:

$$F(x;\mu,\sigma) = \exp\Big\{ -\exp\big(-\frac{x-\mu}{\sigma}\big)\Big\} \tag{25}$$

The Fréchet distribution is used to model the maximum values in a data set. Consisting of all three parameters from the GEV distribution, the Fréchet distribution converges slowly to 1 and has the following CDF when $x > \mu$:

$$F(x;\mu,\sigma,\xi) = \exp\Big\{ -\big(\frac{x-\mu}{\sigma}\big)^{-\xi}\Big\} \tag{26}$$

This CDF is defined from the GEV distribution when $\xi > 0$. Many natural phenomena have been modeled with the GEV Type II distribution to include maximum rainfalls and human lifespans.

Last of the three distributions is the Weibull distribution which is primarily used to model failure times. Extremely useful in reliability, the Weibull actually comprises a family of distributions that take on multiple shapes depending on the parameterization. The GEV distribution CDF becomes the Weibull CDF when $\xi < 0$ and, when $X \geq 0$, takes on the form:

$$F(x;\sigma,\xi) = 1 - \exp\Big\{ -\big(\frac{x}{\sigma}\big)^{\xi}\Big\} \tag{27}$$

EVDs are useful in a variety of contexts. The nature of the primary problem for this research lends itself well to potentially incorporating the use of an extreme value distribution for the distribution of lightning strikes that occur outside the preexisting lightning area.

2.7 Goodness-of-Fit (GoF) Tests

Once the data on the distances outside a preexisting lightning area is fit to a distribution, the next step is to test the fit of that distribution to the data. Common to the statistical world, GoF tests seek to determine if a set of data originates from a population with a specific distribution. This is accomplished by failing to reject the null hypothesis for the test, which specifies that the data follows a conjectured distribution, at some level of confidence. Several tests exist for testing GoF of sample data. This section contains an overview of three tests to include Anderson-Darling (A-D), Kolmogorov-Smirnov (K-S), and the Chi-Square. The pertinent information in this section on GoF tests draws from the Engineering Statistics Handbook (2018).

2.7.1 A-D GoF Test

Often the preferred test because of its more sensitive approach to deviations in the tails than the K-S test, the A-D test may be applied to any distribution. The A-D test gains more power as the sample size increases and is consistently more powerful than the K-S test. When calculating the critical values, the A-D test incorporates the specific distribution being tested. Common distributions for which tables of critical values are readily available include the normal, uniform, lognormal, exponential, Weibull, Gumbel, generalized Pareto, and logistic distributions. The definition of the A-D test follows Equations 28 and 29 where F is the CDF of the specified distribution and $\{Y_i < ... < Y_N\}$ are the ordered data.

H_0 : The data follow a specified distribution.

H_1 : The data do not follow the specified distribution.

Test Statistic:

$$A^2 = -N - S \tag{28}$$

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} [\ln F(Y_i) + \ln(1 - F(Y_{N+1-i}))] \tag{29}$$

After choosing a significance level α and calculating the test statistic, the critical value, which is dependent on the specific distribution being tested, may be found in one of the aforementioned tables for the listed distributions. The A-D test is one-sided and the null hypothesis is rejected if the test statistic A^2 is greater than the critical value.

2.7.2 K-S GoF Test

One of the most well known GoF tests, the K-S test is based on the empirical distribution function (EDF). With N *ordered* data points $Y_1, Y_2, ..., Y_N$ the EDF is defined:

$$E_N = \frac{n(i)}{N} \tag{30}$$

where *n(i)* is the number of points less than Y_i and the Y_i are ordered by value from smallest to largest. Unlike the A-D test, the K-S test statistic does not depend on an underlying distribution; however, several limitations exist for the K-S test. Besides only being applicable for continuous distributions, the K-S test holds more sensitivity near the center of the distribution than at the tails as indicated previously when describing the A-D test. A more difficult limitation though is that the distribution in question must be fully specified for the K-S test, that is the parameters of the distribution may not be estimated from the data because that would invalidate the critical value. This means simulations are often necessary to determine the parameters of the distribution.

The K-S test is defined with the same null and alternative hypotheses as the A-D test. The test statistic for the K-S test follows where F is the fully specified

theoretical CDF of the distribution being tested:

$$D = \max_{1 \leq i \leq N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right) \tag{31}$$

After choosing a significance level α and obtaining the critical value from an associated table, rejection of the null hypothesis occurs when the test statistic D is greater than the critical value. Although the K-S test may often provide adequate results, the A-D test, considered a refinement of the K-S test, is often preferred as it is considered to be more powerful than the K-S test.

2.7.3 The Chi-Square GoF Test

Differing from the two previously described tests, the Chi-Square GoF test may be applied to any univariate distribution for which a CDF can be calculated but must be applied to binned data. The binning requirement is not an incredible restriction as a histogram or frequency table may be generated for non-binned data prior to accomplishing the chi-square test. However, it is important to note that the test statistic value is dependent on the nature of the binning of the data. Another potential limitation lies in the fact that the chi-square test requires a sufficient sample size for the test to be valid. One advantage the chi-square test has over the two previous tests is that it can be applied to discrete distributions.

The null and alternative hypotheses for the chi-square test are the same as for the previous two tests. After the data are divided into k bins the chi-square test statistic follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{32}$$

Here O_i is the observed frequency for bin i and E_i is the expected frequency for bin

i where the expected frequency is calculated:

$$E_i = N(F(Y_u) - F(Y_l)) \tag{33}$$

Equation 33 is based on sample size N where F is the CDF for the distribution being tested and Y_u and Y_l are respectively the upper and lower limits for class i. The test is sensitive to the choice of bins and there is no optimal choice for the bin width; however, most reasonable choices should produce similar results. The expected frequency should be at least five and combining some bins in the tails might be necessary if the counts are less than five.

If $c = p + 1$ is the number of estimated parameters for the distribution plus one, then the test statistic approximately follows a chi-square distribution with $(k - c)$ degrees of freedom, k being the number of non-empty cells. Thus, the null hypothesis will be rejected at the level of significance α if:

$$\chi^2 > \chi^2_{1-\alpha,k-c} \tag{34}$$

where $\chi^2_{1-\alpha,k-c}$ is the chi-square critical value.

Several factors must be taken into account when determining what GoF test should be applied to a set of data. First, considering whether or not the data is continuous or discrete is paramount. The sample size plays an important role as well since some tests require an adequate number of data points to maintain validity. Also, depending on the distribution to which the data is being compared there could be potential concerns about the availability of tables. An additional option to choosing one test over another would be to apply two or more to the data; however, this might result in confounding results and lead to the lack of an overall conclusion on the distributional form of the data.

2.8 Summary

We now have more detailed knowledge of the lightning strike process which helps inform the application of the LDAR data being used in the creation of the ellipse fitting algorithm. An overview of the current methods of prediction reveals to us the number of techniques and plethora of technology being utilized to help safeguard life and property from lightning. A thorough review of past research suggests the importance of the new methodologies being employed in this study. Finally, a summary of the ellipse fitting and statistical methods considered for implementation prepares us for discussing the methodologies actually used in this study. Therefore, we continue to the next chapter where we discuss the detailed methodologies we apply to the data to identify a fitting distribution for lightning strikes beyond a preexisting lightning area.

III. Methodology

3.1 Overview

This chapter details the data used for analysis as well as how that data is processed in preparation for the study. We then discuss the process used to fit ellipses to the data to include how previously outlined techniques from Chapter II are incorporated and any assumptions that are made about the data for purposes of the fitting routine. Lastly, we discuss the analysis process, including an outline of a validation method that tests the chosen safety distance using the actual lightning warning circles of the 45 WS.

3.2 Data

The source of data for this study originates from the LDAR network. As the LDAR network registers many in-cloud events and step leaders as lightning sources or source points, a single lightning flash may contain anywhere from several to thousands of source points. The Applied Meteorology Unit (AMU) at the 45 WS provided the flashified LDAR data for this study. For the years of 2013 through 2016, LDAR data from CCAFS is flashified by grouping source points; if any two source points are within 0.3 seconds and 3,000 meters, they are grouped together as part of the same flash.

For every month, January through December of 2013 through 2016, a text file is generated with this flashified LDAR data where every vertical line break indicates a new flash. Any source of lightning detected by the LDAR system registers a date/time stamp along with a 3D recording of location and an epoch time. Table 1 illustrates the data collected for each source point; in the provided text files, the data are in a comma separated format with the following fields: Date/time, X, Y, Z, and epoch

time. The X, Y, and Z coordinates are in reference to KSC's LDAR-I Central Site. Appendix E provides an example of these monthly text files. From this example in Appendix E we see that a lightning event can range anywhere from one to hundreds or thousands of source points as indicated earlier. An example of the largest flash from June 2013 containing 1,321 source points of can be seen in Figure 9. The MVEE for this flash covers a geographical area of 328.17 NM^2.

Table 1: Format of Data Collected via LDAR system

Date/Time:	MM/DD/YYYY HH:MM:SS:ms (hours are 00-23, milliseconds to 6 digits)
X:	east/west distance from center in meters
Y:	north/south distance from center in meters
Z:	altitude in meters
epoch time:	number of seconds since 01/01/1970 00:00:00

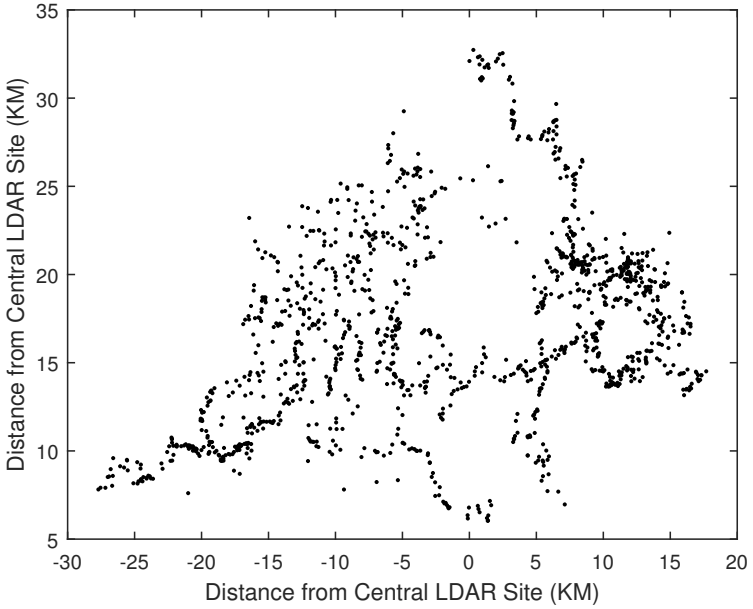


Figure 9: June 2013 Lightning Flash

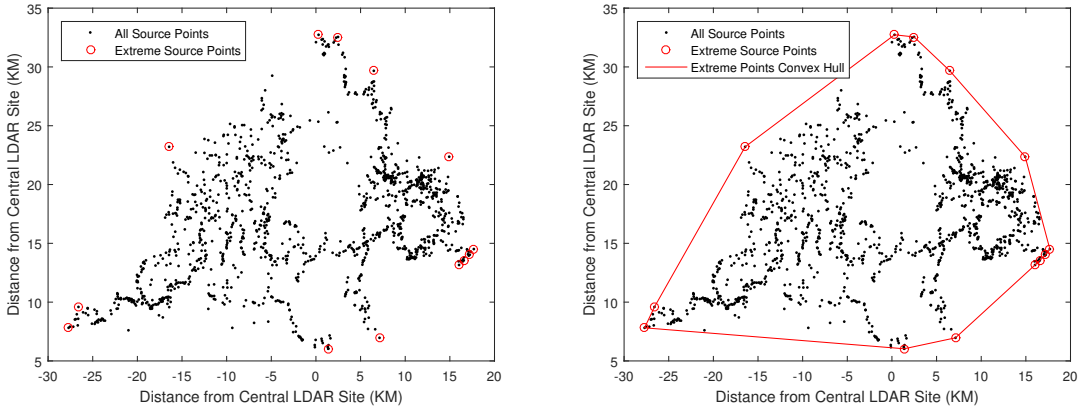
For the purposes of this research we determined after consultation with subject matter experts that we would remove from the data any events with less than five source points. The principal reasoning for this is that those readings with few to no other nearby source points detected by the LDAR system are generally not considered to be full lightning flashes. Also, because the peak lightning season for the central eastern coast of Florida occurs in the summer months, we narrowed the focus of the analysis to the months of May through September for 2013-2016, giving us 20 months of data for the complete analysis process. Lastly, because the accuracy and detection rate for the LDAR network decrease as the distance from the sensors increases, subject matter experts also suggested a reduction in data with a focus on the flashes that occur within a 25 NM (46.3 KM) radius of the central LDAR site. Although this reduction results in a smaller data set, the 25 NM radius still covers all 10 lightning warning circles supported by the 45 WS.

3.2.1 Data Processing

All data processing and analysis discussed throughout the following sections relied on MATLAB R2015a. The relevant code for the following outline of how the flashified LDAR data is processed can be found in Appendix F. The subsequent process is applied to all 20 months of data. After the data is read in, the X and Y coordinates are converted from meters to kilometers (KM). The Z coordinates are removed as we are not considering the height of a lightning flash in this particular study. The number of source points associated with each flash is counted. If a flash does not have five or more source points, it is removed from the data. Two separate matrices are then created to record the x and y coordinates for every source point of each flash. For example, if the first flash has 20 source points, column one of the matrices **flash_x** and **flash_y** would have 20 non-zero entries reflecting the x and y coordinates of the

source points, respectively. The dimension of each of these matrices is the maximum number of source points for the largest flash by the total number of flashes.

The next step is to assign a date/time stamp for each flash. Because any two source points for a single flash occur within 0.3 seconds of each other, we determine that using the time stamp associated with the first source point of the flash is sufficient. Therefore, the time stamp from the first source point of each flash is extracted and associated with each flash. Given the nature of the analysis being performed on the data set, we foresaw the relevance of determining the extreme source points of each flash. That is, using a convex hull MATLAB function we found the outer most source points of each flash. An example of this on the largest flash in June 2013 can be seen in Figure 26 where the resulting number of extreme points for this flash is 13. The code for the convex hull function can be found in Appendix G. The x and y coordinates for each of these extreme points are recorded in a similar fashion as outlined previously for all the source points. The matrices used to record the extreme points of each flash are **flash_x_extremes** and **flash_y_extremes** where the dimension of each of these matrices is the maximum number of extreme points by the total number of flashes.



(a) Extreme Points **(b) Convex Hull**

Figure 10: June 2013 Lightning Flash Extreme Points and Convex Hull

To store all the relevant information for each flash, a single matrix **date0** is created. The **date0** matrix records the following for each flash: the original flash number, the number of source points, the date/time stamp (in six columns: YYYY, MM, DD, HH, mm, SS.mssssss), and the number of extreme points. The total number of columns in this matrix is 10, where the tenth column is reserved for recording the associated ellipse number for each particular flash.

The subsequent step of the data reduction process is to determine which flashes occur within the specified 25 NM (46.3 KM) radius of the central site. For each flash, the distance each extreme point lies from the central site is calculated. If all the extreme points occur within 25 NM of the central LDAR site, the flash is retained. If even one extreme point occurs outside the 25 NM radius, the entire flash is removed from the data set. Figure 11 gives a visual of this data reduction, where the black source points represent flashes with five or more source points.

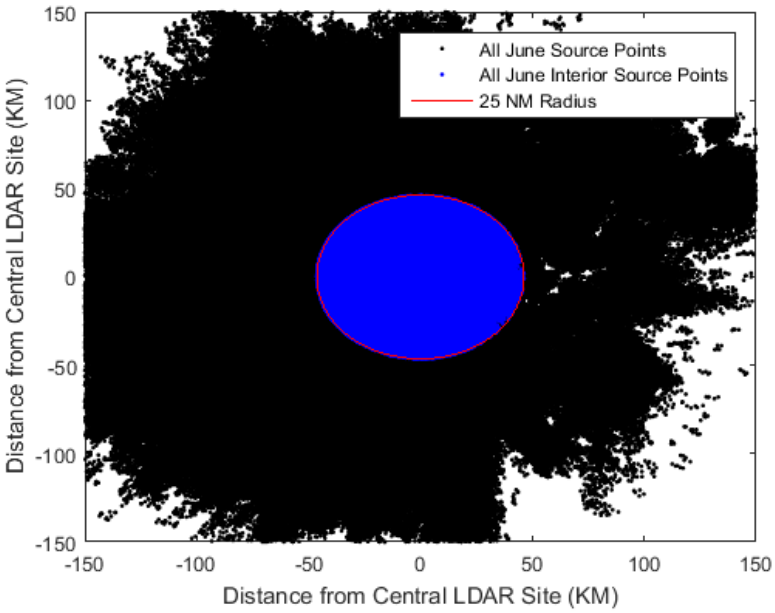


Figure 11: June 2013 Data Reduction

All of the previously specified matrices are then updated to only include flashes that occurred within the 25 NM radius of the central site. A final matrix designated **flash_mat** is also created to maintain a comprehensive listing of all the x and y coordinates of the included flashes. The **flash_mat** matrix dimensions are the total number of source points for all flashes by six. The six columns are: x coordinates for all source points, y coordinates for all source points, x coordinates for all extreme points, y coordinates for all extreme points, the final flash number associated with each source point, and the final flash number associated with each extreme point. These final flash numbers are simply a renumbering of the flashes so that the first flash with five or more source points and within the 25 NM radius is now flash number one and so on; however, the original flash number for each source point is still maintained in column one of the **date0** matrix.

As observed in Figure 11 a considerable amount of data reduction occurs to ready the data for the ellipse fitting routine. Table 2 gives the specifics of the data reduction for the number of flashes. A total of 808,430 flashes, 4% of all the flashes that occurred from May to September of 2013 to 2016, have been extracted to use in the actual ellipse fitting process. July 2014 has the greatest number of flashes in every category (original number of flashes, flashes with five or more source points, and flashes within 25 NM of the central site). June 2013, May 2015, and June 2016 have the greatest percentage of flashes with five or more source points, with each month having 33% of its source points retained through this particular data reduction step. And May 2013 has the greatest percentage of flashes within the 25 NM radius of the central site.

In Table 3 we see the number of source points in each month. Not surprisingly, the month with the greatest number of source points in every category is again July 2014. The largest flash, that is the one with the greatest number of source points, occurs in August 2015 where one flash has 2,560 source points. Lastly, the month

with the largest number of flashes on average is May 2014 with a mean number of source points per flash of 151.

