NASA Contractor Report NASA/CR-2005-212564

Objective Lightning Probability Forecasting for Kennedy Space Center and Cape Canaveral Air Force Station

Prepared by:
Winifred Lambert
Mark Wheeler
Applied Meteorology Unit

Prepared for:
Kennedy Space Center
Under Contract NAS10-01052

NASA
National Aeronautics and
Space Administration

Office of Management

Scientific and Technical
Information Program
2005

THIS PAGE INTENTIONALLY BLANK

Executive Summary

The 45th Weather Squadron (45 WS) forecasters include a probability of lightning occurrence in their daily 24-Hour and Weekly Planning Forecasts, which are briefed at 1100 UTC (0700 EDT). This information is used for general scheduling of operations at Cape Canaveral Air Force Station (CCAFS) and Kennedy Space Center (KSC). Forecasters at the Spaceflight Meteorology Group (SMG) also make thunderstorm forecasts during Shuttle flight operations. Much of the current lightning probability forecast at both groups is based on a subjective analysis of model and observational data. The objective tool currently available is the Neumann-Pfeffer Thunderstorm Index (NPTI, Neumann 1971), developed specifically for the KSC/CCAFS area over 30 years ago. However, recent studies have shown that 1-day persistence provides a better forecast than the NPTI. These issues indicated that the NPTI needed to be upgraded or replaced. Because forecasters require a tool that provides a reliable estimate of the daily thunderstorm probability forecast, they requested that a new lightning probability forecast tool be developed. In response, the AMU developed a set of statistical lightning forecast equations that provide a probability of lightning occurrence at KSC/CCAFS for the current day during the warm season (May–September).

The equation development incorporated results from two research projects that investigated causes of lightning occurrence near KSC/CCAFS and the Florida peninsula. One proved that logistic regression outperformed the linear regression method used in NPTI, even when the same predictors were used. This finding influenced the decision to use logistic regression in this AMU task. The other study found relationships between large scale flow regimes and spatial lightning distributions over Florida. As a result, lightning probabilities based on these flow regimes were used as candidate predictors of lightning occurrence for the equation development in this task.

Fifteen years (1989–2003) of warm season data were used to develop the forecast equations. The data sources included the Cloud-to-Ground Lightning Surveillance System (CGLSS), 1200 UTC Florida synoptic soundings, and the 1000 UTC CCAFS sounding. Data from CGLSS, a local network of cloud-to-ground lightning sensors, were used to determine lightning occurrence for each day. The 1200 UTC Florida soundings were used to calculate the synoptic-scale flow regimes and the 1000 UTC CCAFS soundings were used to calculate local stability parameters. Each of the three datasets was processed and analyzed to create the predictand, the element to be predicted, and candidate predictors needed for the statistical forecast equation development. The CGLSS data were used to create a binary predictand for lightning, where 1 denoted that lightning occurred during the day and 0 denoted that lightning did not occur. The flow regimes and local stability parameters from the sounding datasets were used to calculate the candidate predictors of lightning occurrence. In all, 13 candidate predictors were available for equation development.

The data were stratified into two sub-sets: a development dataset consisting of 13 warm seasons from which the equations were developed, and an independent verification dataset of two warm seasons on which the equations were tested. One equation was developed for each warm season month using an iterative manual technique in which each predictor was tested to determine its ability to explain the variance in the predictand individually and in combination with other predictors. The resulting equations contained five or six predictors each. The daily lightning climatology, persistence, and flow regime lightning probability were common to all five monthly equations. The 800–600 mb layer mean relative humidity was a predictor in four of the five equations. Other predictors included the Lifted Index, K-Index, Total Totals, and Thompson Index.

Four equation performance tests were conducted. The results indicated that the equations showed an increase in skill over several standard forecasting methods, good reliability, an ability to distinguish between non-lightning and lightning days, and improved accuracy measures and skill scores over those for 1-day persistence, which as stated previously was shown to outperform the NPTI. Three of the tests, however, showed a tendency for the equations to over-forecast the probability of lightning occurrence, i.e. high probability values were calculated when no lightning was observed. However, given the overall improved skill over current standard forecast methods, the 45 WS requested that the equations be transitioned to operations and added to the current set of tools used to determine the daily lightning probability of occurrence.

A graphical user interface (GUI) was created to facilitate forecaster access to the equations through user-friendly input and fast, easy-to-read output of the lightning probability for the day. Personnel from the 45 WS were involved in the GUI development by providing comments and suggestions on the design to ensure that the final product addressed their operational needs. The probabilities output by the GUI are meant to be used as first-guess guidance when developing the lightning probability forecast for the day. These probabilities provide an objective base from which forecasters can use other observations, model data, consultation with other forecasters, and their own experience to create the final daily lightning probability for the 1100 UTC briefing.