Table 2: Number of Flashes by Month

Month	Original # Flashes	Flashes >=5 Source Points	Flashes <= 25 NM	% Flashes >= 5 Source Points	% Flashes <= 25 NM
May-13	305,142	97,887	30,311	32%	10%
Jun-13	1,017,798	340,642	64,789	33%	6%
Jul-13	452,182	139,423	23,818	31%	5%
Aug-13	755,052	200,016	24,408	26%	3%
Sep-13	649,287	195,280	22,527	30%	3%
May-14	486,595	142,587	11,825	29%	2%
Jun-14	1,783,266	527,485	58,277	30%	3%
Jul-14	1,849,330	580,752	126,384	31%	7%
Aug-14	1,593,201	483,207	73,902	30%	5%
Sep-14	817,718	180,377	22,761	22%	3%
May-15	299,013	99,105	11,134	33%	4%
Jun-15	1,441,951	459,392	43,668	32%	3%
Jul-15	1,145,037	350,497	51,321	31%	4%
Aug-15	1,597,543	428,337	55,781	27%	3%
Sep-15	673,189	171,177	26,663	25%	4%
May-16	848,919	228,413	21,755	27%	3%
Jun-16	760,832	249,381	33,288	33%	4%
Jul-16	847,517	267,508	31,065	32%	4%
Aug-16	368,214	97,915	8,029	27%	2%
Sep-16	1,015,235	313,125	66,724	31%	7%
Overall	18,707,021	5,552,506	808,430	30%	4%

Table 4 gives the last data summary with the monthly number of extreme points. Again we see July 2014 has the greatest number of extreme source points in every

category. One important note about these data summaries is that they are based on the data after slight modifications are made to ready the data for the ellipse fitting algorithm. Specifically, some flashes are removed from certain months and added to the end of the previous month; the reasoning for this change is explained in further detail later.

Table 3: Number of Source Points by Month

Month	Original # Source Points	Total Source Points flashes >= 5	Total Source Points flashes <= 25 NM	Max # Source Points (Final Data)	Mean # Source Points (Final Data)
May-13	3,138,848	2,786,095	1,084,971	1,601	36
Jun-13	11,231,273	2,828,053	3,493,132	1,321	54
Jul-13	4,152,709	3,614,903	1,028,104	1,067	43
Aug-13	6,528,645	5,596,959	1,638,228	1,031	67
Sep-13	5,980,882	5,198,347	1,206,091	1,060	54
May-14	7,741,792	7,187,196	1,786,503	1,627	151
Jun-14	19,678,385	17,648,540	3,834,333	1,927	66
Jul-14	22,304,479	20,247,683	7,295,701	1,908	58
Aug-14	19,775,169	17,993,453	5,316,811	1,663	72
Sep-14	9,975,539	9,040,597	2,570,368	1,646	113
May-15	3,804,174	3,469,991	859,383	1,460	77
Jun-15	16,610,301	14,978,016	3,258,952	1,384	75
Jul-15	14,596,621	13,298,406	4,521,050	2,056	88
Aug-15	18,859,328	17,015,787	5,158,066	2,560	92
Sep-15	5,845,101	5,048,310	1,376,242	1,453	52
May-16	7,520,125	6,499,344	1,396,893	1,466	64
Jun-16	7,687,396	6,848,603	1,541,787	1,725	46
Jul-16	6,708,425	5,725,809	1,079,353	1,050	35
Aug-16	3,112,428	2,684,850	600,879	1,219	75
Sep-16	11,879,607	10,752,551	3,861,580	2,245	58
Overall	207,131,227	178,463,493	52,908,427	2,560	65

Table 4: Number of Extreme Points by Month

Month	Total Extreme Points Flashes >=5	Total Extreme Points Flashes <= 25 NM	Max # Extreme Points (Final Data)	Mean # Extreme Points (Final Data)
May-13	622,591	200,911	19	6.63
Jun-13	2,159,814	467,706	20	7.22
Jul-13	853,684	161,167	20	6.77
Aug-13	1,212,872	178,995	23	7.33
Sep-13	1,198,246	161,045	20	7.15
May-14	930,780	95,935	21	8.11
Jun-14	3,299,047	417,604	21	7.17
Jul-14	3,663,170	892,645	21	7.06
Aug-14	3,060,281	533,500	21	7.22
Sep-14	1,176,175	176,632	21	7.76
May-15	631,795	83,128	19	7.47
Jun-15	2,877,164	318,416	24	7.29
Jul-15	2,226,579	387,689	25	7.55
Aug-15	2,699,805	419,299	23	7.52
Sep-15	1,041,714	182,911	19	6.86
May-16	1,388,369	160,739	21	7.39
Jun-16	1,541,280	236,656	20	7.11
Jul-16	1,596,892	208,379	18	6.71
Aug-16	601,933	61,745	19	7.69
Sep-16	1,990,370	483,000	25	7.24
Overall	34,772,561	5,828,102	25	7.21

3.3 Selection of the Ellipse Fitting Method

The choice of which ellipse fitting method to use is one of the most critical decisions in the creation of the final algorithm to find the distance lightning strikes beyond the edge of a preexisting lightning area. The code for all of these ellipse fitting algorithms can be found in Appendix H. We do not heavily consider the least squares best fitting ellipse method for this particular application. This is primarily due to the fact that this method seeks to fit the boundary of an ellipse to a given set of data points; that is, the least squares best fit ellipse seeks to minimize the distance of each data

point from the edge of the ellipse. However, for our ellipse formations, we would prefer the ellipse to be drawn around a certain percentage, if not all, of the data points (preferably centered around the data points) as opposed to drawn through the data points. An ideal algorithm for such a task would be a least squares approach that seeks to minimize the distance of each data point from the center of an ellipse; however, to our knowledge, there does not currently exist an algorithm that performs this specific task.

We heavily investigate and apply the method of PCA confidence ellipses to this problem. The reasoning behind using the PCA approach is that it provides a good deal of elliptical coverage over a large percentage of flashes while not producing overtly large areas. For all investigative purposes, the data for the months of May and June 2013 are used. The first necessary step in applying the PCA confidence ellipses is to determine what α value to use for the ellipse formations. To find the desired α value, we use the data from May 2013 and apply an early version of the ellipse fitting algorithm, the details of which we discuss in a later section. At various α values (.20, .15, .10, .05, .01), the algorithm calculates the area of each ellipse that is drawn, as well as the percentage of points that are inside the ellipse. As a way to determine the optimum α level, we then proceed to find the following value for each ellipse drawn:

$$f(\alpha, \%) = \frac{\text{area of ellipse}}{\%\text{ of flashes inside ellipse}} \tag{35}$$

Taking the mean of these values for each α level and selecting the minimum results in $\alpha = .20$ being the optimum alpha value to use. Table 5 gives a breakdown of this analysis.

Table 5: Optimal Alpha Value

α **level**	**Mean Area (KM^2)**	**Mean % Flashes inside**	$f(\alpha, \%)$
.01	2,357	90.6%	2,604
.05	1,114	79.4%	1,419
.10	716	70.2%	1,037
.15	523	62.8%	852
.20	384	56.8%	697

After choosing $\alpha = .20$, we begin to examine plots of the performance of the PCA ellipse fitting method. The first problem we discover using the PCA approach is that when all of the source points are used to create the ellipses, large clusters of source points overpower the smaller outlying source points, as is the nature of PCA. An example of this can be seen in Figure 12 where an ellipse is drawn for a series of lightning flashes using all the source points from each flash.

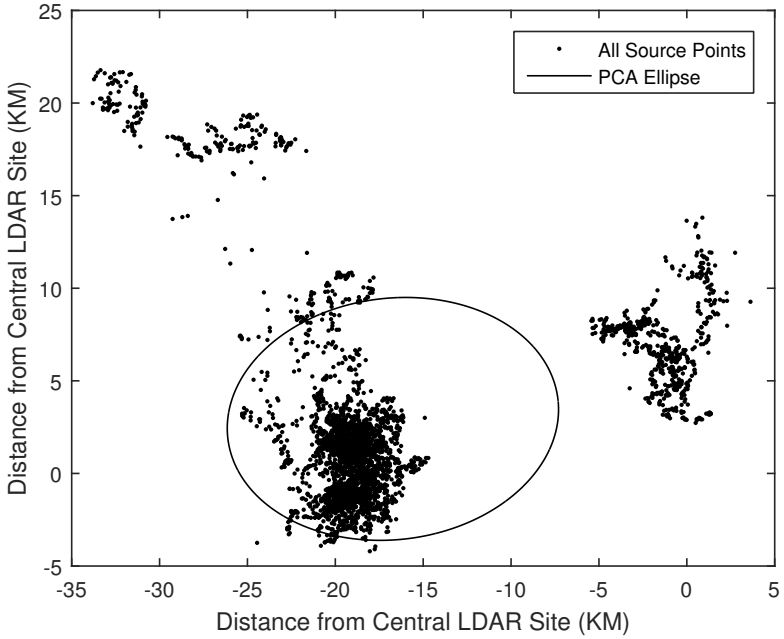


Figure 12: PCA Ellipse using All Source Points

From Figure 12 we see that using all of the source points to draw the PCA ellipse results in a large portion of the source points falling outside of the ellipse. This does not achieve the initial goal of fitting an ellipse to the data. The next attempt we make is to apply PCA ellipses using the extreme points of every flash. However, the same general problem of larger clusters of extreme points overpowering the smaller groups of extreme points persists. Figure 13 shows an updated version of Figure 12 with the extreme points of each flash specifically identified and a new ellipse drawn using just those extreme points.

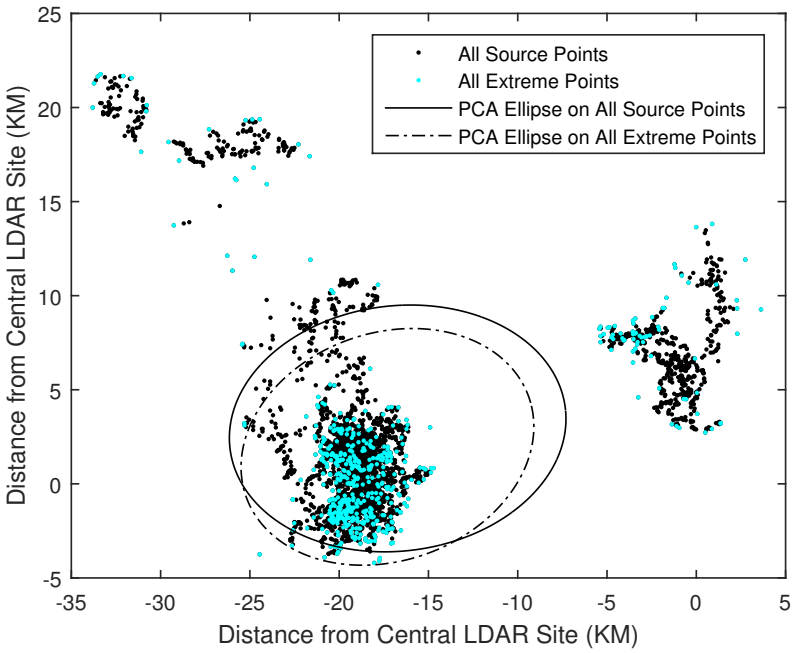


Figure 13: PCA Ellipse using All Extreme Points

The final attempt we make to salvage the PCA approach relies on taking the convex hull of all the extreme points and using those most extreme points to create the PCA ellipse. This method solves a large portion of the problem; however, too many source points still lay outside of the ellipses, creating an inflation in the number and length of the distances that are being recorded for lightning strikes beyond the edge of a preexisting lightning area. Figure 14 demonstrates this last PCA ellipse

approach using the most extreme points. Therefore, as the least squares method and PCA approach do not generate the desired results, the final method chosen is the MVEE technique.

The main concern in using the MVEE method from the onset is the size of the ellipses, or the amount of area that is contained within these ellipses. However, certain choices in the ellipse fitting algorithm help to curtail the formation of extremely large ellipses. In Table 6, we see a comparison between the ellipses created using the PCA technique and those created using the MVEE method for the May 2013 trial data. The PCA ellipses in Table 6 are those created using the last PCA approach where the most extreme points are applied in the creation of the ellipses. Although the MVEE are larger on average than the PCA ellipses, the mean lengths of the axes for both methods are similar and not implausible in regards to the size of actual lightning storms.

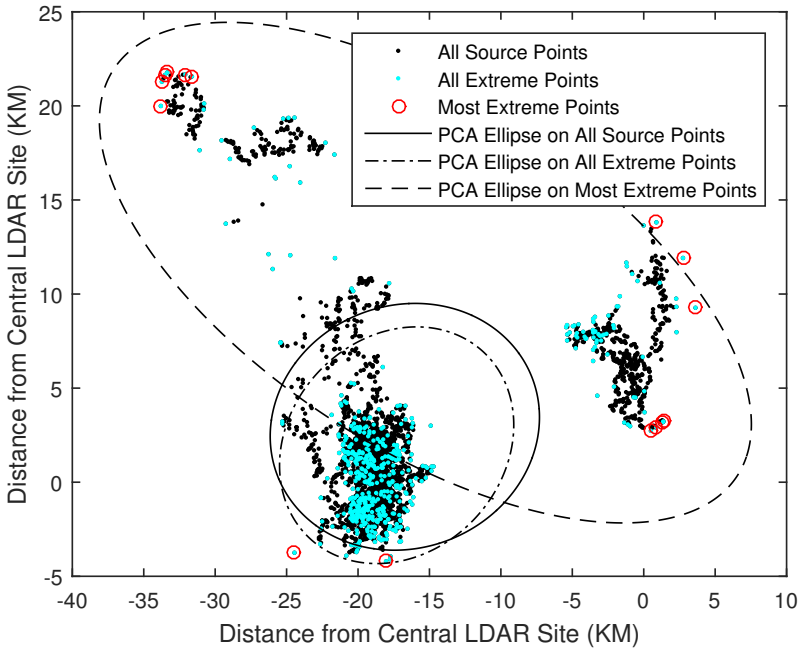


Figure 14: PCA Ellipse using Most Extreme Points

Table 6: Comparison between PCA Ellipses and MVEE

Ellipse Fitting Technique	Mean Area (NM^2)	Max Area (NM^2)	Mean Major Axis (NM)	Mean Minor Axis (NM)
PCA Ellipses	636.5	1,281.9	9.6	5.7
MVEE	815.9	1,745.2	10.5	6.7

The last decision to be made in regards to the use of the MVEE method is to determine the optimum tolerance value for the algorithm. The choice is based on finding a balance between computational speed and accuracy. Table 7 and Figure 15 provide the numerical and visual support for the final choice of a .01 tolerance level, respectively. Considering Figure 15 and Table 7 simultaneously we see that there is not much accuracy lost when moving from a tolerance of .0001 to .01; but, there is a 97% increase in computational speed between the two tolerance levels. At the same time, moving from .01 to .05 we begin to see a deterioration in the accuracy of the algorithm with no speed increase. Thus, the final decision is to use a tolerance of .01 for optimum computational speed and accuracy. We now turn to the ellipse fitting algorithm and the assumptions that are made to create it.

Table 7: Computational Speed of MVEE Algorithm

Tolerance Level	MVEE Algorithm Speed (seconds)
.0001	.336
.001	.035
.01	.010
.05	.010
.10	.009

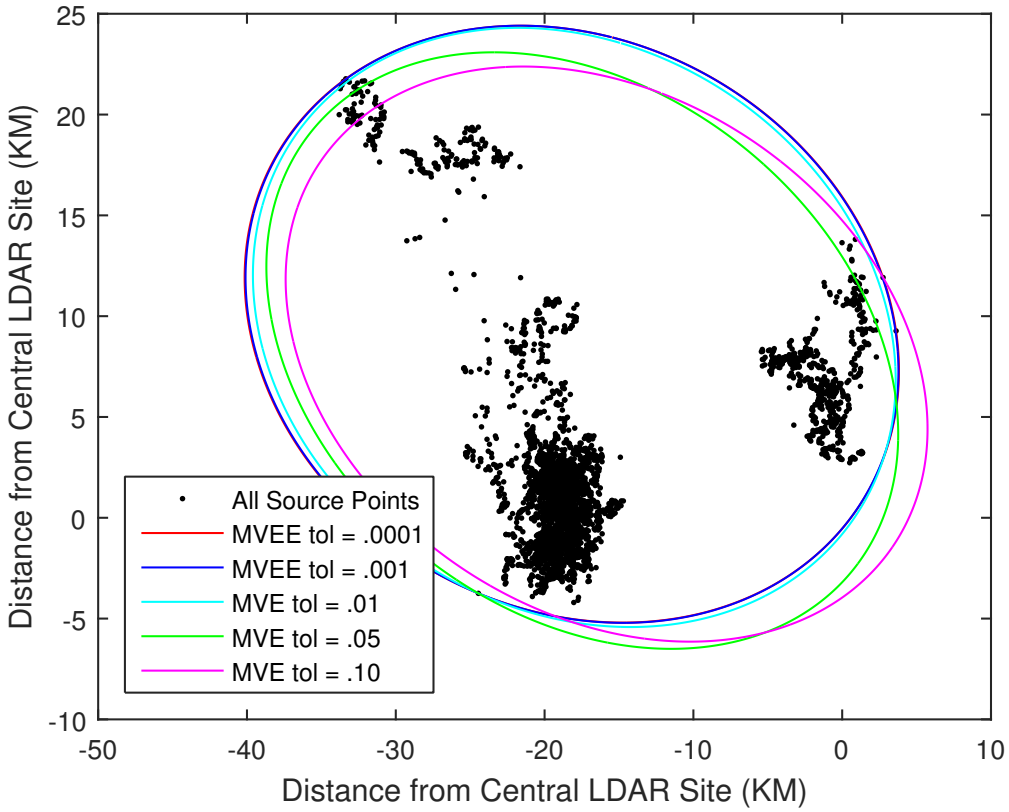


Figure 15: MVEE Tolerance Levels

3.4 Ellipse Fitting Algorithm and Assumptions

When discussing the function of the ellipse fitting algorithm, we consider the overarching goal of this research of finding a distributional representation of lightning strikes beyond a preexisting lightning area. Therefore, the purpose of the ellipse fitting algorithm is to draw elliptical boundaries around active lightning areas and measure the distance a lightning strike occurs outside of these ellipses. As we describe the process of the algorithm, we outline the specific assumptions that are made in its creation. The code for this algorithm can be found in Appendix I. The first assumption lies in the concept of a lightning strike. Recall from Chapter II that the LDAR network does not produce CG strike data, rather it gives X and Y coordinates for all

the source points associated with a lightning strike. Therefore, the first assumption made is that each source point of a flash is considered as a potential CG strike at each source points X and Y coordinate location.

The algorithm functions in accordance with the following series of steps and begins with the first flash from a list of all the flashes for the entire month. When a new flash occurs, if it is the first flash of the month or an existing ellipse is not already present, then an ellipse is drawn around it; if it is not the first flash of the month and an existing ellipse is already present, meaning an ellipse has already been drawn around at least one previous flash, then the time difference between the current flash and the last flash of the current ellipse is found. If the time difference is more than 30 minutes, the current flash represents the start of a new ellipse. This assumption originates from the second part of the 30-30 rule discussed in Chapter II and suggests that if lightning has not occurred within 30 minutes of the previous flash, then it is not a part of the previous storm but rather a new one.

However, if the time difference is less than 30 minutes, the current flash might potentially belong to the current storm for which the current ellipse was drawn. Therefore, the next step is to determine if the lightning flash occurred within or outside the current ellipse. If the lightning occurred within the ellipse, then the distance from the center of the ellipse to the furthest source point of the lightning flash is recorded and the algorithm moves on to the next flash. In contrast, if even one source point of the flash occurred outside the ellipse, then the distance from the edge of the ellipse to each source point striking outside the ellipse is determined and the largest distance is taken as the flash's distance from the edge of the ellipse.