Table of Contents

Executive Summary iii

List of Figures vi

List of Tables vii

1. Introduction 1

1.1. Previous Work 1

1.1.1. Neumann-Pfeffer Thunderstorm Index 1

1.1.2. NPTI Improvements 1

1.1.3. Daily Flow Regimes 2

1.2 Developing the Objective Lightning Probability Forecast Tool 2

2. Data 3

2.1. Cloud-to-Ground Lightning Surveillance System (CGLSS) 3

2.2. Florida 1200 UTC Rawinsondes 4

2.3. CCAFS 1000 UTC Rawinsonde 5

3. Preparation of Equation Elements 7

3.1. Binary Predictand 7

3.2. Candidate Predictors 7

3.2.1. CGLSS Predictors 7

3.2.2. Flow Regime Predictors 8

3.2.3. Stability Index Predictors 14

3.2.4. Summary of Candidate Predictors 15

4. Equation Development and Testing 16

4.1. Data Availability 16

4.1.1. Missing Data 16

4.1.2. Development and Verification Datasets 17

4.2. Equation Development 18

4.2.1. Logistic Regression 18

4.2.2. Residual Deviance Calculation 19

4.2.3. Predictor Selection 20

4.3. Equation Performance 23

4.3.1. Brier Skill Score 23

4.3.2. Probability Distributions 24

4.3.3. Reliability Diagram 25

4.3.4. Contingency Table Statistics 26

4.3.5. Equation Performance Summary 28

5. Graphical User Interface 29
5.1. Excel Workbook 29
5.2. Current Date Dialog Box 29
5.3. Equation Predictor Dialog Boxes 30
5.3.1. Persistence and Flow Regime 30
5.3.2. Sounding Parameters 30
5.4. Equation Output Dialog Box 34
5.5. Predictor Responses 35
5.5.1. May 35
5.5.2. June 36
5.5.3. July 37
5.5.4. August 38
5.5.5. September 39
6. Summary and Conclusions 40
6.1. Equation Performance Review 40
6.2. Graphical User Interface Issues 41
6.3. Future Work 42
References 43
List of Acronyms 45

List of Figures

Figure 1. The locations of the six CGLSS sensors are indicated by the blue circles. 3

Figure 2. The 5 n mi lightning warning circles on KSC/CCAFS and Astrotech used to determine the spatial area for the lightning occurrence prediction and verification. 4

Figure 3. The red dots on the map show the locations of all soundings used in this task. 5

Figure 4. (a) The daily raw (light blue) and smoothed (dark blue) climatological probability values of lightning occurrence for the warm-season months in the POR 1989–2003, and (b) The Gaussian weight values used in the 15-day smoothing equation. 8

Figure 5. Bar charts showing the number of days each flow regime was observed in (a) May, (b) June, (c) July, (d) August, (e) September, and (f) all warm season months in the period of record 1989–2003. 11

Figure 6. Illustration of linear (dashed line) vs. logistic (solid curve) regression probability forecasting for a binary predictand and one predictor. 19

Figure 7. The total percent reduction in residual deviance from that of the NULL model as each predictor was added to the equation using the July development dataset. 21

Figure 8. The forecast probability distributions for lightning (pink) and non-lightning (blue) days in the verification data. 24

Figure 9. The reliability diagram of the probability forecasts for all months. 25

Figure 10. Graph showing the values in the four contingency table cells in Table 8 for the range of probability values 0.1–0.9 in increments of 0.01. 27

Figure 11. The first dialog box in the GUI queries the user for the Month and Day values. Month and Day are chosen by clicking on the down arrows next to each and choosing from the drop-down lists. The Cancel button exits from the GUI, the Continue button brings up the next dialog box. 30

Figure 12. This dialog box contains choices for the predictors in the May equation. 31

Figure 13. Same as Figure 12 except for June, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with sounding parameters LI and RH. 31

Figure 14. Same as Figure 12 except for July, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with sounding parameters TT and RH. 32

Figure 15. Same as Figure 12 except for August, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with sounding parameters KI, TT, and RH. 32

Figure 16. Same as Figure 12 except for September, and with sounding parameters LI and RH. 33

Figure 17. The equation output dialog box in the GUI containing the probability of lighting for the day based on the inputs from the date and equation predictor dialog boxes. 34

Figure 18. Equation response charts for May 15. 35

Figure 19. Equation response charts for June 15. 36

Figure 20. Equation response charts for July 15. 37

Figure 21. Equation response charts for August 15. 38

Figure 22. Equation response charts for September 15. 39

List of Tables

Table 1. List of the flow regime names used in this study and the corresponding sectors showing the average 1000–700 mb wind directions at each of the stations. 10

Table 2. Example of the tables containing the lightning probabilities based on flow regime....12

Table 3. Monthly probabilities of lightning occurrence based on the flow regimes that were used as candidate predictors. 13

Table 4. Summary of missing and available data in the POR....17

Table 5. Summary of missing and available data for equation development and testing. 18

Table 6. The final predictors for each monthly equation, in order of their contribution to the reduction in residual deviance. 22

Table 7. The percent improvement in skill of the logistic regression equation forecasts over the reference forecasts of persistence, daily and monthly lightning climatologies, and flow regime probabilities. 23

Table 8. Basic contingency table for calculating categorical accuracy measures and skill scores (Wilks 1995). The equations for the accuracy measures and skill scores are in the bottom row. 26

Table 9. Contingency table for the cutoff probability value of 0.61. 28

Table 10. Summary values for each of the predictors in the POR 1989–2003. 34

THIS PAGE INTENTIONALLY BLANK

1. Introduction

The 45th Weather Squadron (45 WS) forecasters include a probability of lightning occurrence in their daily 24-Hour and Weekly Planning Forecasts, which are briefed in the morning at 1100 UTC (0700 EDT). This information is used for general scheduling of operations at Cape Canaveral Air Force Station (CCAFS) and Kennedy Space Center (KSC). Forecasters at the Spaceflight Meteorology Group (SMG) also make thunderstorm forecasts during Shuttle operations. Much of the current lightning probability forecast at both groups is based on a subjective analysis of model and observational data. The objective tool currently available operationally is the Neumann-Pfeffer Thunderstorm Index (NPTI), developed specifically for the KSC/CCAFS area over 30 years ago (Neumann 1971; Pfeffer 1967). However, recent studies have shown that the NPTI has several shortcomings. It was proven to under-forecast lightning occurrence by Wohlwend (1998) although a bias-correction technique was applied by Roeder (1998) to improve performance. Howell (1998) and Everitt (1999, hereafter referred to as Everitt) showed that the 1-day persistence (hereafter persistence) forecast outperforms NPTI by ~10%. These issues indicated that the NPTI needed to be upgraded or replaced.

Because the forecasters require a tool that increases the reliability and objectivity of the daily thunderstorm probability forecast, they requested that a new lightning probability forecast tool be developed. In response, the AMU developed statistical lightning forecast equations using recent data and more sophisticated techniques now possible with more computing power than was available in the late 1960's. These equations provide a probability of lightning occurrence at KSC/CCAFS for the current day during the warm season (May–September).

1.1. Previous Work

Several studies have been conducted that address the issue of current-day thunderstorm forecasting specifically for the KSC/CCAFS area, and for the Florida peninsula as a whole. Aspects of those studies were important for the tool development in this task.

1.1.1. Neumann-Pfeffer Thunderstorm Index

Neumann (1971) used 13 years of CCAFS 1200 UTC sounding data (XMR) to develop the NPTI, using hourly surface observations of thunder as the predictand. A separate linear regression equation was developed for each month in the warm season using five predictors: wind vectors at 850 and 500 mb, average relative humidity (RH) in the 800–600 mb layer, Showalter Stability Index (SSI), and day number. The study accounted for non-linear effects by representing the predictors with second and third order polynomials rather than the predictor values themselves. The coefficients in the polynomials and the coefficients for the predictors in the linear regression equations varied by month. The NPTI outputs the probability of thunderstorm occurrence for the day, and also estimates the time of thunderstorm development.

The NPTI output was incorporated into the Meteorological Interactive Data Display System (MIDDS) and can be accessed from any computer connected to MIDDS. This is the objective lightning probability forecast tool currently in use at the 45 WS.

1.1.2. NPTI Improvements

One attempt to update NPTI and possibly improve its performance was done by Howell (1998) at the Air Force Institute of Technology (AFIT). This work used 2 more years of data than in Neumann (1971) for a total of 15 years. The same procedure used by Neumann (1971) to develop the equations was used in Howell (1998). The goal of the study was to determine whether more data in the development of NPTI would improve its performance. Forecasts from the original NPTI, the new NPTI, and persistence were compared. The new and current NPTI algorithms performed similarly, and persistence outperformed both by a small amount. Given the relatively small change in performance of the new NPTI, Howell (1998) recommended the current NPTI continue to be used, and also that a more reliable and accurate forecasting technique should be developed.

Everitt, also at AFIT, attempted to develop a new technique to replace the NPTI using ~25 years of hourly surface observations at the Shuttle Landing Facility (TTS) and morning XMR data. The TTS observations of thunderstorm occurrence were used as the predictand and observed and derived variables from the XMR sounding as predictors. Logistic regression was used instead of linear regression to develop the Stratified Logistic Thunderstorm Index (SLTI). As in the NPTI, the SLTI used the same predictors for each month. The predictors for the SLTI were different, however, and included the 850/700/600 mb winds, Thompson Index, K-Index, 800–600 mb mean RH, 6-day conditional climatology, and daily climatology. Another version of the NPTI (LNPTI) was created using logistic instead of linear regression with the NPTI predictors. An initial test in the study showed that persistence outperformed NPTI by ~11%. Tests with LNPTI and SLTI showed improved skill over persistence by 43% and 44%, respectively. Other performance indicators showed that SLTI was the superior forecast method.

Given the Everitt results, the AMU was tasked to assist the 45 WS in implementing the SLTI software (Wheeler 2001). The code was written using the MathCAD® software package, which the 45 WS did not have. The AMU converted and tested the code and created a program that could be run on any personal computer (PC) in the Weather Operations Center (WOC). Once in operations, the same performance results as those in Everitt could not be duplicated. In addition, the procedure to run the code was cumbersome due to the complex nature of creating the necessary predictor variables and transferring them from the MIDDS to a PC. After several months of testing the code with poor results, the 45 WS decided to stop pursuing the SLTI.