Next, the algorithm uses that distance to establish if the current flash is within 16 KM of the current ellipse. If the current flash is not within 16 KM of the current ellipse, it is not included in the current ellipse, but annotated as still needing assign-

ment to an ellipse, and the algorithm will then move on to the next flash in the list. If the current flash is within 16 KM of the current ellipse then the flash will be included in the current ellipse and the next step of the process is initiated. The reasoning behind using a 16 KM distance threshold is to decipher between different storms that might be active at the same point in time. Several distances were considered for this particular problem; however, using a shorter length arbitrarily curtails lightning that strikes at greater distances beyond a preexisting lightning area, and larger distances create ellipses that are immense and encompassing far too much area. Previous studies outlined in Chapter II also referenced similar distance thresholds; for example, Parsons (2000) used a 17 KM distance threshold for her clustering criteria. Subject matter experts were presented these details and agreed that 16 KM is the best choice for the algorithm.

The subsequent step of the algorithm is to go through all of the flashes in the current ellipse and remove those that are older than 10 minutes from the time of the current flash. This step accounts for the movement of a storm over time and also helps to ensure that the ellipses being drawn are not overly large. Subject matter experts were also consulted on this point and did not contend with its implementation as it was offered that removal at both 5 and 15 minutes were also attempted and the differences in the distributional results between all three time removal thresholds were minimal. Therefore, removal of flashes at 10 minutes is selected as the moderate choice.

At this point in time, a new ellipse is drawn using the MVEE algorithm with a .01 tolerance level. A matrix entitled **ellipses** records the summary information of each ellipse that is drawn. Specifically, the **ellipses** matrix includes for each flash: the ellipse number, the area of the ellipse, the distance the flash occurred from the edge of the ellipse, the distance the flash occurred from the center of the ellipse, the

number of flashes in the ellipse, and the parameters of the ellipse (a, b, h, k, τ).

The algorithm continues this process until one of three events happen. The first was already mentioned, that the time difference between the last flash of the current ellipse and the current flash being considered is greater than 30 minutes. The second event that interrupts this process deals with the number of flashes in a row that occur further than 16 KM from the edge of the current ellipse. Every time a flash occurs beyond 16 KM from the edge of the current ellipse, a counter is incremented; this counter is reset to zero if a subsequent flash occurs within 16 KM of the current ellipse. However, if the counter reaches five, meaning five flashes in a row have occurred at a distance of 16 KM or more beyond the edge of the current ellipse, the algorithm process is again interrupted. The reasoning for this is twofold. First, if there are no more relevant flashes for the current ellipse, curtailing the process at this point saves valuable computation time as the algorithm does not have to continue cycling through a larger amount of flashes without drawing ellipses. Secondly, this allows that there might be other storms in the same area, currently further than 16 KM away, that may be moving towards the current ellipse; thus, it would be better to interrupt the process before a flash from another storm reaches that 16 KM threshold and is included, incorrectly, in the current ellipse.

The third event that interrupts the algorithm's status queue is reaching the end of the existing list of unused flashes (flashes that have not been assigned or used in the creation of an ellipse). As the algorithm runs, besides storing information in the **ellipses** matrix, the ellipse number is also recorded in the last column of the **date0** matrix for each flash. If the algorithm's process is interrupted for one of the three events previously described, a new list of unused flashes is generated by looking at the last column of the **date0** matrix. This process continues using this new list of flashes until every flash of the month has been included in an ellipse.

We made note earlier that some flash data is removed from certain months and placed at the end of the previous month. This happens because, for multiple pairs of months, the same storm spanned both months. For example, if there is a lightning storm late on the night of 30 June that continues into the early morning hours of 1 July, the flashes that occur in July inherently belong with the flashes that occur on the last day of June. Therefore, if there is any overlap between the last day of one month and the first day of the next month, we find the first time break between flashes of 30 minutes or more on the first day of the month and move all the flashes prior to that time break to the end of the previous month.

3.5 Validation of the New Warning Distance

Once the process for fitting ellipses is complete, we fit the distances obtained from the algorithm, which represent the distance lightning strikes beyond the edge of a preexisting lightning area, to a distribution. After finding the best distributional representation of these distances, we determine the appropriate distance for initiating a lightning warning given the level of increased risk we are willing to accept. After this new stand-off distance is ascertained, we empirically decipher how often this new stand-off distance fails as opposed to the previous 5 NM distance.

To complete this empirical process, we use the actual warning circles from the 45 WS. Given the latitude and longitude of each warning circle center, we calculate the geographical distance of each warning circle center from the central LDAR site. As all of the LDAR data are X and Y coordinates from the central LDAR site, we also need the warning circle centers as X and Y coordinates from the central LDAR site. These conversions are made using the National Deep Submergence Facility (NDSF) Coordinate Conversion Utility tool (NDSF, 2018).

After obtaining the X and Y coordinates, we proceed to calculate the euclidean distance of each warning circle center to the central LDAR site. We then compare the euclidean and geographical distances and determine the difference between the two to be marginal. Table 8 provides all of this information as well as references the radius of each warning circle (recalling that some are 5 NM while others are 6 NM). Figure 16 depicts the location of each warning circle in relation to the central LDAR site. As all the circles fall completely within the 25 NM radius of the central LDAR site, we determine to complete the empirical validation process for all the warning circles.

Table 8: 45 WS Warning Circles Distance from Central LDAR Site

Warning Circle	Radius(NM)	Geographical Distance From LDAR Central Site (KM)	Euclidean Distance From LDAR Central Site (KM)	Difference Between Geographical and Euclidean Distance (KM)
HAULOVER	5	24.57	24.52	0.05
ASTROTECH	5	17.13	17.16	0.03
LC39	6	7.38	7.36	0.02
SLF	6	9.90	9.88	0.02
IA	6	2.18	2.17	0.01
40/41	5	7.21	7.22	0.00
37/ITL	5	6.67	6.68	0.01
CENTRAL	6	10.81	10.8	0.01
PORT	5	14.52	14.48	0.04
PAFB	5	34.00	33.89	0.11

To determine how well the newly selected warning radius performs as compared to the previous 5 or 6 NM radii, we execute the following process for each warning circle for all 20 months of data. The code for this process can be found in Appendix J. The first objective of this process is to establish how much time is saved if the warning circle radius is reduced from 5 or 6 NM to the newly identified radius length. A second objective of this process is to detect how many false alarms are saved using the new radius. We define a false alarm as a warning that would occur at the current 5 or 6

NM radii distance but would not be issued using the new radius distance because no flash occurs within the new radius for the duration of the storm. The final objective is to ascertain how many warning failures occur at the new warning radius distance. That is, how many times does a lightning strike occur within .5 NM of the center of the warning circle before a warning would have been called at the new warning radius. The distance of .5 NM is used as a failure radius as the assets actually being protected fall within .5 NM of the center of each warning circle.

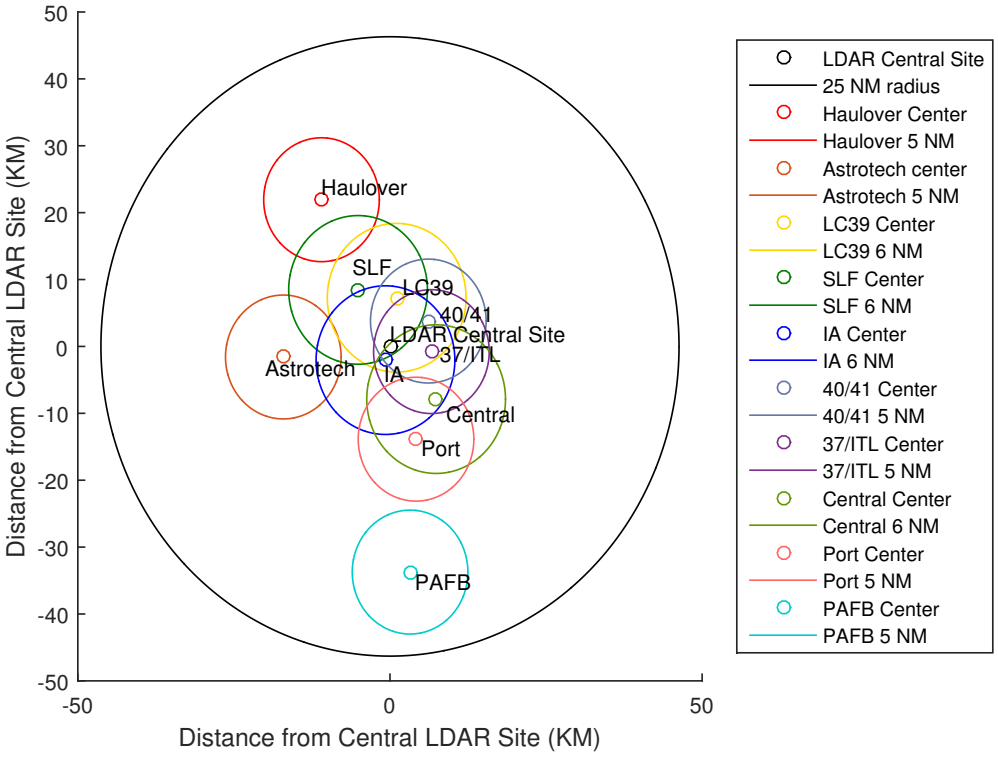


Figure 16: Lightning Warning Circles in Relation to Central LDAR Site

The process begins by initially finding for each flash, the closest source point to the center of the warning circle; this distance is then designated as the only distance associated with each flash. Next, only the flashes that fall within 5 or 6 NM (respective of the warning circle's radius) are retained for that particular warning circle. The flashes are then separated into storms where any time there is at least a 30 minute

break between flashes designates a new storm. Then each flash is assigned a number in accordance with Table 9, a visual of which can be seen in Figure 17.

Table 9: Number Assignment for Flash Distance from Center of Warning Circle

Distance Flash X occurs from Warning Circle Center	**Number Assignment**
new radius $< X \leq 5$ or 6 NM	1
.5 NM $< X \leq$ new radius	2
$X \leq .5$ NM	3

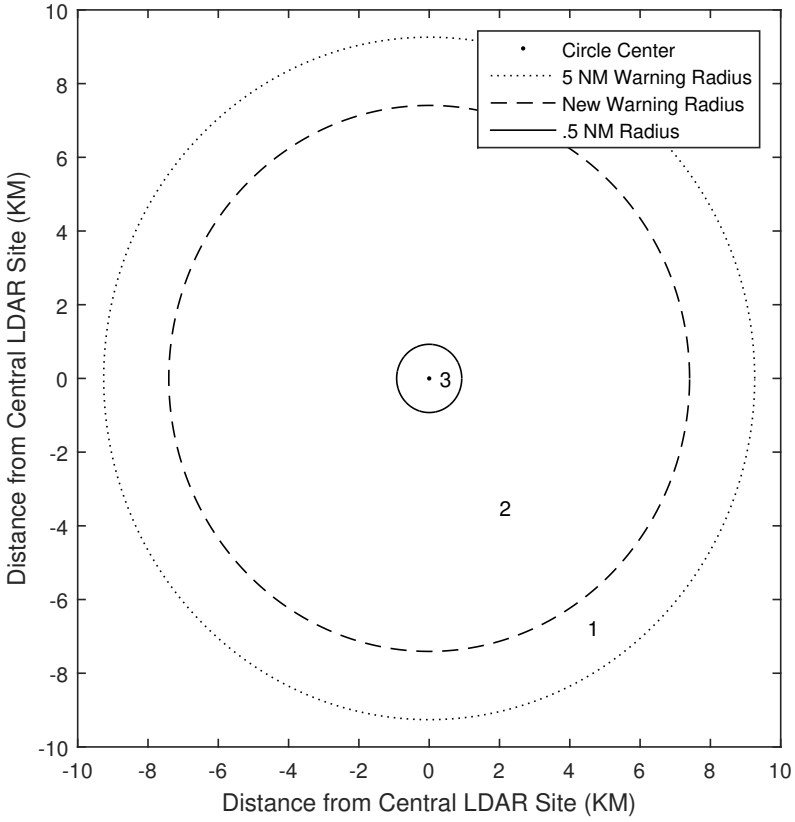


Figure 17: Number Assignment for Flash Distance from Center of Warning Circle

Once each flash is assigned a number based on its distance from the center of the warning circle, the following information is collected for each storm: the first flash assigned a two, the time between that first flash assigned a two and the first flash of the storm (as the first flash of the storm indicates when the original 5 or 6 NM warning is called), the time of the first flash assigned a three, a logical indicator determining if a three was recorded before a two (indicating a failure in the new warning radius), and the total warning time for the original 5 or 6 NM radius. From this information we gather the necessary statistics on how much time is saved, how many false alarms are averted, and how many failures occur both by month and by warning circle.

3.6 Summary

We have now discussed the particular methods being employed to preprocess the data, create the ellipse fitting algorithm, and validate the new warning distance. The next chapter provides the results and analysis of these methods. Specifically, in Chapter IV we see how the choices made in the ellipse fitting algorithm assumptions affect the formation of ellipses. We also use results from these methods to address the secondary research question about the mean distance the initial strikes of a storm occur from the center of the storm. Finally, applying the empirical validation routine provides an answer to the primary research question as to whether or not a new and shorter warning distance may be safely selected and implemented.

IV. Results and Analysis

4.1 Overview

In this section we present descriptive statistics regarding the ellipses formed using the ellipse fitting algorithm. We also answer our secondary research question regarding the mean distance of strike for the initial flashes in a storm before continuing on to the primary objective. We then fit the distances ascertained from the ellipse fitting algorithm to a distribution and determine if a new warning distance can be implemented with the appropriate level of safety. Finally, we provide the empirical validation results on the new warning distance. For all statistical testing we apply an alpha value of .05.

4.2 Ellipse Data

After executing the ellipse fitting algorithm for all 20 months of data, we examine several different outcomes for how the algorithm processes the data. Table 10 provides a compilation of what we find for each month as well as overall for all 20 months. Particularly, column one of Table 10 references the total number of ellipses for each month; this can also be considered the total number of storms in each month as each ellipse represents a single storm. August and July of 2015 have the greatest number of storms with 256 and 252 respectively.

One especially interesting result of the algorithm is that 25% of all ellipses are ellipses with only one flash. This means that there are single flashes either too far away from another storm and/or the time difference between these flashes and those before or after it are greater than 30 minutes. A closer examination of the data provides that of the 730 single flash ellipses, 25% of those are a result of time and/or distance isolation. The other 75% are a result of the assumptions and limitations of

the algorithm. We discuss an alternative ellipse fitting method as a possible solution to this issue in Chapter V.

Table 10: Ellipse Statistics for Each Month

Month	Total # Ellipses	Mean # Flashes in Each Ellipse	Largest # Flashes in an Ellipse	Total # Ellipses with 1 Flash	% Ellipses with 1 flash
May-13	84	361	20,266	29	35%
Jun-13	148	438	10,962	26	18%
Jul-13	127	188	3,750	39	31%
Aug-13	174	140	3,484	44	25%
Sep-13	79	285	12,113	24	30%
May-14	98	121	3,597	25	26%
Jun-14	150	389	12,587	35	23%
Jul-14	216	585	21,675	54	25%
Aug-14	151	489	11,147	37	25%
Sep-14	173	132	3,990	46	27%
May-15	45	247	6,228	14	31%
Jun-15	162	270	11,986	42	26%
Jul-15	252	204	9,765	59	23%
Aug-15	256	218	17,695	56	22%
Sep-15	157	170	6,212	40	25%
May-16	109	200	9,891	25	23%
Jun-16	176	189	6,208	38	22%
Jul-16	101	308	19,339	20	20%
Aug-16	96	84	1,382	23	24%
Sep-16	209	319	9,878	54	26%
Overall	2,963	273	21,675	730	25%

To answer our secondary research question, we next investigate the behavior of the initial flashes in a lightning storm from our data. To begin this part of the analysis we determine to only examine ellipses with 10 or more flashes. Of the 2,963 ellipses

created from the algorithm, 1,178 ellipses, about 40% of the ellipses, have 10 or more flashes. For each ellipse in each month we find the mean distance from the center for the first five flashes as well as the first ten flashes. Table 11 provides a summary of these means where the values reported are the means of the means and the overall mean is calculated for the entirety of the data, that is for all 1,178 ellipses.

Table 11: Mean Distance from Center for initial Flashes in a Lightning Storm

Month	**Mean Distance (NM) for 5 Flashes**	**Mean Distance (NM) for 10 Flashes**
May-13	3.26	3.65
Jun-13	2.94	3.28
Jul-13	3.61	4.16
Aug-13	3.09	3.47
Sep-13	2.97	3.36
May-14	3.70	4.30
Jun-14	3.38	3.83
Jul-14	2.69	2.96
Aug-14	2.69	2.99
Sep-14	3.07	3.74
May-15	3.11	3.45
Jun-15	3.05	3.39
Jul-15	3.19	3.70
Aug-15	3.20	3.67
Sep-15	3.83	4.01
May-16	3.31	3.99
Jun-16	3.10	3.48
Jul-16	1.80	2.11
Aug-16	3.22	3.44
Sep-16	3.56	3.94
Overall	3.13	3.53

From Table 11 we see that the mean distance for the first few flashes is between 3.13 and 3.53 NM depending on the number of flashes being considered. The 95% confidence intervals for these means are (3.03,3.22) and (3.42,3.63) respectively. Further descriptive statistics to include the histograms for these mean distances of the initial flashes in a lightning storm can be found in Appendix K. Conducting a paired t-test on the means of five and ten flashes gives us a p-value of $< .0001$ which suggests at a significance level of $\alpha = .05$ that there is enough evidence to conclude that the difference between the mean distances using five flashes and the mean distances using ten flashes is not zero. This difference can also be observed by the lack of overlap in the respective confidence intervals. Although there is a detectable difference in the mean distance from the center when considering five flashes versus ten flashes, we now have numerical evidence that the first few flashes of a storm, where we define the first few flashes as anywhere from five to ten flashes, occur within a 4 NM radius of the center of the storm on average. Now we consider the main research question of the distribution of lightning strikes beyond the edge of a preexisting area.

4.3 Distribution Fitting

Using the results of the ellipse fitting algorithm we extracted a total of 48,134 distances from the edge of an ellipse. A histogram of these distances can be seen in Figure 18. The maximum distance being less than 16 is due to the distance between storms restriction that was included in the ellipse fitting algorithm.

The two primary distributions we consider fitting to this data are the GEV and the Weibull. The application of the GEV to the data can be seen in Figure 19. The parameters of this particular GEV make it a Fréchet distribution since $\xi = .834 > 0$. The GEV fits fairly well, however, it does not capture the full extent of the higher number of observations at shorter distances. Conversely, the Weibull

distribution, seen in Figure 20, also offers an adequate fit while capturing the higher frequency of shorter distances. We did not attempt to find the distribution in terms of order statistics as this would require looking at the maximum order statistics from the distribution of the distances from the center of the ellipse and then taking a transformation to characterize a parent distribution of these distance measurements.