1.1.3. Daily Flow Regimes

Lericos et al. (2002, hereafter referred to as Lericos) used 10 years of data to develop lightning distributions over the Florida peninsula, stratified by flow regime. The 1200 UTC soundings at Miami (MIA), Tampa (TBW), and Jacksonville (JAX), Florida were used to define the flow regimes and data from the National Lightning Detection Network (NLDN) were used to determine the distributions. In all, six flow regimes were defined, with four relying on the latitudinal position of the subtropical ridge extending westward from the Atlantic Ocean. The mean 1000–700 mb wind directions at the three stations were calculated and combined to determine the flow regime for each day in the dataset. Then the NLDN data were used to determine the lightning distributions over the Florida peninsula for each of the flow regimes. Distinct maxima in lightning activity were noted near and over KSC/CCAFS when the ridge was south of the area, and minimal activity with other flow regimes. These results suggested that the daily flow regime is an important predictor of lightning occurrence for KSC/CCAFS.

1.2 Developing the Objective Lightning Probability Forecast Tool

The equation development in this work incorporated the results from Everitt and Lericos. In particular, the logistic regression approach from Everitt and the flow regimes from Lericos were considered significant new approaches not used in the NPTI. While the SLTI developed by Everitt was not transitioned to operations, the proof that the logistic regression method outperformed linear regression, even when using the same predictors, was a significant finding. This influenced the decision to use logistic regression in the AMU task. Based on the results in Lericos, flow regime information was added to the candidate predictor set in addition to the daily lightning climatology and several XMR sounding parameters as in Neumann (1971) and Everitt. The flow regimes developed by Lericos (2002) replaced the wind fields used in the NPTI.

The details of all data sources used and how they were manipulated to create the equation elements are provided in Section 2. Section 3 describes how the data were processed and manipulated to create the predictand and candidate predictors. The equation development and testing are described in Section 4, which shows how the predictors were chosen for the equations and how well the equations performed. Section 5 describes the graphical user interface (GUI) developed to provide an efficient way for the forecasters to manipulate the equations. Finally, Section 6 contains the conclusions and suggestions for future work.

2. Data

The period of record (POR) for the data used to develop the forecast equations was the 15 years 1989–2003 and the warm season months of May–September. The data sources include the

- Cloud-to-Ground Lightning Surveillance System (CGLSS),
- 1200 UTC JAX, TBW, and MIA soundings, and
- 1000 UTC XMR sounding.

Data from CGLSS, a local network of cloud-to-ground lightning sensors, were used to determine lightning occurrence for each day. The 1200 UTC JAX, TBW, and MIA soundings were used to calculate the daily flow regimes and the 1000 UTC XMR soundings were used to calculate the standard stability parameters that are readily available to the forecasters. The following sections describe each data type and how they were processed prior to the creation of the predictors and predictand for the statistical forecast equations. All data were processed using the S-PLUS® software package (Insightful Corporation 2001a).

2.1. Cloud-to-Ground Lightning Surveillance System (CGLSS)

The CGLSS is a network of six sensors (Figure 1) that collects date/time, latitude/longitude, strength, and polarity information of cloud-to-ground lightning strikes in the local area. These data were provided by Mr. Johnny Weems of the 45 WS and Mr. Paul Wahner of Computer Sciences Raytheon (CSR). The CGLSS data were used to determine whether or not lightning occurred on each day in the POR. Everitt used the TTS hourly surface observations of thunder for this purpose, but the CGLSS data are more reliable indicators of lightning occurrence in the local area than the surface observations (Mr. Weems, personal communication). Lericos used NLDN data to create lightning distributions based on flow regime. However, the CGLSS provides greater spatial accuracy and flash detection than the NLDN in the local KSC/CCAFS area (Harms et al. 1998). Using the Lightning Detection and Ranging (LDAR; Maier et al. 1995) data would have been more consistent with how the 45 WS issues lightning advisories. However, the LDAR data were not used due to the considerable size of the datasets and the shorter POR available. The primary purpose of the CGLSS data was to create the binary predictand for the equations. The data were also used to create a daily climatological frequency and persistence forecast that would be used as candidate predictors and forecast benchmarks against which to test the new equations.

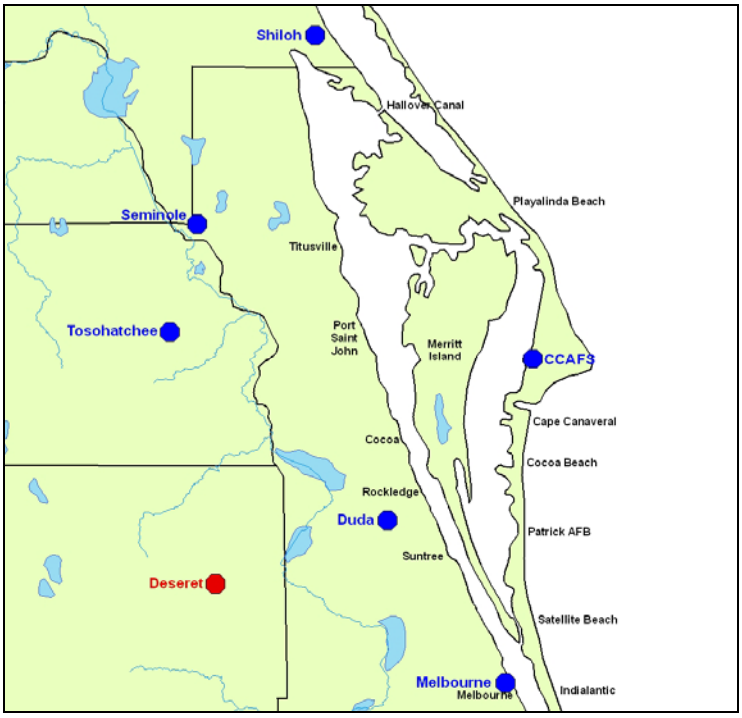

Figure 1. The locations of the six CGLSS sensors are indicated by the blue circles. The location names are next to the circles. The Duda sensor was moved to the Deseret site (red circle) in 2005.

As requested by the 45 WS, the equations were to forecast lightning within warning circles surrounding specific asset locations, each with a diameter of 5 n mi (Figure 2). This is analogous to a 45 WS Phase II lightning warning in which lightning is imminent or occurring within one of the circles. Ideally, the data should have been filtered to include only lightning strikes occurring within the circles in Figure 2. However, due to the complexity of computing the latitude/longitude boundaries of the intersecting circles, the area for this study was a rectangle surrounding the circles, defined by the outermost points of all the circles. The rectangle defining the spatial area of interest for this task encompasses the entire area shown in Figure 2. With this technique, a portion of the area enclosed in the rectangle is outside all of the 5 n mi circles, but any lightning within the rectangle would be sufficiently close as to cause the 45 WS to consider issuing a lightning advisory and may be reasonably included in the probabilities for lightning forecasts (Mr. William Roeder, 45 WS, personal communication). The latitude/longitude information in the CGLSS data was used to include only ground strikes that occurred within the area of the rectangle.

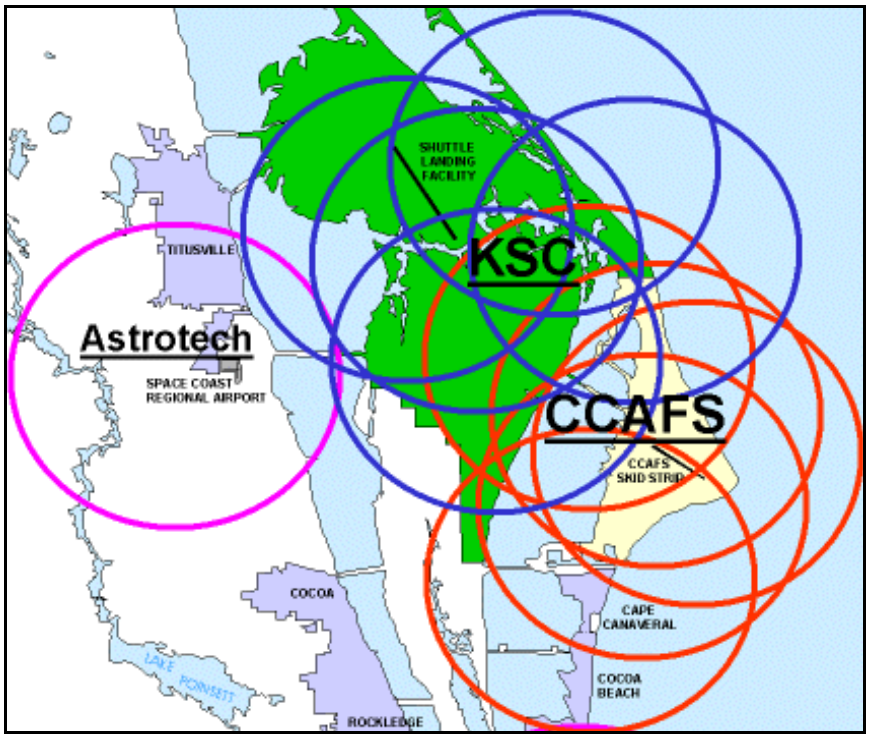

Figure 2. The 5 n mi lightning warning circles on KSC/CCAFS and Astrotech used to determine the spatial area for the lightning occurrence prediction and verification. The actual area is a rectangle surrounding all the intersecting circles at their outermost points, represented by the boundaries of this figure.

After the spatial filtering, the CGLSS data were filtered temporally to include only lightning strikes recorded in the time period 0700–0000 EDT. The 45 WS morning forecast is issued at 0700 EDT and is valid for 24 hours. However, the 45 WS verification procedure is for the current day, or Day 1, to end at midnight (0000 EDT). Times after midnight are considered Day 2. Since the goal of this task is to develop equations for Day 1 forecasting, lightning occurring between midnight and 0700 EDT were not considered.

2.2. Florida 1200 UTC Rawinsondes

These data were collected to determine the daily flow regimes using the procedure outlined in Lericos. Rawinsonde data for the period 1989–1997 were available on the CD-ROMs "Radiosonde Data of North America 1946–1996" (NCDC 1996) and "Radiosonde Data of North America 1994–1997" (NCDC 1997). Data from 1998–2003 were downloaded from the Forecast Systems Laboratory (FSL) web site http://raob.fsl.noaa.gov/.