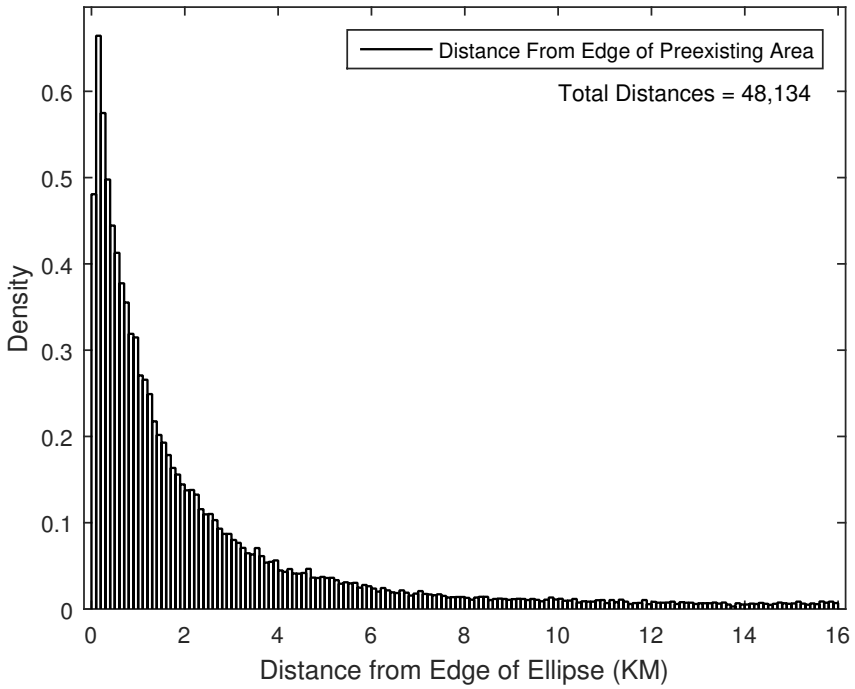


Figure 18: Histogram of the Distance (KM) From the Edge of a Preexisting Area

Conducting a GoF test for each of these distributions is the next logical step; however, because our sample size is so high, any GoF test applied to the data would result in a rejection of the null hypothesis. That is, the p-values from any GoF test would be so low that any test would determine the hypothesized distribution is not an adequate fit. This is due to the fact that such a large sample size is indicative of having an entire population rather than just a sample. Therefore, we conclude based on the general fit of both distributions that the Weibull is the best choice when looking at the probability of strike beyond the edge of a preexisting area. The specific

details of the particular Weibull distribution fit to the data, along with a table of the probability of strike can be found in Appendix L.

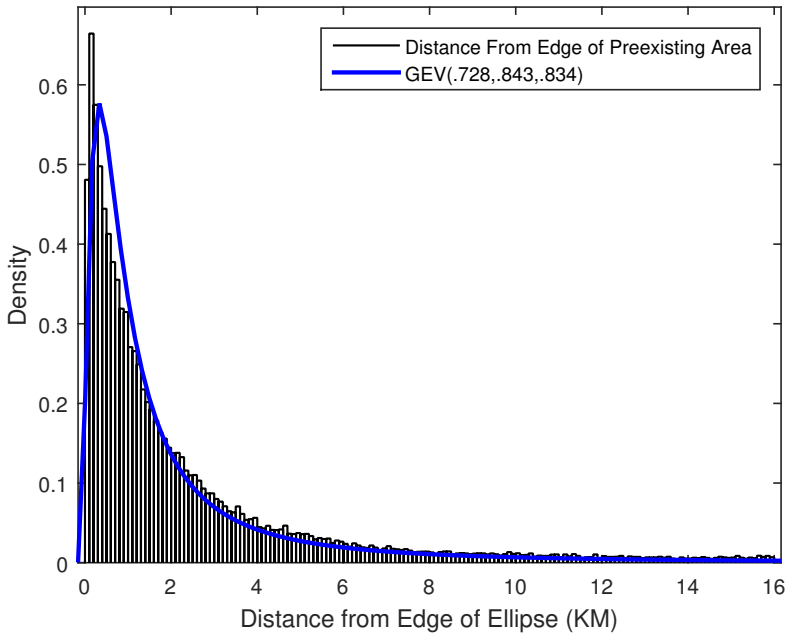


Figure 19: GEV fit to Distance from the Edge of a Preexisting Area

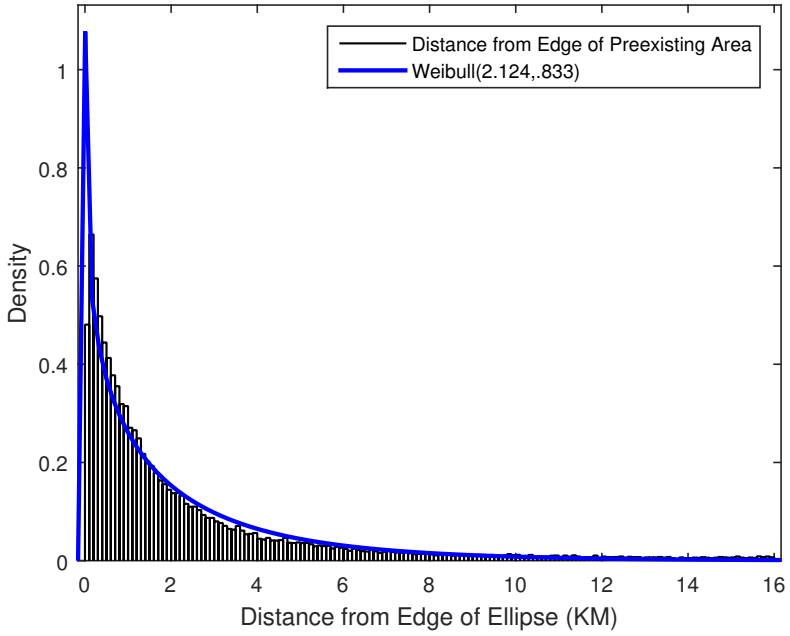


Figure 20: Weibull fit to Distance from the Edge of a Preexisting Area

The next step of our analysis process is to determine if a new warning distance may be safely established. The primary concern in decreasing the distance of warning for lightning strike is the extra risk incurred with such a reduction. As we decrease the warning distance we can determine from the established Weibull distribution the percentage of risk increase. This percentage of risk increase is then compared to the amount of area decreased when considering the warning distance as the radius of a warning circle as utilized by the 45 WS. Table 12 displays the percentage of risk increase and area decrease from the current 5 NM radius.

Table 12: Percentage of Area Decrease and Risk Increase from 5 NM Radius

Radius (NM)	**Percentage of Area Decrease from 5 NM**	**Percentage of Risk Increase from 5 NM**
4.75	10%	15%
4.5	19%	33%
4.25	28%	54%
4	36%	78%
3.75	44%	107%
3.5	51%	140%
3.25	58%	179%
3	64%	226%

From Table 12 we calculate the percentage of rate gained for both the risk increase and area reduction. The plot of these percentages can be seen in Figure 21. From Figure 21 we see that the amount of rate gained for both risk and area diminishes substantially around 4 NM. Therefore, we elect 4 NM as the new alternative warning distance that offers an appropriate balance of area reduction and increased risk. Now that a new radius has been established, we move to testing this radius empirically

using the 45 WS warning circles. It is important to note here that the specific Weibull distribution that is fit to this set of data is not intended to perfectly model the strike distance of lightning outside a preexisting area; rather, this distribution serves as a stepping tool to select a new potential radius. The results of the empirical validation process provide the strongest justification for accepting or rejecting a shorter warning distance.

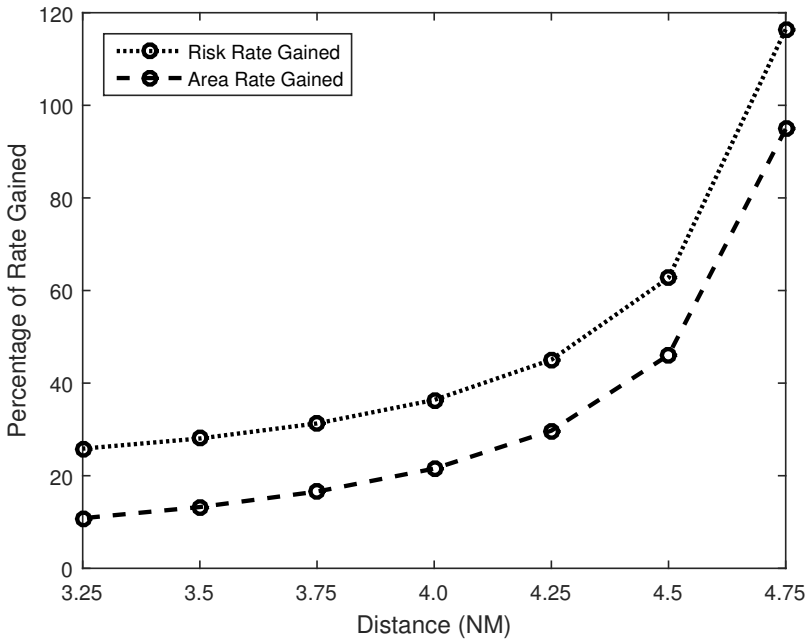


Figure 21: Percentage of Risk and Area Rate Gained from 5 NM Warning Distance

4.4 Empirical Validation Results

The first empirical testing we apply is the chosen 4 NM radius. Table 13 provides a breakdown of the results of this test by warning circle, while Table 14 provides a breakdown of the results by month. From Table 14, we find a reduction for all the warning circles from 5 or 6 NM to 4 NM provides a savings of 720.15 hours or 90.02 8-hour man days with only .277% of failure. This is a savings of about 22.5 8-hour man days a year just in the five summer months. We also find that the 4

NM reduction produces a total of 724 false alarms saved; this gives us an average of 181 false alarms saved a year in the five summer months. The hours saved from these prevented false alarms are included in the overall time savings and account for roughly 60% of the total time saved. The breakdown by circle reveals that eight of the ten warning circles experienced at least one failure, with the IA, 37/ITL, and Port circles having two failures each. The highest percentage of failures by warning circle is 0.571% for the 37/ITL circle as it saw two failures out of three hundred fifty total storms. To ensure that a reduction to 4 NM is the appropriate choice, we also complete the empirical test for several other reductions.

Table 13: Results by Circle for Reduction of Warning Radius to 4 NM

Warning Circle	Radius (NM)	Hours Saved (8 hr man days)	# False Alarms Saved	# Failures	Total Storms	% Failures
HAULOVER	5	42.96 (5.37)	45	1	376	0.266%
ASTROTECH	5	62.11 (7.76)	66	1	445	0.225%
LC39	6	112.67 (14.08)	124	1	421	0.238%
SLF	6	115.50 (14.44)	112	0	449	0.000%
IA	6	98.17 (12.27)	87	2	431	0.464%
40/41	5	40.35 (5.04)	43	1	349	0.287%
37/ITL	5	44.37 (5.55)	45	2	350	0.571%
CENTRAL	6	104.10 (13.01)	96	0	399	0.000%
PORT	5	49.17 (6.15)	47	2	380	0.526%
PAFB	5	50.74 (6.34)	59	1	370	0.270%

Specifically, we test a reduction from the 5 or 6 NM to 4.5, 4.25, 3.75, and 3.5 NM; we also compare the 4 NM reduction failure count to the number of failures seen if the warning circles have 6 and 7 NM radii reduced to the current 5 and 6 NM radii respectively. Finally, we compare the time saved and number of failures for reducing 5 and 6 NM to 4 and 5 NM respectively in order to determine if making the larger jump from 6 to 4 NM is having a greater impact on the total number of failures. The results of these tests can be seen in Table 15.

Table 14: Results by Month for Reduction of Warning Radius to 4 NM

Month	Hours Saved (8 hr man days)	# False Alarms Saved	# of Failures	Total Storms	% of Failures
May-13	23.01 (2.88)	29	0	106	0.000%
Jun-13	36.98 (4.62)	37	1	240	0.417%
Jul-13	32.07 (4.01)	31	0	181	0.000%
Aug-13	49.35 (6.17)	61	0	229	0.000%
Sep-13	24.77 (3.10)	26	0	140	0.000%
May-14	14.64 (1.83)	16	0	86	0.000%
Jun-14	37.62 (4.70)	39	1	198	0.505%
Jul-14	58.59 (7.32)	45	1	352	0.284%
Aug-14	48.90 (6.11)	45	1	265	0.377%
Sep-14	41.01 (5.13)	42	1	217	0.461%
May-15	11.58 (1.45)	13	0	65	0.000%
Jun-15	30.69 (3.84)	29	0	165	0.000%
Jul-15	48.01 (6.00)	43	0	282	0.000%
Aug-15	49.85 (6.23)	57	3	284	1.056%
Sep-15	43.52 (5.44)	45	1	213	0.469%
May-16	23.44 (2.93)	20	0	173	0.000%
Jun-16	37.77 (4.72)	37	2	219	0.913%
Jul-16	22.14 (2.77)	22	0	102	0.000%
Aug-16	26.81 (3.35)	28	0	151	0.000%
Sep-16	59.40 (7.43)	59	0	302	0.000%
Total	720.15 (90.02)	724	11	3,970	0.277%

From Table 15 we see that the current radii of 5 and 6 NM produces six failures, only five fewer than the newly selected 4 NM radius. We also see a decrease of 30% and 28% respectively in the number of hours saved and the number of false alarms saved, with only two fewer failures when reducing the 5 and 6 NM radii to 4 and 5 NM radii as compared to reducing all the circles to 4 NM.

Table 15: Number of Failures, Hours Saved, and False Alarms Saved at Different Warning Distances

Original Radii (NM)	New Radii (NM)	Hours Saved (8-hr man days)	# False Alarms Saved	# of Failures	% of Failures
6 and 7	5 and 6	-	-	6	0.151%
5 and 6	4.5	463.25 (57.91)	479	8	0.202%
5 and 6	4.25	588.51 (73.56)	600	9	0.227%
5 and 6	4 and 5	505.59 (63.20)	523	9	0.227%
5 and 6	4	720.15 (90.02)	724	11	0.277%
5 and 6	3.75	848.32 (106.04)	837	19	0.479%
5 and 6	3.5	966.62 (120.83)	932	27	0.680%

A visualization of the data from Table 15 can be seen in Figure 22 where we see the number of failures at each new warning distance plotted simultaneously with the number of days saved and hundreds of false alarms saved. The conversions from hours saved to days saved and the scaling of the number of false alarms saved by 100 were made to allow for concurrent plotting of the three statistics. From Figure 22 we see that when specifically considering the number of failures at each warning distance, 4 NM again proves to be the pivotal distance. This is primarily due to the number of failures more than doubling with an additional eleven failures from the current 5 and 6 NM when reduced to 3.75 NM as opposed to the increase of only five failures when moving from 5 and 6 NM to 4 NM.

Part of this investigation poses the question as to whether or not the current warning circles that have 6 NM radii are increasing the number of failures when moving directly from a 6 NM to 4 NM warning distance. When the reduction is made to 4 and 5 NM respectively from the previous 5 and 6 NM, there are only two fewer failures. Both of these failures can be attributed to warning circles currently at 6 NM. However, Table 16 and Figure 23 demonstrate that overall, it is not the 6 NM warning circles contributing to the bulk of the failures. From Table 16 and Figure 23 we see that the increase in failures for the 6 NM warning circles is relatively

consistent while the greater number of failures occurring at the shorter radii are due to the larger increase in failures for the 5 NM warning circles.

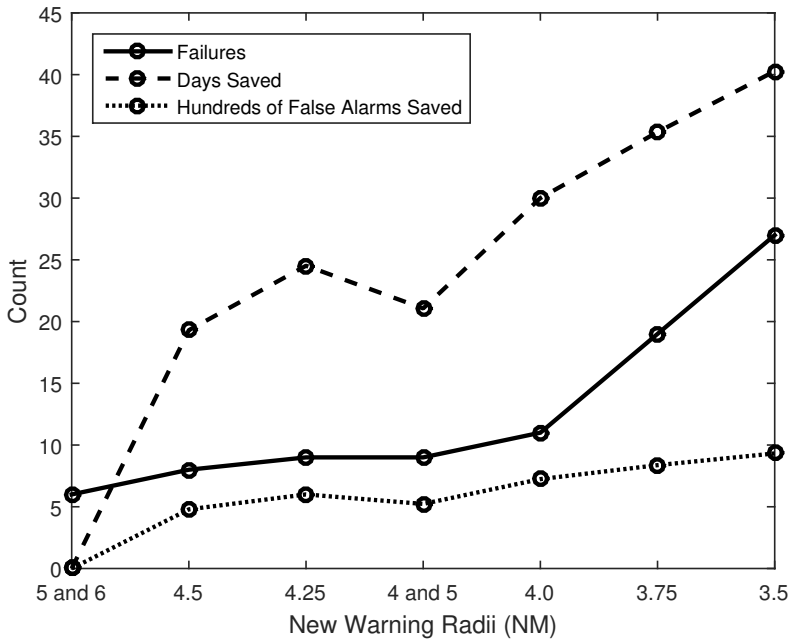


Figure 22: Number of Failures, Days Saved, and Hundreds of False Alarms Saved at Different Warning Distances

Table 16: Number of Failures by Current Warning Circle Radii (NM)

Current Warning Circle Radii (NM)	**New Warning Circle Radii (NM)**					
	4.5	4.25	4 and 5	4	3.75	3.5
5	6	7	8	8	14	20
6	2	2	1	3	5	7
Total	8	9	9	11	19	27

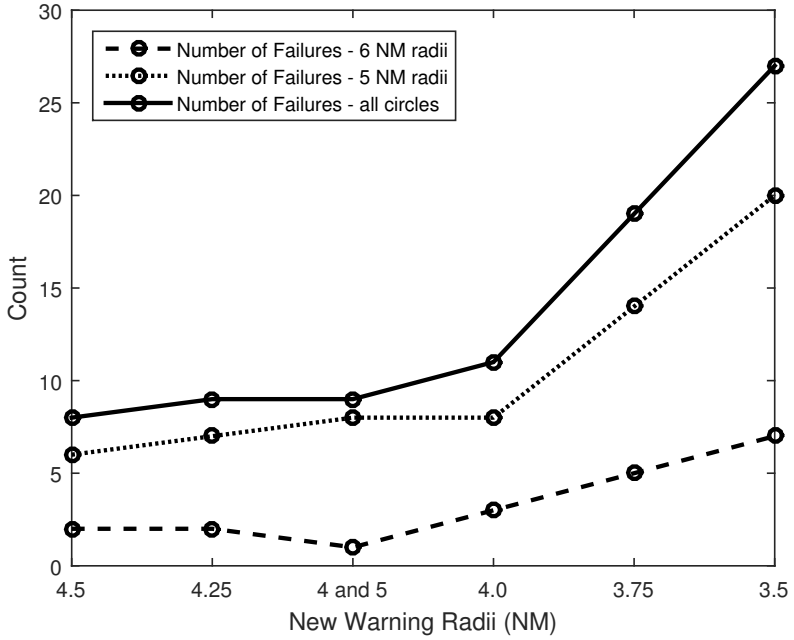


Figure 23: Number of Failures by Current Warning Circle Radii (NM)

4.5 Summary

In conclusion, we found using the ellipse fitting algorithm results that the initial strikes of a storm on average occur within 4 NM of the center of the storm. Completing the process of fitting the distances from the edge of a preexisting lightning area yielded a Weibull distribution fit. Though the particular Weibull distribution chosen is not meant to be a precise model of the actual probability of strike, we used it as a tool to determine the increase in probability of strike. Comparing the rate gained of risk increase and area reduction led to the selection of 4 NM as the new warning radius. Finally, the empirical validation results confirmed the selection of the 4 NM warning radius as the best choice when considering the amount time and the number of false alarms saved versus the extra risk incurred. In the next chapter, we discuss the results from this study as compared to past results. We also suggest further areas of study to include possible changes or follow-on studies from this work.

V. Conclusions

5.1 Overview

The purpose of this section is to provide a summary of the method applied to the primary research problem as well as the results of that method. We also discuss the differences and similarities of this method in regards to past research on the same topic. Additionally, we present an alternative methodology for solving the same problem, including the advantages and disadvantages of using a different method. We also suggest future work that can be explored regarding this research. Finally, we offer our conclusions as to the operational application of the results of this study.

5.2 Results and Comparison to Past Research

Because of the discrepancy of past research methodologies and the manner in which lightning warnings are issued, meteorologists at the 45 WS recognized the need for a study that considered the distribution of the distance lightning travels beyond the edge of a preexisting area. We found, using ellipses as the boundary, that an extreme value distribution, specifically the Weibull distribution, may be used to fit the distance lightning strikes beyond the edge of a preexisting area. Applying the Weibull distribution and ascertaining the amount of additional risk incurred at shorter distances, the distance of 4 NM was chosen as the new radius to empirically test. Comparing the empirical results of 4 NM to various other choices, the 4 NM radius continued to present as the appropriate choice based on the amount of additional risk it produced.

Past studies focused on the distance lightning travels from the center of an area designated as a storm; thus, there are fewer comparisons to be made as the purpose of the studies differed slightly. We applied similar time and spatial constraints to our

data as that of previous work. We used 20 months of data from a single location, a considerable amount more than most other studies. The amount of data we used fell third in magnitude behind Parsons (2000) who used four years worth of data covering the majority of the US and McNamara (2002) whose study also included nearly four years worth of data from the 45 WS.

One similarity between our study and the majority of previous work, save McNamara (2002), is the apparent problem of not accounting for storm origin. Recall from Chapter II that McNamara (2002) corrected for this by pairing flashified LDAR data, much like what we used in our study, with NLDN data, which provides actual strike location. Our omission of the concept of storm origin was opposite that of the other studies in that we used the in-cloud data and assumed ground strike location at any of the various LDAR source point locations. We discuss a solution to this lack of origin problem in the next section.

Briefly discussed in Chapter IV was the issue of isolated flashes. Various other studies mentioned in Chapter II provide the single flash results of their studies. Parsons (2000) was particularly concerned with the number of isolated flashes in studies predating hers and attempted to reduce the number of isolated flashes her method produced. Although we mention this as a similarity to other studies, it is not a primary concern to us; however, we do suggest in the next section a different method that might alleviate some of the isolated flashes.

5.3 Alternative Methodology and Future Research

While creating the primary ellipse fitting algorithm an alternative method that could be applied to this data became perceptible. One of the hardest portions of the phenomena of the lightning storm to model is the concept of multiple storms occurring at the same time as well as the merging of two or more storm cells. The

assumptions and conditions placed in the ellipse fitting algorithm of this research attempted to correct for these issues as much as possible. Therefore, an alternative method that arose as a potential solution to these problems is the idea of a dynamic clustering analysis.

Clustering analysis would deal with similar issues as those we discovered in the ellipse fitting algorithm. This includes how to decipher to which storm a particular flash belongs when a flash occurs between two or more storms, or what is the criteria for merging multiple storm cells. Although similar decisions must be made using a dynamic clustering analysis, the results of such an application would be interesting to observe, especially as a comparison to the results we found with our ellipse fitting algorithm. Two specific advantages of a dynamic clustering algorithm would be a reduction in the mean area of the ellipses as well as a reduction in the number of isolated flashes. An example of a k-means cluster analysis on the same set of flashes from Chapters II and III can be seen in Figure 24. To create this example we chose to find three distinct clusters and then applied the MVEE to these specific clusters. From Figure 24 we already see the problem of storms merging between storms one and two.

In addition to a comparative study between cluster analysis and the method we chose to adapt, further work could be done with our data. Primarily, a pairing of the strike data, collected by the updated CGLSS, with our LDAR data to account for the idea of lightning origin and provide true ground strike location would offer additional justification of the warning distance reduction. Also, further empirical testing of the new 4 NM warning radius using the current MERLIN system would be useful. Finally, similar research in other geographic locations would help to solidify the reduction in warning radius.

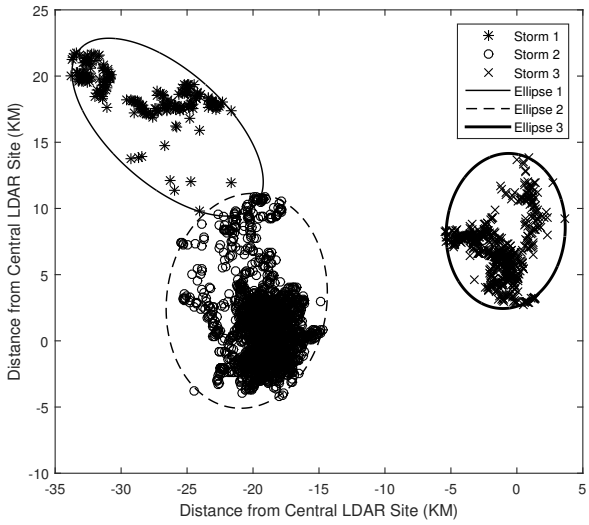


Figure 24: K-Means Cluster Analysis Ellipse Fitting Approach

5.4 Final Remarks

While there is further research to be done on this particular topic, the results of this study are very revealing. The amount of risk incurred with a reduction of 4 NM from the current 5 NM lightning warning distance of .277% is acceptable, especially given that there is presently risk at the current distance. We saw with a 4 NM warning radius at the 45 WS an immense reduction in the amount of time saved with 22.5 8-hr man days recovered as well as an average of 181 false alarms saved in five months a year alone. As suggested previously, further studies of different locations and storm types would offer even greater support of a reduction in the lightning warning radius. However, because the 45 WS has so much invested in the assurance of safety of life and property due to the nature of their mission, much of the decision to update the AFI 91-203 lightning safety standards has historically been dictated by the 45 WS. Therefore, with the results of this study, we suggest that the AFI 91-203 lightning safety standards be changed to reflect a 4 NM watch and warning distance across the United States Air Force (USAF).

Appendices

A. List of Acronyms

2D 2-Dimensional

3D 3-Dimensional

45 WS 45 Weather Squadron

4DLSS Four Dimensional Lightning Surveillance System

A-D Anderson-Darling

AF Air Force

AFI 91-203 Air Force Instruction 91-203

AMS American Meteorological Society

CCAFS Cape Canaveral Air Force Station

CDC Center for Disease Control

CDF Cumulative Distribution Function

CG Cloud-to-Ground

CGLSS Cloud-to-Ground Lightning Surveillance System

DBSF Distance Between Successive Flash

DoD Department of Defense

EDF Empirical Distribution Function

EVD Extreme Value Distribution

EVT Extreme Value Theorem

GEV Generalized Extreme Value

GoF Goodness-of-Fit

GUI Graphical User Interface

IDL Interactive Data Language

IID Independent and Identically Distributed

KM Kilometers

K-S Kolmogorov-Smirnov

KSC Kennedy Space Center

LDAR Lightning Detection and Ranging

LPLWS Launch Pad Lightning Warning System

LSG Lightning Safety Group

M Statute Miles

MERLIN Mesoscale Eastern Range Lightning Information Network

MVEE Minimum Volume Enclosing Ellipsoid

NASA National Aeronautics and Space Administration

NDSF National Deep Submergence Facility

NLDN National Lightning Detection Network

NM Nautical Miles

NOAA National Oceanic and Atmospheric Administration

NSSL National Severe Storms Laboratory

NWS National Weather Service

OSHA Occupational Safety and Health Administration

PAFB Patrick Air Force Base

PCA Principal Component Analysis

SCIT Storm Cell Identification and Tracking

TOA Time of Arrival

US United States

USAF United States Air Force

VHF Very High Frequency

WATADS WSR-88D Algorithm Testing and Display System

WSR-88D Weather Surveillance Radar - 88 Delta

B. Lightning Strike Process

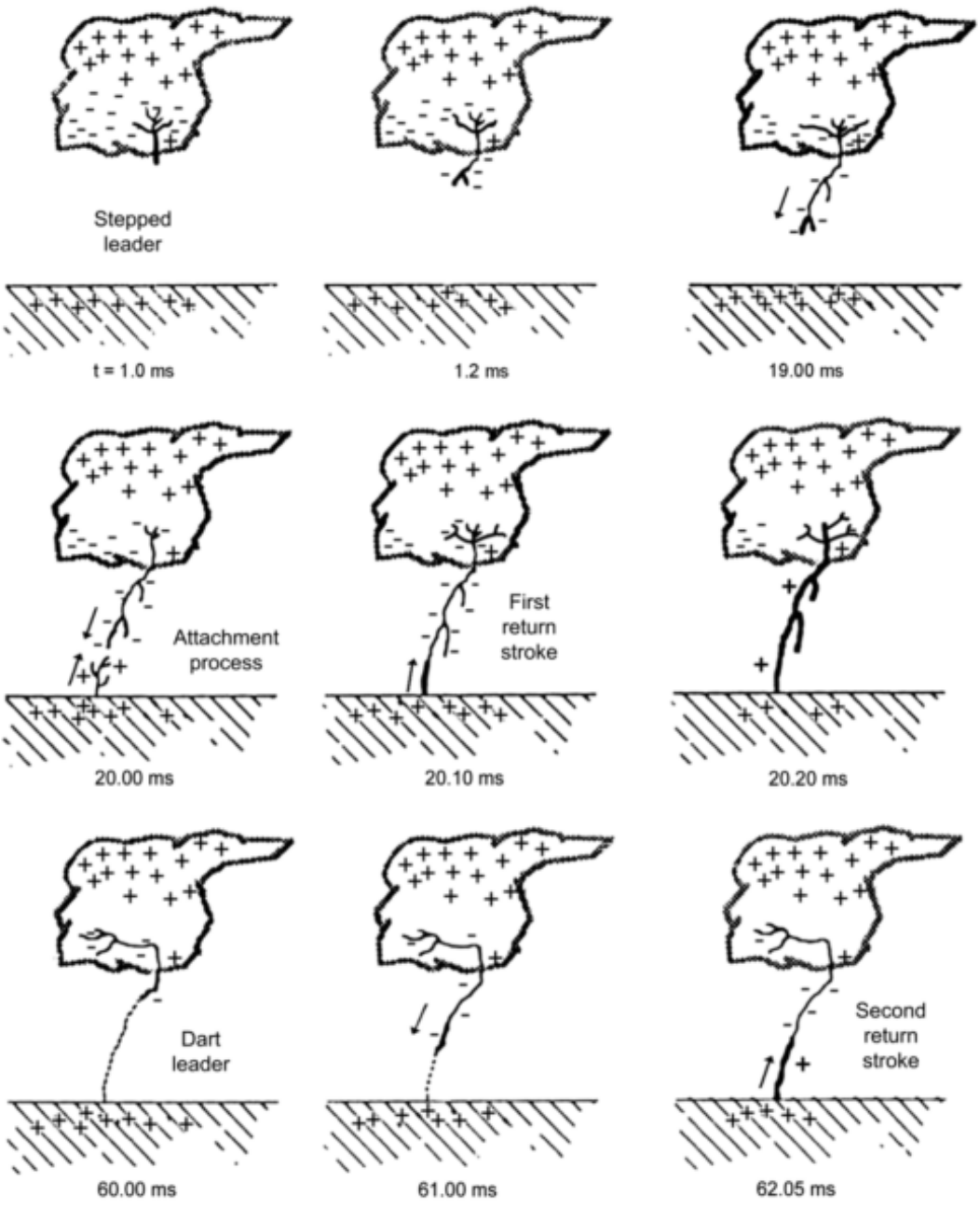


Figure 25: Steps of the Lightning Strike Process

C. Ellipse Representations and Conversions

Parametric: *h, k, a, b,* and τ

$$\boldsymbol{\varepsilon} = \left\{ (x,y) \mid \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} h \\ k \end{bmatrix} + \begin{bmatrix} \cos(\tau) & -\sin(\tau) \\ \sin(\tau) & \cos(\tau) \end{bmatrix} \begin{bmatrix} a\cos(t) \\ b\sin(t) \end{bmatrix}, 0 \le t \le 2\pi \right\}$$

Conic: *B, C, D, E,* and *F*

$$\boldsymbol{\varepsilon} = \left\{ (x,y) \mid x^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \right\}$$

Foci/String: $\alpha_1, \beta_1, \alpha_2, \beta_2$, and s

$$\boldsymbol{\varepsilon} = \left\{ (x,y) \mid \sqrt{(x-\alpha_1)^2 + (y-\beta_1)^2} + \sqrt{(x-\alpha_2)^2 + (y-\beta_2)^2} = s \right\}$$

Conic $\rightarrow$ Parametric

If $Ax^2+Bxy+Cy^2+Dx+Ey+F=0$ specifies an ellipses and we define the matrices:

$$M_0 = \begin{bmatrix} F & D/2 & E/2 \\ D/2 & A & B/2 \\ E/2 & B/2 & C \end{bmatrix} \qquad M = \begin{bmatrix} A & B/2 \\ B/2 & C \end{bmatrix}$$

then,

$$a = \sqrt{-\det(M_0)/(\det(M)\lambda_1)} \qquad b = \sqrt{-\det(M_0)/(\det(M)\lambda_2)}$$

$$h = (BE - 2CD)/(4AC - B^2) \qquad k = (BD - 2AE)/(4AC - B^2)$$

$$\tau = \arctan((A-C)/B)/2$$

where λ_1 and λ_2 are the eigenvalues of M ordered so that $|\lambda_1 - A| \le |\lambda_1 - C|$. (This ensures that $|\lambda_2 - C| \le |\lambda_2 - A|$.)

Parametric → Conic

If $c = \cos(\tau)$ and $s = \sin(\tau)$, then the ellipse

$$x(t) = h + c[a\cos(t)] - s[b\sin(t)]$$

$$y(t) = k + s[a\cos(t)] + c[b\sin(t)]$$

is also specified by $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$ where

$$A = (bc)^2 + (as)^2$$

$$B = -2cs(a^2 - b^2)$$

$$C = (bs)^2 + (ac)^2$$

$$D = -2Ah - kB$$

$$E = -2Ck - hB$$

$$F = -(ab)^2 + Ah^2 + Bhk + Ck^2$$

Parametric → Foci/String

Let $\boldsymbol{\varepsilon}$ be the ellipse

$$x(t) = h + \cos(\tau)[a\cos(t)] - \sin(\tau)[b\sin(t)]$$

$$y(t) = k + \sin(\tau)[a\cos(t)] + \cos(\tau)[b\sin(t)]$$

If $c = \sqrt{a^2 - b^2}$ then $\boldsymbol{\varepsilon}$ has foci

$$F_1 = (h - \cos(\tau)c, k - \sin(\tau)c) \qquad F_2 = (h + \cos(\tau)C, k + \sin(\tau)C)$$

and string length

$$s = 2a$$

Foci/String → Parametric

If $F_1 = (\alpha_1, \beta_1), F_2 = (\alpha_2, \beta_2)$ and s defines an ellipse then

$$a = s/2$$

$$b = \sqrt{s^2 - ((\alpha_1 - \alpha_2)^2 + (\beta_1 - \beta_2)^2)}/2$$

$$h = (\alpha_1 + \alpha_2)/2$$

$$k = (\beta_1 + \beta_2)/2$$

$$\tau = \arctan((\beta_2 - \beta_1)/(\alpha_2 - \alpha_1))$$

D. Design, Constraint, and Scatter Matrices of the Fitzgibbon Least Squares Ellipse Fitting Approach

Design Matrix

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{x_1^2} & \boldsymbol{x_1 y_1} & \boldsymbol{y_1^2} & \boldsymbol{x_1} & \boldsymbol{y_1} & \boldsymbol{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{x_i^2} & \boldsymbol{x_i y_i} & \boldsymbol{y_i^2} & \boldsymbol{x_i} & \boldsymbol{y_i} & \boldsymbol{1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{x_N^2} & \boldsymbol{x_N y_N} & \boldsymbol{y_N^2} & \boldsymbol{x_N} & \boldsymbol{y_N} & \boldsymbol{1} \end{bmatrix}$$

Constraint Matrix

$$C = \begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Scatter Matrix

$$S = D^T D$$

$$= \begin{bmatrix} S_{x^4} & s_{x^3y} & S_{x^2y^2} & S_{x^3} & S_{x^2y} & S_{x^2} \\ S_{x^3y} & S_{x^2y^2} & S_{xy^3} & S_{x^2y} & S_{xy^2} & S_{xy} \\ S_{x^2y^2} & S_{xy^3} & S_{y^4} & S_{xy^2} & S_{y^3} & S_{y^2} \\ S_{x^3} & S_{x^2y} & S_{xy^2} & S_{x^2} & S_{xy} & S_x \\ S_{x^2y} & S_{xy^2} & S_{y^3} & S_{xy} & S_{y^2} & S_y \\ S_{x^2} & S_{xy} & S_{y^2} & S_x & S_y & S_1 \end{bmatrix}$$

E. LDAR Data Sample

Date	Time	X	Y	Z	Epoch Time
01/07/2013	00:58:08:442365	-23634	-126106	10450	1357538288.442365
01/07/2013	01:02:40:561306	-20468	-58349	6492	1357538560.561306
01/07/2013	01:02:40:56143	-20846	-57710	6262	1357538560.56143
01/07/2013	01:02:40:563187	-20955	-57658	6968	1357538560.563187
01/07/2013	01:02:40:563633	-18778	-58969	7399	1357538560.563633
01/07/2013	01:02:40:564347	-19129	-61794	10385	1357538560.564347
01/07/2013	01:02:40:564678	-19100	-59666	6212	1357538560.564678
01/07/2013	01:02:40:565384	-19065	-56984	7530	1357538560.565384
01/07/2013	01:02:40:565833	-18570	-57859	7201	1357538560.565833
01/07/2013	01:02:40:568385	-18826	-55372	8458	1357538560.568385
01/07/2013	01:02:40:569221	-18177	-56798	7685	1357538560.569221
01/07/2013	01:02:40:569344	-17871	-55716	7164	1357538560.569344
01/07/2013	01:02:40:569954	-18403	-56330	7539	1357538560.569954
01/07/2013	01:02:40:573603	-19857	-55328	7232	1357538560.573603
01/07/2013	01:02:40:574448	-21581	-56418	6816	1357538560.574448
01/07/2013	01:02:40:574825	-17578	-57205	7528	1357538560.574825
01/07/2013	01:02:40:575795	-21655	-57388	6873	1357538560.575795
01/07/2013	01:02:40:577273	-19918	-56176	7115	1357538560.577273
01/07/2013	01:02:40:578929	-19581	-59359	6973	1357538560.578929
01/07/2013	01:02:40:580822	-19288	-54003	7130	1357538560.580822
01/07/2013	01:02:40:588608	-18725	-54068	7022	1357538560.588608
01/07/2013	01:02:40:592467	-18857	-52932	8758	1357538560.592467
01/07/2013	01:02:40:603366	-20878	-58533	6172	1357538560.603366