Following the procedure in Lericos, the 1200 UTC soundings from MIA, TBW, and JAX were used to determine the large scale flow regime for the day. As noted in Lericos the current MIA and JAX sites were located at West Palm Beach, FL (PBI) and Waycross, GA (AYS), respectively, prior to 1995. The PBI and AYS data were used as proxies for MIA and JAX, respectively, during the period 1989–1994. All future references to MIA and JAX include the 1989–1994 data from AYS and PBI. The map in Figure 3 shows the locations of all the soundings used in this task.

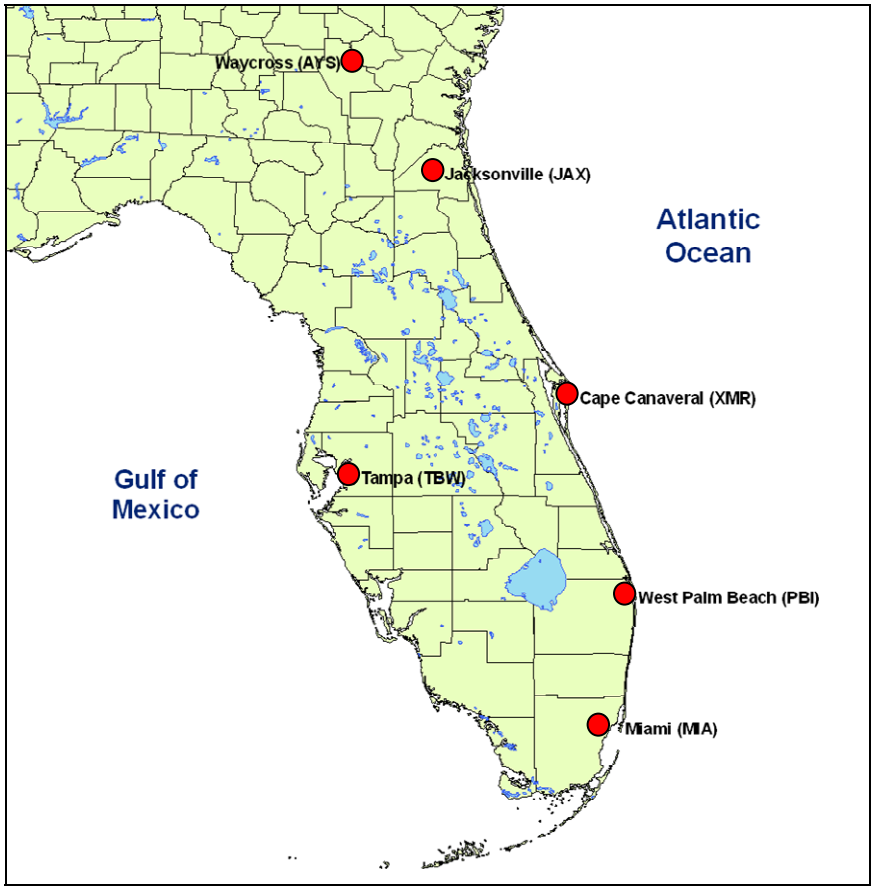

Figure 3. The red dots on the map show the locations of all soundings used in this task.

Use of the 1200 UTC sounding may seem inappropriate as it cannot provide data in time for the 1100 UTC briefing. Use of the 0000 UTC sounding from the day before was ruled out as the 1000–700 mb flow during the Florida warm season could be contaminated by afternoon convective circulations that mask the larger scale flow pattern. For the purpose of determining the flow regimes for each day in the POR, the 1200 UTC sounding provided the most reliable data. Due to the weak synoptic patterns during the Florida warm season, it is not likely that a flow regime change would take place in the two-hour period between 1000–1200 UTC. In an operational setting, the 45 WS can use several data sources, including model output and surface observations, to help determine the flow regime of the day before the morning 1100 UTC briefing. Specific suggestions for data sources and procedures that can be used to determine the flow regime will be discussed in Section 6.

2.3. CCAFS 1000 UTC Rawinsonde

The CCAFS sounding location (XMR) is shown in Figure 3. Data from the 1000 UTC sounding are used for the 1100 UTC morning briefing since it contains the most recent information on the state of the atmosphere over the area. These data were used to calculate the sounding parameters normally available to the forecasters through MIDDS. The parameters were used as candidate predictors in the equation development. Mr. Wahner of CSR supplied these data to the AMU.

The original dataset included all soundings taken on every day in the warm season. The data were first filtered to include only 1000 UTC soundings. After noting that there were many days with missing 1000 UTC soundings, a check was done to determine all sounding times and the number of soundings at each time. Many of the soundings for the "missing" days were from rawinsondes released at 1015 or 1020 UTC, a few between 1020 and 1030 UTC, and very few after 1030 UTC. To include all days with 1000 UTC soundings that may have been released late, soundings taken at times between 1000–1030 UTC were included in the dataset. The cutoff time of 1030 UTC was chosen for two reasons. First, very few soundings were taken after this time. The second reason was to ensure the simulation of a real-time situation in which the sounding data must be available to the 45 WS forecasters for the 1100 UTC weather briefing. A sounding released after 1030 UTC may not provide the data in time for the briefing. An automated check was developed to ensure that only one sounding occurred during the 30-minute period on each day. On the few days in which two soundings were found, their times were often only 1-2 minutes apart. Mr. Weems of the 45 WS said that these were most likely re-transmissions of the sounding due to a possible error in the first transmission, and that the latter sounding should be used in the analysis. Hereafter, references to the 1000 UTC sounding include soundings taken in the time period 1000–1030 UTC, inclusive.