01/07/2013 01:02:40:61154,-21118,-59062,5938,1357538560.61154

01/07/2013 01:02:40:617129,-21432,-58926,6225,1357538560.617129

01/07/2013 01:04:53:424723,64603,4312,7624,1357538693.424723

01/07/2013 01:04:53:416047,59911,4527,6052,1357538693.416047

01/07/2013 01:04:53:416221,59598,4388,6051,1357538693.416221

01/07/2013 01:04:53:425215,59188,5224,6842,1357538693.425215

01/07/2013 01:04:53:432122,59371,5021,6771,1357538693.432122

01/07/2013 01:04:53:432539,59801,5129,6892,1357538693.432539

01/07/2013 01:04:53:43475,61915,5119,6413,1357538693.43475

01/07/2013 01:04:53:436361,60388,5108,6988,1357538693.436361

01/07/2013 01:04:53:437636,58233,5579,8100,1357538693.437636

01/07/2013 01:04:53:439928,61687,5327,6498,1357538693.439928

01/07/2013 01:04:53:442614,62915,5535,6481,1357538693.442614

01/07/2013 01:04:53:44505,59129,6168,6945,1357538693.44505

01/07/2013 01:04:53:468761,63627,6067,9137,1357538693.468761

01/07/2013 01:04:53:480643,59279,5822,7077,1357538693.480643

01/07/2013 01:04:53:483627,60567,6893,7237,1357538693.483627

01/07/2013 01:04:53:488159,63754,7208,6855,1357538693.488159

01/07/2013 01:04:53:489197,61920,7748,7243,1357538693.489197

01/07/2013 01:10:21:532066,-19170,-124046,10521,1357539021.532066

01/07/2013 01:14:25:251286,67865,10952,5758,1357539265.251286

01/07/2013 01:14:24:927772,61464,16105,3153,1357539264.927772

01/07/2013 01:14:24:899732,69816,17468,8373,1357539264.899732

01/07/2013 01:14:25:174882,68290,16900,2157,1357539265.174882

01/07/2013 01:14:24:886718,71644,14324,9189,1357539264.886718

01/07/2013 01:14:24:835782,69105,12175,5141,1357539264.835782

01/07/2013 01:14:24:83985,67512,12172,8240,1357539264.83985

01/07/2013 01:14:24:91162,68162,12695,4846,1357539264.91162

01/07/2013 01:14:24:832529,60837,10795,3821,1357539264.832529

01/07/2013 01:14:24:834874,63576,11024,6775,1357539264.834874

01/07/2013 01:14:24:835276,64821,11005,7807,1357539264.835276

01/07/2013 01:14:24:836762,62665,10501,6142,1357539264.836762

01/07/2013 01:14:24:837259,65252,10304,7519,1357539264.837259

01/07/2013 01:14:24:840787,65873,10842,8243,1357539264.840787

01/07/2013 01:14:24:846928,62411,10853,4729,1357539264.846928

01/07/2013 01:14:24:848087,64968,11256,6517,1357539264.848087

01/07/2013 01:14:24:851182,65615,11624,3553,1357539264.851182

01/07/2013 01:14:24:855456,65620,12453,7702,1357539264.855456

01/07/2013 01:14:24:864968,64424,13113,6762,1357539264.864968

01/07/2013 01:14:24:871735,65060,13328,6934,1357539264.871735

F. Data Processing Code

```
%Import Data
[date x0 y0 z0 epoch] = textread('F:\LDAR Thesis Data\LDAR_2013\LDAR_2013_06.txt','%s %f %f %f %f'
    , 'delimiter', ',');
[flash] = textread('F:\LDAR Thesis Data\LDAR_2013\LDAR_2013_06.txt','%s', 'delimiter', '\n');
x0 = x0/1000.;  %convert to km
y0 = y0/1000.;  %convert to km
%create all the counters
row_count = 1;                      %row counter
num_sourcepts = 0;                  %number of source points in flash counter
flashtot = 0;                       %counter for total number of flashes in first loop
pcount = 0;                         %counter for the number of source points in first loop for
    each flash
fpcount = zeros(1,length(flash));   %preallocate size of matrix that counts number of sourcepoints
     in each flash
%This loop goes through the flash data and counts the number of
%sourcepoints in each flash and keeps track of the total number of flashes
%the total number of source points for each flash is stored in fpcount
for i = 1:length(flash)
    if (~isempty(flash{i}))
        pcount = pcount + 1;
    else
        flashtot = flashtot + 1;
        fpcount(flashtot) = pcount;
        pcount = 0;
    end
end
flashtot = flashtot+1;       %total number of flashes in file (since last one isn't counted during
     the loop we add one)
fpcount(flashtot)= pcount;   %we have to input the last number of source points for the last flash
     as it's not accounted for in loop
max_pcount = max(fpcount)   %just gives max number of sourcepoints in a flash
           %total sourcepoints for all flahses w/ >=5 sourcepoints
          %total flashes >= 5 source points
%count number of sourcepoints and flashes to preallocate sizes for faster
%processing
fpcount_2 = find(fpcount >=5);
total_flashes = length(fpcount_2)
extra = zeros(1, total_flashes);
fpcount_2 = [fpcount_2; extra];
fpcount_2(2,:)= fpcount(fpcount_2(1,:));
total_srcpts = sum(fpcount_2(2,:))
date0 = zeros(total_flashes,11);    %preallocate size of date matrix
%date0: col 1 = flash #, col 2 = # srcpts in flash, col 3:8 = yyyy MM DD hh
%mm ss SSSSSS,
flash_x = zeros(max_pcount,total_flashes);
flash_y = zeros(max_pcount,total_flashes);
%this loop goes through and seperates flashes into useable form so that
%date/time stamps are accounted for for each flash/source point. It's only
%looking for flashes with >= 5 source points
last = 0;
dcount=1;                           %date counter
```

```
row_count2 = 1;                     %second row counter
flash_count = 0;                    %number of flashes in the file greater than/equal to 5 source
    points counter
col_count = 1;                      %column counter
source_start =1;                    %first starting point of loop and subsequent starting point
    after num_sourcepts is added
for i = 1:length(fpcount)
    num_sourcepts = fpcount(i); %makes current number of source points fpcount(i)
   if num_sourcepts >= 5
      last =  source_start + num_sourcepts-1;
      flash_count= flash_count + 1;
      date0(dcount,1)= flash_count;
      date0(dcount,2)=num_sourcepts;
      dcount = dcount+1;
      for n = source_start:last
         %flash_mat(row_count,1) = x0(n);
        % flash_mat(row_count,2) = y0(n);
          flash_x(row_count2,col_count)=x0(n);
          flash_y(row_count2,col_count)=y0(n);
        % row_count = row_count +1;
          row_count2 = row_count2+1;
      end
      source_start = last+1;
      col_count = col_count + 1;
      row_count2 = 1;
   else
       last =  source_start + num_sourcepts-1;
       source_start = last + 1;
   end
end
%reset counters for last loop
row_count = 1;                      %row counter
source_start = 1;                    %first starting point of loop and subsequent starting point
    after num_sourcepts is added
dates = cell(total_flashes,1); %preallocate the cell list for all the date/time stamps of the
    first sourcepoint of a strike
%this loop goes through and pulls the first time stamp for every flash w/ >= 5
%source points
for i = 1:length(fpcount)
    num_sourcepts = fpcount(i);
    if (num_sourcepts >= 5)
      %last =  source_start + num_sourcepts;
      %flash_count= flash_count + 1;
     dates{row_count,1} = date{source_start};
      source_start = source_start + num_sourcepts;
      row_count = row_count + 1;
    else
      source_start = source_start + num_sourcepts;
    end
end
%this loop goes through, converts the dates to readable form and then puts
%them into the date0 matrix so that the flash #, # source points in flash,
%and date/time stamp for each flash are all in the same
for i = 1:length(dates)
```

```
    conv_date = datetime(dates(i),'InputFormat', 'MM/dd/yyyy HH:mm:ss:SSSSSS','Format','dd-MM-yyyy
        HH:mm:ss.SSSSSS');
    %put dates into vector in date0 matrix
    date0(i,3:8) = datevec(conv_date);
end
%creates all the counters
num_sourcetpts = 0;
flash_extreme = zeros(max_pcount,total_flashes); %creates a matrix that lists all the extreme
    points for each flash
row_count = 1;
col_count = 1;
ext_start = 1;
last = 0;
tot = 0;
%this loop goes through and finds all the extreme points for each flash, it
%stores which points they are in the flash_extreme matrix and records the
%number of extreme points
for i = 1:total_flashes
num_sourcetpts = date0(i,2);
extreme = convhull(flash_x(1:num_sourcetpts,i),flash_y(1:num_sourcetpts,i));
tot = nnz(extreme)-1;
flash_extreme(1:tot,col_count) = extreme(1:tot);
date0(i,9) = tot;
col_count = col_count + 1;
end
most_extremes = max(date0(:,9)); %records the max number of extreme sourcepoints for any given
    flash
total_extremes = sum(date0(:,9)); %records the total number of extreme sourcepoints for all the
    flashes
%this loop goes through and creates seperate flash_x and flash_y matrices
%for the extreme points of each flash
flash_x_extremes = zeros(most_extremes,total_flashes);
flash_y_extremes = zeros(most_extremes,total_flashes);
row_count = 1;
for i = 1:total_flashes
  final = nnz(flash_extreme(:,i));
  for n = 1:final
      flash_x_extremes(row_count,i) = flash_x(flash_extreme(n,i),i);
      flash_y_extremes(row_count,i) = flash_y(flash_extreme(n,i),i);
      row_count = row_count + 1;
  end
  row_count = 1;
end
original_srcpts = sum(fpcount);

numextpts = 0; %counts number of extreme points for each flash
x1 = 0; %x coordinate for central site
x2 = 0; %x coordinate for each ext srcpt
y1 = 0; %y coordinate for central site
y2 = 0; %y coordinate for each ext srcpt
d = 0; %distance from srcpt to central site
date00 = zeros(total_flashes,2);
old_tot_srcpts = total_srcpts;
old_tot_flashes = total_flashes;
```

```
old_tot_extremes = total_extremes;
%this loop goes through and identifies all of the flashes that are outside a
%25 mile radius of the central site
for i = 1:total_flashes
    numextpts = date0(i,9);
    distance_check = zeros(numextpts,1);
    for w = 1:numextpts
        x2 = flash_x_extremes(w,i);
        y2 = flash_y_extremes(w,i);
        d = sqrt((x2-x1)^2 + (y2-y1)^2);
        if d > 40.2336
        distance_check(w,1) = 1;
        end
    end
    if sum(distance_check(:,1)) == 0
    date00(i,1) = 1; %indicates the flash is within the boundary
    date00(i,2) = d;
    end
end
inside_flashes = sum(date00(:,1));
first_inside_flash = find(date00(:,1)==1,1,'first');
last_inside_flash = find(date00(:,1)==1,1,'last');
in_flash_mat = zeros(inside_flashes,1);
in_flash_mat(:,1) = find(date00(:,1) == 1);
date0 = date0(in_flash_mat(:,1),:); %gives the date matrix of only those flashes that are within
    25 miles
flash_extreme = flash_extreme(:,in_flash_mat(:,1));
flash_x = flash_x(:,in_flash_mat(:,1));
flash_y = flash_y(:,in_flash_mat(:,1));
flash_x_extremes = flash_x_extremes(:,in_flash_mat(:,1));
flash_y_extremes = flash_y_extremes(:,in_flash_mat(:,1));
%this loop goes through and adds all of the extreme sourcepoints for each
%flash into the flash_mat matrix where col 1 = x-values for all source
%points, col 2 = y-values for all source points, col 3 = x-values for
%extreme sourcepoints, col 4 = y-values for extreme sourcepoints
row_count = 1;
row_count2 = 1;
total_srcpts = sum(date0(:,2));
flash_mat = zeros(total_srcpts,6); %preallocate size of flash matrix
for i = 1:inside_flashes
  final = nnz(flash_extreme(:,i));
  for n = 1:final
flash_mat(row_count,3) = flash_x(flash_extreme(n,i),i);
flash_mat(row_count,4) = flash_y(flash_extreme(n,i),i);
row_count = row_count + 1;
  end

  lastone = date0(i,2);
  for p = 1:lastone
     flash_mat(row_count2,1) = flash_x(p,i);
     flash_mat(row_count2,2) = flash_y(p,i);
     row_count2 = row_count2 + 1;
  end
end
```

```
%this loop goes through and adds the flash number associated with each
%sourcepoint to the flash_mat matrix
start = 1;
start_ext = 1;
last_ext = 0;
last = 0;
num_srcpts = 0;
num_extpts = 0;
for i = 1:length(date0)
  num_srcpts = date0(i,2);
  num_extpts = date0(i,9);
  last = start + num_srcpts - 1;
  last_ext = start_ext + num_extpts -1;
  for n = start:last
      flash_mat(n,5) = i;
  end
  for p = start_ext:last_ext
     flash_mat(p,6)=i;
  end
  start_ext = start_ext + num_extpts;
  start = start + num_srcpts;
end
total_srcpts = sum(date0(:,2));
total_flashes = inside_flashes;
total_extremes = sum(date0(:,9));
most_extremes = max(date0(:,9));
max_pcount = max(date0(:,2));
```

G. Convex Hull Code

```
function convexhull
a=imread('impcon.bmp');
a=a(:,:,3);
o=a;
subplot(2,2,1)
imshow(o)
title('original image');
[r,c]=size(a);
c=[1 0 0;1 0 0;1 0 0 ]
c1=[1 1 1;0 0 0;0 0 0 ];
c2=[0 0 1; 0 0 1;0 0 1];
c3=[0 0 0;0 0 0;1 1 1];
b=a;
for u=1:10
d=(imerode(b,c))|a;
if(b==d)
break;
end
b=d;
end
%------------
m=a;
for u=1:10
n=(imerode(m,c1))|a;
if(m==n)
break;
end
m=n;
end
%------------
p=a;
for u=1:10
l=(imerode(p,c2))|a;
if(p==l)
break;
end
p=l;
end
%------------
z=a;
for u=1:10
v=(imerode(z,c3))|a;
if(z==v)
break;
end
z=v;
end
%------------
H=z|p|m|b;
subplot(2,2,2)
imshow(H)
title('convex Hull of image');
```

H. Ellipse Fitting Functions

Least Squares Best Fit Ellipse Code:

```
function A = EllipseDirectFit(XY,color);
%  Direct ellipse fit, proposed in article
%    A. W. Fitzgibbon, M. Pilu, R. B. Fisher
%     "Direct Least Squares Fitting of Ellipses"
%     IEEE Trans. PAMI, Vol. 21, pages 476-480 (1999)
%  Our code is based on a numerically stable version
%  of this fit published by R. Halir and J. Flusser
%     Input:  XY(n,2) is the array of coordinates of n points x(i)=XY(i,1), y(i)=XY(i,2)
%     Output: A = [a b c d e f]' is the vector of algebraic
%             parameters of the fitting ellipse:
%             ax^2 + bxy + cy^2 +dx + ey + f = 0
%             the vector A is normed, so that ||A||=1
%  This is a fast non-iterative ellipse fit.
%  It returns ellipses only, even if points are
%  better approximated by a hyperbola.
%  It is somewhat biased toward smaller ellipses.
centroid = mean(XY);   % the centroid of the data set
D1 = [(XY(:,1)-centroid(1)).^2, (XY(:,1)-centroid(1)).*(XY(:,2)-centroid(2)),...
      (XY(:,2)-centroid(2)).^2];
D2 = [XY(:,1)-centroid(1), XY(:,2)-centroid(2), ones(size(XY,1),1)];
S1 = D1'*D1;
S2 = D1'*D2;
S3 = D2'*D2;
T = -inv(S3)*S2';
M = S1 + S2*T;
M = [M(3,:)./2; -M(2,:); M(1,:)./2];
[evec,eval] = eig(M);
cond = 4*evec(1,:).*evec(3,:)-evec(2,:).^2;
A1 = evec(:,find(cond>0));
A = [A1; T*A1];
A4 = A(4)-2*A(1)*centroid(1)-A(2)*centroid(2);
A5 = A(5)-2*A(3)*centroid(2)-A(2)*centroid(1);
A6 = A(6)+A(1)*centroid(1)^2+A(3)*centroid(2)^2+...
     A(2)*centroid(1)*centroid(2)-A(4)*centroid(1)-A(5)*centroid(2);
A(4) = A4;  A(5) = A5;  A(6) = A6;
A = A/norm(A);
delta = A;
rx = XY(:,1);
ry = XY(:,2);
%color = 'black';
drawellip(delta,rx,ry,color)
end  %  EllipseDirectFit
```

Draw Ellipse Code (used in Least Squares Best Fit Ellipse Code):

```
% draws an ellipse with a(1)x^2 + a(2)xy + a(3)y^2 + a(4)x + a(5)y + a(6) = 0
% overlaid with original points (rx,ry)
function drawellip(delta,rx,ry,color)
%delta = check;
```

```
% rx = plotting2(1:last2,1);
% ry = plotting2(1:last2,2);
A = delta(1);
B = delta(2);
C = delta(3);
D = delta(4);
E = delta(5);
F = delta(6);
Mnot = [F D/2 E/2; D/2 A B/2; E/2 B/2 C];
M = [A B/2; B/2 C];
[Vnot, Dnot] = eig(Mnot);
[V, Dnew ] = eig(M);
if abs(Dnew(1,1)-A) <= abs(Dnew(1,1)-C)
    lambda_1 = Dnew(1,1);
    lambda_2 = Dnew(2,2);
else
    lambda_1 = Dnew(2,2);
    lambda_2 = Dnew(1,1);
end
a = sqrt((-det(Mnot))/(det(M)*lambda_1));
b = sqrt((-det(Mnot))/(det(M)*lambda_2));
h = (B*E - 2*C*D)/(4*A*C - B^2);
k = (B*D - 2*A*E)/(4*A*C - B^2);
tau = (acot((A-C)/B))/2;
t = 0:.01:2*pi;
x = h + cos(tau)*a*cos(t)-sin(tau)*b*sin(t);
y = k + sin(tau)*a*cos(t)+cos(tau)*b*sin(t);
hold on
plot(x,y,'Color',color);
plot(rx,ry,'.','Color','black');
```

PCA Ellipse Code:

```
function[h,k,a,b,tau,x,y,charlie,F_1,F_2,s,area] = PCA_Ellipse(x,y,conf)
% x =plotting(1:last,1);
% y = plotting(1:last,2);
% conf = .75;
% color = 'green';
XY = [x y];
center = mean(XY);
h = center(1);
k = center(2);
% c = convhull(XY); %NEW
% c = c(1:end-1); %NEW
% XY = XY(c,:); %NEW
CovMat = cov(XY);
[eigVec, eigVal] = eig(CovMat);
[eigVec, eigVal] = sortem(eigVec,eigVal);
%conf = .95;
n = length(XY);
p = 2;
c = finv(conf,p,n-p);
a = sqrt(c*eigVal(1,1));
b = sqrt(c*eigVal(2,2));
```

```
t = 0:.01:2*pi;
tau = atan(eigVec(2)/eigVec(1));
x = h + cos(tau)*a*cos(t)-sin(tau)*b*sin(t);
y = k +sin(tau)*a*cos(t)+cos(tau)*b*sin(t);
charlie = sqrt(a^2 - b^2);
F_1 = [h - cos(tau)*charlie,k-sin(tau)*charlie];
F_2 = [h + cos(tau)*charlie, k + sin(tau)*charlie];
s = 2*a;
area = pi*a*b;
% hold on
% plot(XY(:,1),XY(:,2),'.','Color','black');
% plot(x,y,'-','Color','blue');
%plot(h,k,'o','Color','green');
```

Sort Function Code (used in PCA Ellipse Code):

```
function [P2,D2]=sortem(P,D)
% this function takes in two matrices P and D, presumably the output
% from Matlab's eig function, and then sorts the columns of P to
% match the sorted columns of D (going from largest to smallest)
% EXAMPLE:
% D =
%    -90     0     0
%      0   -30     0
%      0     0   -60
% P =
%      1     2     3
%      1     2     3
%      1     2     3
% [P,D]=sortem(P,D)
% P =
%      2     3     1
%      2     3     1
%      2     3     1
% D =
%    -30     0     0
%      0   -60     0
%      0     0   -90
D2=diag(sort(diag(D),'descend')); % make diagonal matrix out of sorted diagonal values of input D
[c, ind]=sort(diag(D),'descend'); % store the indices of which columns the sorted eigenvalues come
    from
P2=P(:,ind); % arrange the columns in this order
```

Minimum Volume Enclosing Ellipsoid Code (2D):

```
function[h,k,a,b,tau,x,y,charlie,F_1,F_2,s,area]= MVE(P, tolerance,color)
% [A , c] = MinVolEllipse(P, tolerance)
% Finds the minimum volume enclsing ellipsoid (MVEE) of a set of data
% points stored in matrix P. The following optimization problem is solved:
% minimize        log(det(A))
% subject to      (P_i - c)' * A * (P_i - c) <= 1
% in variables A and c, where P_i is the i-th column of the matrix P.
% The solver is based on Khachiyan Algorithm, and the final solution
% is different from the optimal value by the pre-spesified amount of 'tolerance'.
```

```
% inputs:
%---------
% P : (d x N) dimnesional matrix containing N points in R^d.
% tolerance : error in the solution with respect to the optimal value.
% outputs:
%---------
% A : (d x d) matrix of the ellipse equation in the 'center form':
% (x-c)' * A * (x-c) = 1
% c : 'd' dimensional vector as the center of the ellipse.
%%%%%%%%%%%%%%%%%%%% Solving the Dual problem%%%%%%%%%%%%%%%%%%%%%%%%%%%%5
% ---------------------------------
% data points
% ---------------------------------
[d N] = size(P);
Q = zeros(d+1,N);
Q(1:d,:) = P(1:d,1:N);
Q(d+1,:) = ones(1,N);
% initializations
% ---------------------------------
count = 1;
err = 1;
u = (1/N) * ones(N,1);          % 1st iteration
% Khachiyan Algorithm
% ---------------------------------
while err > tolerance,
    X = Q * diag(u) * Q';       % X = \sum_i ( u_i * q_i * q_i')  is a (d+1)x(d+1) matrix
    M = diag(Q' * inv(X) * Q);  % M the diagonal vector of an NxN matrix
     [maximum j] = max(M);
     step_size = (maximum - d -1)/((d+1)*(maximum-1));
     new_u = (1 - step_size)*u ;
     new_u(j) = new_u(j) + step_size;
     count = count + 1;
     err = norm(new_u - u);
     u = new_u;
end
%%%%%%%%%%%%%%%%%%%% Computing the Ellipse parameters%%%%%%%%%%%%%%%%%%%%%%
% Finds the ellipse equation in the 'center form':
% (x-c)' * A * (x-c) = 1
% It computes a dxd matrix 'A' and a d dimensional vector 'c' as the center
% of the ellipse.
U = diag(u);
% the A matrix for the ellipse
% ----------------------------------------
A = (1/d) * inv(P * U * P' - (P * u)*(P*u)' );
% center of the ellipse
% ----------------------------------------
c = P * u;
%N = 50;
[U D V] = svd(A);
a = 1/sqrt(D(1,1));
b = 1/sqrt(D(2,2));
t = [0:.01:2*pi];
% Parametric equation of the ellipse
    %---------------------------------------
```

```
%      state(1,:) = a*cos(t);
%      state(2,:) = b*sin(t);
    % Coordinate transform
    %--------------------------------------
    h = c(1);
    k = c(2);
    tau = atan(V(2)/V(1));
%     X = V * state;
%     X(1,:) = X(1,:) + h;
%     X(2,:) = X(2,:) + k;
charlie = sqrt(a^2 - b^2);
F_1 = [h - cos(tau)*charlie,k-sin(tau)*charlie];
F_2 = [h + cos(tau)*charlie, k + sin(tau)*charlie];
s = 2*a;
area = pi*a*b;
x = h + cos(tau)*a*cos(t)-sin(tau)*b*sin(t);
y = k +sin(tau)*a*cos(t)+cos(tau)*b*sin(t);
% plot(X(1,:),X(2,:),'Color','magenta');
       hold on
   plot(x,y,'LineWidth',2,'Color',color);
 %plot(P(1,:),P(2,:),'o','Color','green');
```

I. Ellipse Fitting Algorithm Code

```
penny = 1;
FINALS = cell(20,2);
for tango = 15:16
    for foxtrot = 5:9
        fullFileName = sprintf('%d_%d_IF.mat', foxtrot,tango);
        sFileName = sprintf('%d_%d_ellipse_data_MVE.mat',foxtrot,tango);
        load(fullFileName);
%alpha = .80;
d_val = 16; %20;
time_val = 600;
N = total_flashes; %total number of flashes in date0 matrix (# of rows)
ellipses = zeros(total_flashes,12); %create the output matrix for all the data
current_time = zeros(1,6); %current time of current flash
last_flash_time = zeros(1,6); %time of previous flash
time_elapse = 0; %time between current flash and previous flash
p1 = 0; %x coordinate of current extreme point
p2 = 0; %y coordinate of current extreme point
distance = 0; %distance from edge of last ellipse to current extreme point
dist_away_last_ellipse = 0; %max of all distances = distance new flash is from last ellipse
current_ellipse = zeros(total_srcpts,3); %matrix that holds all the sourcepoints for the current
    ellipse
inside = 0; % srcpts that occur inside current ellipse
outside = 0; % srcpts that occur outside ellipse
all_dist_from_center = 0; %the distance from the center for each extreme point of the current
    flash
dist_from_center = 0; %max distance from the center for the current flash
flashes_in_curr_ell = zeros(N,9); %matrix that keeps track of current flashes in the ellipse
flash_time = zeros(1,6);
t_d = 0;
curr_srcpts = 0;
P = zeros(2,total_srcpts);
tol = .01;
%LIST OF COUNTERS%
m = 0; %total number of flashes used where "i" is the current flash we are looking at
curr_flashes_in_ell = 0; %total number of flashes in current ellipse
extpts = 0; %count of extpts for each flash
    %EXTPT Counters for loop checking distance from last ellipse
count_far_away = 0;
row_count = 1;
    %EXTPT Counters for loop checking distance from last ellipse
ell_count = 1; %count of number of ellipses that have been created
new_srcpts = 0; %count of srcpts for current flash
    %counters for loop creating current_ellipse matrix
     curr_ell_start = 1; %current place to start counting for current_ellipse
     curr_ell_last = 0; %last entry in current_ellipse matrix
     row_counter = 1;
    %counters for loop creating current_ellipse matrix
%all_srcpts = 0; %count of all srcpts in current ellipse
%LIST OF COUNTERS%
date0(:,10) = zeros(total_flashes,1); %changes values of the tracker for if a flash has/has not
    been used all back to unused
```

```
%First while loop keeps everything going until all flashes have been used
%in an ellipse
while m < N
    %hold off
    DD = find(date0(:,10)==0);
    dusk = length(DD);
        for i = 1:dusk %run through all flashes
        whiskey = DD(i);
       % if date0(whiskey,10) == 0 %if flash has not been used, continue
            current_time = date0(whiskey,3:8); %record current time of the current flash
            %if there is already a flash in the ellipse this loop will go
            %through and see what the time difference is
            if curr_flashes_in_ell > 0
                last_flash_time = flashes_in_curr_ell(curr_flashes_in_ell,4:9); %
                    %%%%%%%%%%%%%%%%%%%%%%%%%
                time_elapse = etime(current_time,last_flash_time);
                if time_elapse > 1800 || count_far_away >= 5 %if time difference is greater than
                    30 minutes we reset everything and start over at the beginning of the flash
                    list
                    %INCLUDE UPDATE TO COUNTERS HERE
                    flashes_in_curr_ell = zeros(N,9);
                    dist_away_last_ellipse = 0;
                    current_ellipse = zeros(total_srcpts,3);
                    curr_ell_start = 1;
                    curr_ell_last = 0;
                    curr_flashes_in_ell = 0;
                    date0(whiskey,10) = 0;
                    ell_count = ell_count + 1;
                    break
                end
            extpts = date0(whiskey,9); %finds the number of extreme points for the flash (stored
                in date0 col #9)
                        dist = zeros(extpts,2); %creates a matrix to record the distances from the
                            edge of the ellipse and from the center for each extreme point in the
                            flash                for n = 1:extpts
                p1 = flash_x_extremes(n,whiskey); %x coord of current extreme point
                p2 = flash_y_extremes(n,whiskey);%y coord of current extreme point
                    %if statment checks if point is inside or outside
                    %the ellipse, if it is inside, we can disregard it, if
                    %outside we calculate distance and record in dist
                    %matrix
                    if (sqrt((p1-F_1(1,1))^2 + (p2 - F_1(1,2))^2) + sqrt((p1-F_2(1,1))^2 + (p2-F_2
                        (1,2))^2) > s)
                        d_t = sqrt((p1-x).^2 + (p2 - y).^2);
                        distance = min(d_t);
                        dist(n,1) = distance;
                    end
                   %find the distance from the center for each extreme point and record it in col
                        #2 of the dist matrix
                    all_dist_from_center = sqrt((p1-h)^2 + (p2-k)^2);
                    dist(n,2) = all_dist_from_center;
                end
                row_count = 1; %resets counter for dist matrix
                dist_away_last_ellipse = max(dist(:,1)); %determine the distance away from current
```

```
      ellipse
dist_from_center = max(dist(:,2)); %determine the distance away from the center of
      the current ellipse
ellipses(whiskey,4:6) = [dist_away_last_ellipse;dist_away_last_ellipse;
     dist_from_center]; %record in ellipses matrix            end
if dist_away_last_ellipse < d_val %16 KM = apprx. 9.94 Miles
    count_far_away = 0;
    %ELLIPSE CALCULATIONS HERE
    new_srcpts = date0(whiskey,9); %find the number of new sourcepoints (extreme
         points now) for the current flash
   curr_flashes_in_ell = nnz(flashes_in_curr_ell(:,1)); %determine number of
        flashes currently in the ellipse
    %this loop goes through and finds the first flash in
    %the ellipse that is at least 10 minutes old and
    %removes that flash from both matrices
    if curr_flashes_in_ell > 0
        for w = curr_flashes_in_ell:-1:1
           flash_time = flashes_in_curr_ell(w,4:9);
           t_d = etime(current_time,flash_time);
           if t_d > time_val
               rmflash = flashes_in_curr_ell(w,1);
               flashes_in_curr_ell(1:w,:) = [];
               last_in_list = find(current_ellipse(:,3) == rmflash,1,'last');
               current_ellipse(1:last_in_list,:) = [];
               break
           end
        end
    end
    curr_flashes_in_ell = nnz(flashes_in_curr_ell(:,1));
    curr_srcpts = nnz(current_ellipse(:,1));
    curr_ell_start = curr_srcpts + 1; %update starting point in current_ellipse
         matrix
    curr_ell_last = curr_ell_start + new_srcpts - 1;
    new_flash = curr_flashes_in_ell + 1;
    flashes_in_curr_ell(new_flash,1) = whiskey;
    flashes_in_curr_ell(new_flash,2) = date0(whiskey,2);
    flashes_in_curr_ell(new_flash,3) = new_srcpts;
    flashes_in_curr_ell(new_flash,4:9) = date0(whiskey,3:8);
    %%%this goes through and adds all the relative sourcepoints to the
    %%%current_ellipse matrix to be used for creating the current ellipse
    current_ellipse(curr_ell_start:curr_ell_last,1) = flash_x_extremes(1:
         new_srcpts,whiskey);
    current_ellipse(curr_ell_start:curr_ell_last,2) = flash_y_extremes(1:
         new_srcpts,whiskey);
    current_ellipse(curr_ell_start:curr_ell_last,3) = whiskey;
    row_counter = 1; %reset row_counter for previous loop
    cell_count = nnz(current_ellipse(:,1)); %finds the number of nonzero rows in
         current_ellipse matrix which is a recording of the number of ext srcpts in
          the ellipse
    P = [current_ellipse(1:cell_count,1),current_ellipse(1:cell_count,2)]';
    if (dist_away_last_ellipse > 0 && curr_flashes_in_ell >0) ||
         curr_flashes_in_ell == 0
    KK = convhull(P'); %convexhull pts for faster MVE algorithm
    KK = KK(1:end-1,1);%find only unique entries for convexhull
```

```
                    %KK = unique(KK(:));
                    Q = P(:,KK);%Q = P(KK,:); %select those entries as what the ellipse will be
                        bounded around
                %   plot(x,y,'Color','white');
                    [h,k,a,b,tau,x,y,charlie,F_1,F_2,s,area] = MVE(Q,tol); %PCA_Ellipse(Q(:,1),Q
                        (:,2),alpha); %  creates a Minimum area ellipse based on the
                        current_ellipse matrix
                    %%%%%%%%%%%%%%PLOT ELLIPSE & POINTS%%%%%%%%%%%%%%%%%
                   % plot(Q(1,:),Q(2,:),'o','Color','black');
                   % hold on
                   % %plot(Q(:,1),Q(:,2),'o','Color','black');
                   % plot(x,y,'Color','blue');
                    %%%%%%%%%%%%%%PLOT ELLIPSE & POINTS%%%%%%%%%%%%%%%%%
                    %ELLIPSE CALCULATIONS HERE
                    %CALCULATE NUMBER OF POINTS INSIDE ELLIPSE
%                       for p = 1:cell_count
%                           if (sqrt((current_ellipse(p,1)-F_1(1,1))^2 + (current_ellipse(p,2)-F_1
    (1,2))^2)+sqrt((current_ellipse(p,1)-F_2(1,1))^2 + (current_ellipse(p,2)-F_2(1,2))^2)<= s)
%                           inside = inside + 1;
%                           else
%                           outside = outside + 1;
%                           end
%                       end
%                       total = inside + outside; %records total of all points...should match
    cell_count
%                       ellipses(whiskey,3) = inside/total; %records the percentage of points inside
     the current ellipse
%                       inside = 0; %resets counter
%                       outside = 0; %resets counter
                    %CALCULATE NUMBER OF POINTS INSIDE ELLIPSE
                    end
                   % plot(P(1,:),P(2,:),'.','Color','magenta');
                    ellipses(whiskey,2)=area; %records ellipse area
                    ellipses(whiskey,8:12) = [a;b;h;k;tau]; %records relative information for
                        newly created ellipse
                    m = m+1; %update the number of points that have been used
                    date0(whiskey,10) = ell_count; %record the ellipse number for tracking
                    ellipses(whiskey,1) = ell_count; %record the ellipse number for data
                    ellipses(whiskey,7) = new_flash; %record the number of flashes in current
                        ellipse
                   curr_flashes_in_ell = new_flash;
                else
                    count_far_away = count_far_away + 1;
                    new_srcpts = date0(whiskey,9);
               % plot(flash_x_extremes(1:new_srcpts,whiskey),flash_y_extremes(1:new_srcpts,whiskey
                    ),'o','Color', 'red');
                    date0(whiskey,10)=0;
                    ellipses(whiskey,4) = 0; %set distance from last ellipse to zero
                    ellipses(whiskey,6) = 0; %set distance from center to zero
                    %INCLUDE UPDATE TO COUNTERS HERE
                end
            %end
            %%%this loop tells us if the distance away from the last ellipse is
            %%%greater than 16 and we are on the final flash, then we have to
```

```
        %%%reset and start back at the beginning of the for loop
        if dist_away_last_ellipse >= d_val && whiskey == N
            flashes_in_curr_ell = zeros(N,9);
            dist_away_last_ellipse = 0;
            current_ellipse = zeros(total_srcpts,3);
            curr_ell_start = 1;
            curr_ell_last = 0;
            curr_flashes_in_ell = 0;
            date0(whiskey,10) = 0;
            ell_count = ell_count + 1;
        end
    end

end
 FINALS{penny,1} = ellipses;
 FINALS{penny,2} = date0;
save(sFileName,'ellipses','date0');
clearvars -except 'FINALS' 'penny' 'tango' 'foxtrot'
penny = penny + 1;
    end

end
save('Finals.mat', 'FINALS');
```