Each individual sounding was separated into three groups in the original dataset: thousand-foot, mandatory-level, and significant-level data. The algorithms in MIDDS use a combination of the mandatory- and significant-level data. Therefore, the mandatory- and significant-level observations in each sounding were combined into one group and sorted by height to create complete individual daily soundings beginning at the surface and extending to the highest observed level. The thousand-foot data were not used.

3. Preparation of Equation Elements

After the data were processed as described in Section 2, each of the three datasets was manipulated and analyzed to create the elements needed for the statistical forecast equation development. The necessary elements include a predictand and candidate predictors. The predictand is the element to be predicted from a predictor or group of predictors. The filtered CGLSS data provided the ground truth indicating whether or not lightning occurred and were used to create the predictand. The sounding datasets were used to calculate the candidate predictors of lightning occurrence, consisting of stability indices and lightning probabilities based on flow regime.

3.1. Binary Predictand

Calculation of the predictand was straightforward: the predictand value was set to '1' if lightning was detected within the defined time period and spatial area on a specific day, otherwise a '0' was assigned. A binary predictand was used because the prediction would be for lightning occurrence, not the number of strikes. Although a larger number of lightning strikes increases the probability of a hit in a sensitive area, the 45 WS verification procedure only requires one strike for a lightning warning to be validated.

3.2. Candidate Predictors

The candidate predictors were tested prior to and during equation development to determine which predictors in what combination would provide the best probability forecast of lightning occurrence. They included a 1-day lightning persistence and daily climatological lightning frequency calculated from the CGLSS binary predictand, the flow regimes from the Florida rawinsondes, and 10 stability parameters calculated from the XMR rawinsonde.

3.2.1. CGLSS Predictors

The binary predictand was used to create two candidate predictors: a binary persistence and a daily climatological frequency of lightning occurrence. Calculation of the persistence was straightforward. If lightning occurred on a particular day, the persistence value for the next day was '1'. If lightning did not occur, the persistence value was '0'. The lightning occurrence information for April 30 was used to create the persistence value for May 1 in each year. A persistence value was created for each individual day in the POR.

A 15-year climatological probability of lightning occurrence was calculated for each of the days in the warm season, 1 May–30 September 1989–2003. The number of years that each day experienced lightning was determined first. Then, a raw climatology was calculated by dividing this number by 15, the number of years in the POR. This yielded a fractional value between 0 and 1 for each day. The light blue jagged curve in Figure 4a is the raw 15-year climatological probability for the warm season. The noisy appearance of this curve is likely due to the few number of years in the POR: 15 is a small number of observations from which to calculate a climatology. A common procedure to minimize the noisiness of such a curve is to use a weighted average of the observations several days before and after the day of interest, artificially increasing the number of observations used to determine the daily lightning probability to infer what the long-term climatology would be if enough observations were available. Following Everitt, a 15-day time period from seven days before to seven days after the day of interest was used with, at the suggestion of Mr. Roeder, a Gaussian weighting function with a scale factor of 3-days. The dark blue curve in Figure 4a is a smoothed climatology that was calculated with the equation

$$P = \frac{1}{N}\left\{\frac{\sum_{k=1}^{7}\left[W(F_{n-k} + F_{n+k})\right] + F_n}{\sum_{k=1}^{7}\left[W * 2\right] + 1}\right\} \text{ (Everitt)}, \tag{1}$$

where W is the Gaussian weighting function

$$W = \exp\left[\frac{-\left(k^2\right)}{2 * \sigma^2}\right] \text{ (Wilks 1995) with } \sigma = 3, \tag{2}$$

P = climatological probability on the day of interest,

N = number of years in the POR (15),

F = raw probability on day of interest,

n = day number of interest, and

k = number of days distant from *n*.

Figure 4b shows the weight values (W) used in the calculations. The weight value for the day of interest was 1, giving it full weight for the frequency calculation. The weight values for the seven days before and after the day of interest decreased normally as the temporal space increased. The lightning probabilities for the last seven days in April and first seven days in October were used to calculate the probabilities at the beginning of May and end of September, respectively.

The probabilities were small at the beginning and end of the season, but approached a maximum near 70% in mid-July. The significance and cause of the fluctuations in the climatology curve in Figure 4a are not known. A similar pattern also appeared in the climatology calculated by Everitt. The fluctuations in the curve from May to the end of June and the end of August through September might reflect yearly differences in the onset and conclusion of the convective season, respectively.