J. Empirical Validation Code

```
%loads the points of each warning circle 2-10 are the ones we are going to
%use
All_Done = cell(20,1); %final storage place for all the data
%load('C:\Users\Admin\Documents\Thesis\Daily Updates\19 Dec 18\warning_circles.mat')
penny = 1; %counter for final storage
for tango = 13:16 %loop through all years
    for foxtrot = 5:9 %loop through all months
        fullFileName = sprintf('%d_%d_IF.mat', foxtrot,tango);
         load(fullFileName);
last = length(date0);
extra = zeros(last,9); %add extra columns for the distances from all the lightning warning circles
date0 = [date0 extra];
%%this code goes through and finds the distance each source point is from
%%the center of each warning circle
col_count = 11;
for i = 2:11
    x1 = points(i,1);
    y1 = points(i,2);
    for p = 1:last
        ext = date0(p,9);
        distances = zeros(ext,1);
        for m = 1:ext
            x2 = flash_x_extremes(m,p);
            y2 = flash_y_extremes(m,p);
            dist = sqrt(((x2-x1)^2)+((y2-y1)^2));
            distances(m,1) = dist;
        end
        smallest = min(distances(:,1));
        date0(p,col_count) = smallest;
    end
    col_count = col_count + 1;
end
%this code goes through and pulls out the flashes that are within 5 or 6
%miles respectively of each site (of course using km here)
Ultimates = cell(10,4);
for i = 2:11
    mileage = points(i,3); %record of the radius of each warning circle
    col_count = i+9;
    winners = find(date0(:,col_count)<=mileage);
    dunzo = date0(winners,:);
    entry = i-1;
    Ultimates{entry,1} = dunzo;
end
first_break = 0;
rcheck = 7.408;
for i = 1:10
%     if points(i+1,3) == 11.1120
%         rcheck = 9.26;
%     else
%         rcheck = 7.408;
%     end
```

```
%rcheck = points(i+1,3);
 inside_circle =  Ultimates{i,1};
 col_count = i + 10;
 last = length(inside_circle);
 extra = zeros(last,2);
 condensed = [inside_circle(:,3:8) inside_circle(:,col_count) extra];
 %%%%%%%%%Sets up the data for the time portion
 for m = 2:last
    prev_flash = condensed(m-1,1:6);
    curr_flash = condensed(m,1:6);
    td = abs(etime(curr_flash,prev_flash));
    condensed(m,8) = td;
 end
 time_breaks = find(condensed(:,8)>=1800)-1;
 time_intervals = zeros(length(time_breaks)+1,9);
 time_intervals(1,1) = 1;
 time_intervals(1:size(time_intervals,1)-1,2) = time_breaks(:,1);
 for w = 2:size(time_intervals,1)
     time_intervals(w,1) = time_intervals(w-1,2) + 1;
 end
  time_intervals(size(time_intervals,1),2) = last;
  %%%%%%%%%
  for delta = 1:size(time_intervals,1)
      start = time_intervals(delta,1);
      finish = time_intervals(delta,2);
      for t = start:finish
          if condensed(t,7) <= rcheck && condensed(t,7) > .926 %7.408
              status = 1;
          elseif condensed(t,7) <=.926
              status = 2;
          else
              status = 5;
          end
          condensed(t,9) = status;
      end
      cessation = etime(condensed(finish,1:6),condensed(start,1:6)) + 1800;
      check_mat = condensed(start:finish,:);
      first_four = find(check_mat(:,9)==1,1,'first');
      if isempty(first_four) == 1
          first_four = 0;
      end
      time_intervals(delta,3) = first_four;
      if first_four > 1
      time_gap = etime(check_mat(first_four,1:6),check_mat(1,1:6));
      time_intervals(delta,4) = time_gap;
      else
          time_gap = 0;
          time_intervals(delta,4) = time_gap;
      end
      first_break = find(check_mat(:,9)==2,1,'first');
      if isempty(first_break) == 1
          first_break = 0;
      end
      if first_break > 1
```

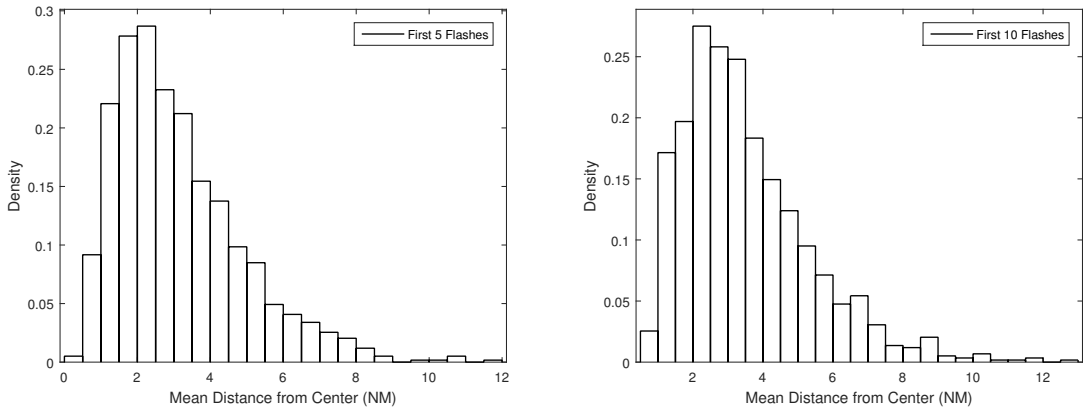
```
      time_break = etime(check_mat(first_break,1:6),check_mat(1,1:6));
      time_intervals(delta,5) = time_break;
      else
          time_break = 0;
          time_intervals(delta,5) = time_break;
      end
      if first_break < first_four && first_break ~= 0
          time_intervals(delta,6) = 1;
      end
      total_breaks = sum(check_mat(:,9)==2);
      time_intervals(delta,7) = first_break;
      time_intervals(delta,8) = total_breaks;
      time_intervals(delta,9) = cessation;

  end
Ultimates{i,2} = condensed;
Ultimates{i,3} = time_breaks;
Ultimates{i,4} = time_intervals;
end
All_Done{penny,1} = Ultimates;
penny = penny + 1;
clearvars -except 'All_Done' 'points' 'tango' 'foxtrot' 'penny'
    end
end

%analyze All_Done
%go through and analyze for each month first
data_by_month = zeros(20,7);
circ_data = cell(20,1);
%bigcirc = [3 4 5 8];
for i = 1:20
  select_1 = All_Done{i,1};
  data_by_circ = zeros(4,7);
  for p = 1:10
      %p = bigcirc(hotshot);
     time_saved = 0;
     time_before_bust = 0;
     num_busts = 0;
     no_warnings = 0;
     time_saved_no_warnings = 0;
     checking = select_1{p,4};
     last = size(checking,1);
     for m = 1:last
         if checking(m,3) == 0 && checking(m,8)== 0;
             time_saved = time_saved + checking(m,9);
         elseif checking(m,3) > 0 && checking(m,6) == 0
             time_saved = time_saved + checking(m,4);
         end
         if checking(m,6) == 1 && checking(m,7) ~= 1
         num_busts = num_busts + 1;
         end
         no_busts = sum(checking(:,8)==0);
        if checking(m,4) > 0 && checking(m,5) > 0 && checking(m,6) == 0
            time_before_bust = time_before_bust + checking(m,5) - checking(m,4);
```

```
        end
        if checking(m,3) == 0 && checking(m,7) == 0
            no_warnings = no_warnings + 1;
            time_saved_no_warnings = time_saved_no_warnings + checking(m,9);
        end
      end
      data_by_circ(p,1) = time_saved;
      data_by_circ(p,2) = num_busts;
      data_by_circ(p,3) = no_busts;
      data_by_circ(p,4) = time_before_bust;
      data_by_circ(p,5) = last;
      data_by_circ(p,6) = no_warnings;
      data_by_circ(p,7) = time_saved_no_warnings;
  end
  circ_data{i,1} = data_by_circ;
  data_by_month(i,1) = sum(data_by_circ(:,1));
  data_by_month(i,2) = sum(data_by_circ(:,2));
  data_by_month(i,3) = sum(data_by_circ(:,3));
  data_by_month(i,4) = sum(data_by_circ(:,4));
  data_by_month(i,5) = sum(data_by_circ(:,5));
  data_by_month(i,6) = sum(data_by_circ(:,6));
  data_by_month(i,7) = sum(data_by_circ(:,7));
end
failures = zeros(sum(data_by_month(:,2)),12);
rc = 1;
for i = 1:20
    select_1 = All_Done{i,1};
    for p = 1:10
        checking = select_1{p,4};
        last = size(checking,1);
        for m = 1:last
         if checking(m,6) == 1 && checking(m,7) ~= 1
         failures(rc,:) = [i p m checking(m,:)];
         rc = rc + 1;
         end
        end
    end
end
circle_results = zeros(10,7);
run = zeros(1,7);
for p = 1:10
for i = 1:20
   select = circ_data{i,1};
   run(1,:) = run(1,:)+ select(p,:);
end
circle_results(p,:) = run;
run = zeros(1,7);
end
```

K. Descriptive Statistics for the Mean Distance from the Center of Initial Flashes in a Lightning Storm



(a) Mean Values for First 5 Flashes **(b) Mean Values for First 10 Flashes**

Figure 26: Histograms for the Mean Distance from the Center (NM) of the Initial Flashes in a Lightning Storm

Table 17: Quantile Values for the Mean Distance from the Center (NM) of the Initial Flashes in a Lightning Storm

Quantile	Mean Distance (NM) First 5 Flashes	Mean Distance (NM) First 10 Flashes
Minimum	0.375	0.633
25th Quartile	1.807	2.192
Median	2.774	3.154
75th Quartile	4.066	4.484
97.5 Percentile	7.469	8.167
Maximum	11.89	12.963

Table 18: Summary Statistics for the Mean Distance from the Center (NM) of the Initial Flashes in a Lightning Storm

Summary Statistics	Mean Distance (NM) First 5 Flashes	Mean Distance (NM) First 10 Flashes
Mean	3.125	3.526
Standard Deviation	1.735	1.837
Standard Error	0.051	0.054
Upper 95% CI	3.224	3.631
Lower 95% CI	3.026	3.421
Sample Count	1,178	1,178

L. Weibull Distribution for Distance from Edge of Ellipse

Shape Estimate and 95% Confidence Interval:

$$k = .833 \quad (.827, .839)$$

Scale Estimate and 95% Confidence Interval:

$$\lambda = 2.124 \quad (2.093, 2.170)$$

Mean Estimate and 95% Confidence Interval:

$$\mu = 2.36 \quad (2.334, 2.388)$$

PDF:

$$f(x; k, \lambda) = \frac{k}{\lambda}\big(\frac{x}{\lambda}\big)^{k-1} e^{-(\frac{x}{\lambda})^k}$$

$$f(x; .833, 2.124) = .392\big(\frac{x}{2.124}\big)^{-.167} e^{-(\frac{x}{2.124})^{.833}}$$

CDF:

$$F(x; k, \lambda) = 1 - e^{-(\frac{x}{\lambda})^k}$$

$$F(x; .833, 2.124) = 1 - e^{-(\frac{x}{2.124})^{.833}}$$

Table 19: Cumulative Probability of Strike Using Weibull Distribution

Distance (NM)	Probability of Strike
16	0.013%
15	0.020%
14	0.033%
13	0.053%
12	0.086%
11	0.140%
10	0.231%
9	0.385%
8	0.648%
7	1.10%
6	1.89%
5	3.31%
4	5.90%
3	10.79%
2	20.42%
1	40.98%

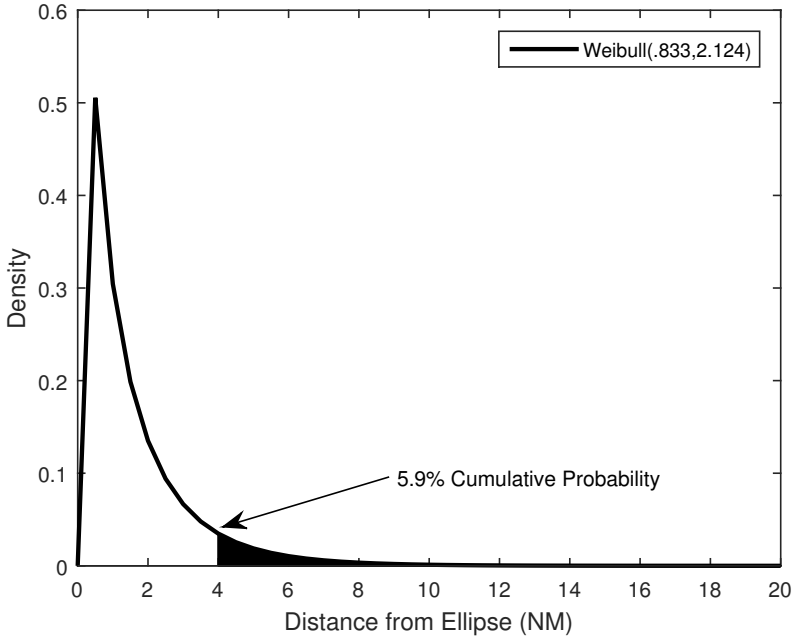


Figure 27: Example of Cumulative Probability of Strike Using Weibull Distribution

Bibliography

Bothwell, Phillip D., and Lindsey M. Richardson. "Forecasting Lightning Using a Perfect Prog Technique Applied to Multiple Operational Models," *Proceedings of the 25th International Conference on Atmospheric Electricity.* 1-17. Norman OK, 2014.

Britt, Thomas O., Carl L. Lennon and Launa M. Maier. "Lightning Detection and Ranging System," T*echbriefs.com.* NASA Tech Briefs Archive. 1 April 1998. Retrieved on 5 August 2018.

Calhoun, Kristin M., T.C. Meyer, K. Berry, H. Obermeier, S.J. Sanders, C.A. Shivers, C.D. Karstens, J.P. Wolfe, and K.E. Klockow. "Cloud-to-Ground Lightning Probabilities and Warnings within an Integrated Warning Team," *Proceedings of the 98th Annual Meeting of the American Meteorological Society.* Austin TX, 2018.

Cox, C.C. *A Comparison of Horizontal Cloud-to-Ground Lightning Flash Distance Using Weather Surveillance Radar and the Distance Between Successive Flashes Method.* MS Thesis, AFIT/GM/ENP/99M-03, Department of Engineering Physics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 1999.

Department of the Air Force. *Safety: Air Force Consolidated Occupational Safety Instruction.* AFI 91-203. Washington: HQ USAF, 15 June 2012.

Dillon, William R. and Matthew Goldstein. *Multivariate Analysis: Methods and Applications.* New York: Jon Wiley & Sons Inc., 1984.

Dwyer, Joseph R. and Martin A. Uman. "The Physics of Lightning," *Physics Reports Journal,* 534: 147-241 (2014).

"Engineering Statistics Handbook," *Nist.gov.* NIST/SEMATECH e-Handbook of Statistical Methods. April, 2012. Retrieved on 3 August 2018.

Fitzgibbon, Andrew W., Maurizio Pilu and Robert B. Fisher. "Direct Least Squares Fitting of Ellipses," *Proceedings of the 13th International Conference on Pattern Recognition.* 1-5. Vienna, Austria: IEEE Press, 1996.

Halíř, R. and J. Flusser. "Numerically Stable Direct Least Squares Fitting of Ellipses," *Proceedings of the 6th International Conference in Central Europe on Computer Graphics and Visualization.* 125-132. Plzen, Czech Republic: WSCG, 1998.

Holle, Ronald L., Raúl E. López, and Christoph Zimmermann. "Updated Recommendation for Lightning Safety-1998," *Bulletin of the American Meteorological Society,* 80: 2035-2041 (October 1999).

Kalair, A., N. Abas, and N. Khan. “Lightning Interactions with Humans and Lifelines,” *Journal of Lightning Research*, 5: 11-28 (2013).

Kotz, Samuel and Saralees Nadarajah. *Extreme Value Distributions: Theory and Applications.* London: Imperial College Press, 2000.

López, Raúl E. and Ronald Holle. *The Distance Between Successive Lightning Flashes.* NOAA Technical Memorandum ERL NSSL-105; Norman OK: National Severe Storms Laboratory, 1999.

Mata, C.T. and J.G. Wilson. “Future Expansion of the Lightning Surveillance System at the Kennedy Space Center and the Cape Canaveral Air Force Station, Florida, USA,” *Proceedings of the 31st International Conference on Lightning Protection (ICLP).* 1-3. Vienna, Austria: IEEE Press, 2012.

McNamara, Todd M. *The Horizontal Extent of Cloud-to-Ground Lightning over the Kennedy Space Center*. MS Thesis, AFIT/GM/ENP/02M-06. Department of Engineering Physics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2002.

NASA Facts. “Lightning and the Space Program.” *NASA.gov*. 2006. Retrieved on 28 August 2018.

National Deep Submergence Facility (NDSF). *Coordinate Conversion Utility*. Woods Hole Oceanographic Institution. Retrieved on 19 December 2018.

Parsons, Tamara L. *Determining the Horizontal Distance Distribution of Cloud-to-Ground Lightning.* MS Thesis, AFIT/GM/ENP/00M-09. Department of Engineering Physics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2000.

Renner, Steve L. *Analyzing Horizontal Distances Between WSR-88D Thunderstorm Centroids and Cloud-to-Ground Lightning Strikes*. MS Thesis, AFIT/GM/ENP/98M-09. Department of Engineering Physics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 1998.

Roeder, William P. “Point Paper on 5 NM Lightning Warnings.” Unpublished Point Paper. 45th Weather Squadron, Cape Canaveral Air Force Station, Cape Canaveral FL, 2008.

Roeder, William P. “The Four Dimensional Lightning Surveillance System,” *Proceedings of the 21st International Lightning Detection Conference.* 1-15. Orlando FL: 2010.

Roeder, William P., Todd M. McNamara, Mike McAleenan, Katherine A. Winters, Launa M. Maier, and Lisa L. Huddleston. “The 2014 Upgrade to the Lightning Warning Circles Used by 45th Weather Squadron,” *Proceedings of the 18th Conference on Aviation, Range, and Aerospace Meteorology.* 1-13. Seattle WA: AMS, 2017.

Roeder, William P. and Clark S. Pinder. “Lightning Forecasting Empirical Techniques for Central Florida in Support of America’s Space Program,” *Proceedings of the 16th Conference on Weather Analysis and Forecasting,* 11-16. Phoenix AZ: AMS, 1998.

Shivalli, Sanketa. “Lightning Phenomenon, Effects and Protection of Structures from Lightning,” *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE),* 11: 2320-3331 (May-June 2016).

Starr, Stan, David Sharp, Francis Merceret, and Martin Murphy. *LDAR, a Three-Dimensional Lightning Warning System: Its Development and Use by the Government, and Transition to Public Availability.* NASA Technical Report. Cocoa Beach FL: NASA Kennedy Space Center, 1998.

Todd, Michael J. and E. Alper Yildirim. “On Khachiyan’s Algorithm for the Computation of Minimum-Volume Enclosing Ellipsoids,” *Discrete Applied Mathematics,* 155: 1731-1744 (April 2007).

Van Loan, Charles F. “Using the Ellipse to Fit and Enclose Data Points,” *Department of Computer Science*, New York: Cornell University, 2006.

Weatherdata, Inc. “Lightning Prediction System.” *Weatherdata.com.* Retrieved on 28 August 2018.

Wijewickrema, Sudanthi N.R. and Andrew P. Papliński. “Principal Component Analysis for the Approximation of a Fruit as an Ellipse,” Australia: Monash University, 2004.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
21–03–2019	Master's Thesis	Oct 2017 — Mar 2019

4. TITLE AND SUBTITLE

Modeling the Distribution of Lightning Strike Distances Outside a Preexisting Lightning Area

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Sanderson, Dawn L., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENC-MS-19-M-003

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Mr. William Roeder
45 WS/SYS
1201 Edward H. White II St. MS 7302
Patrick AFB, FL 32925-3238
DSN 467-8410
Email: william.roeder@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

45 WS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The primary objective of this study is to investigate if the 5 NM safety radius for lightning warnings can be reduced while maintaining a desired level of safety. The research uses processed Lightning Detection and Ranging (LDAR) data to map the movement of preexisting lightning storms using ellipses which are updated with every lightning flash. A systematic recording ensues for the distance from the ellipse boundary of each flash occurring outside the ellipse. All of those exterior flash distances are then used to find the best-fit distribution from which the stand-off distance for the desired level of safety can be calculated. The distances from the edge of the ellipse are fit to a Weibull distribution and a new warning distance of 4 NM is selected as the most appropriate distance to balance safety and increase productivity. The 4 NM radius is tested with a resulting failure rate of .277%, with a savings of 22.5 8-hour man days a year for the months of May through September.

15. SUBJECT TERMS

Lightning Strike, Extreme Value Distribution, Lightning Detection and Ranging (LDAR), Safety, Ellipses

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Edward White, AFIT/ENC
U	U	U	UU	130	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636 x4540; edward.white@afit.edu