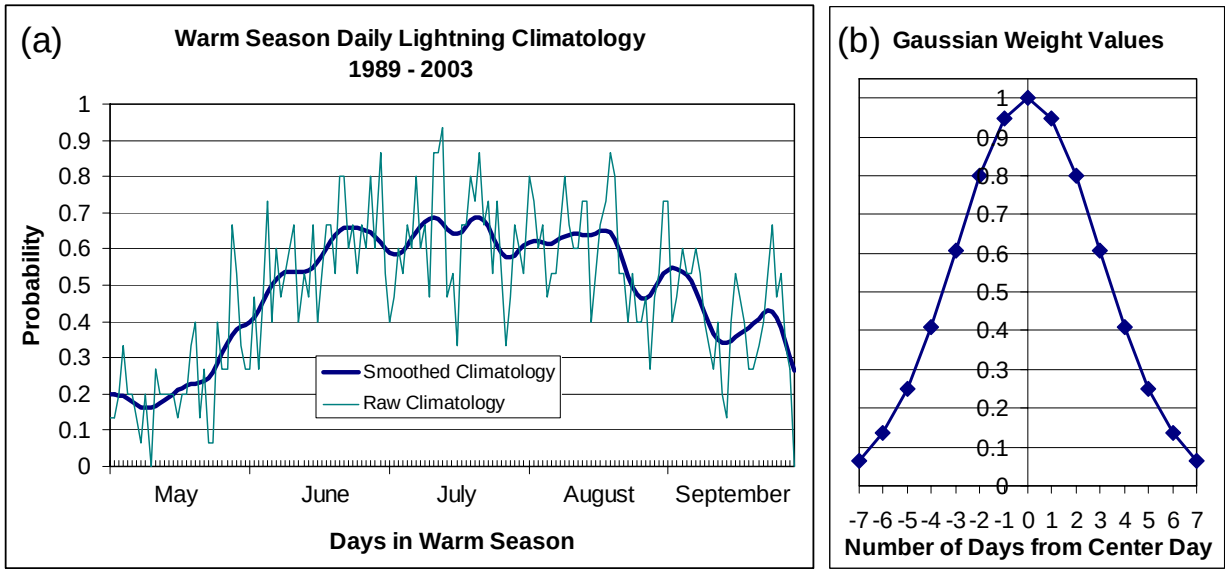

Figure 4. (a) The daily raw (light blue) and smoothed (dark blue) climatological probability values of lightning occurrence for the warm-season months in the POR 1989–2003, and (b) The Gaussian weight values used in the 15-day smoothing equation.

3.2.2. Flow Regime Predictors

The 1000–700 mb layer-average winds were calculated and a peninsular-scale flow regime for each day was determined following the procedures outlined in Lericos. Then, lightning probabilities based on flow regime were calculated using the binary CGLSS predictand.

3.2.2.1. 1000–700 mb Layer-Average Winds

The average wind direction in the 1000–700 mb layer at MIA, TBW, and JAX was calculated with a depth-weighted averaging method using all observations in the layer. The wind speeds and directions at each level were first separated into u- and v-components with the equations

$$u_i = \text{Speed}_i * \cos\left[(270 - \text{Direction}_i) * \frac{\text{pi}}{180}\right] \text{ and}$$

$$v_i = \text{Speed}_i * \sin\left[(270 - \text{Direction}_i) * \frac{\text{pi}}{180}\right], \quad (3)$$

where ‘i’ represents the level number in the layer and (pi/180) is the factor to convert direction from degrees to radians. The depth, D_i, was a region of influence about the observation at height H_i in the sounding and was calculated using the equation

$$D_i = \frac{(H_{i+1} - H_{i-1})}{2}, \tag{4}$$

where H_{i+1} is the height at the observation directly above and H_{i-1} is the height at the observation directly below H_i. In this calculation, the observation has influence in the region encompassing half the distance to the adjacent observations above and below it. The layer-averaged value for each component was computed with the equations

$$u_{avg} = \frac{\sum(u_i * D_i)}{\sum D_i} \text{ and } v_{avg} = \frac{\sum(u_i * D_i)}{\sum D_i}. \tag{5}$$

The layer-averaged components were combined in the following equations to determine the average wind speed and direction in the layer:

$$Speed_{avg} = \sqrt{(v_{avg})^2 + (u_{avg})^2}, \tag{6}$$

$$\text{if } u_{avg} \geq 0,\ Direction_{avg} = 270 - \left[\arctan\left(\frac{v_{avg}}{u_{avg}}\right) * \frac{180}{pi}\right], \text{ or}$$

$$\text{if } u_{avg} < 0,\ Direction_{avg} = 90 - \left[\arctan\left(\frac{v_{avg}}{u_{avg}}\right) * \frac{180}{pi}\right], \tag{7}$$

where (180/pi) is the factor to convert from radians to degrees.

3.2.2.2. Flow Regime Determination

The flow regime for each day depended on the combined layer-averaged wind directions at the three stations. There were six flow regimes in Lericos, including one named Calm in which the average wind speed at all three stations was < 2 m/s. There were a large number of Calm days in Lericos, accounting for almost 20% of all days in the POR. Calculations with the dataset in this task identified less than 1% of days as Calm. Subsequent conversations with Dr. Henry Fuelberg and Ms. Jessica Stroupe at the Florida State University revealed that, while Ms. Stroupe was working on her thesis (Stroupe 2003), she discovered an error in the original Lericos algorithms. Once corrected, the Lericos code calculated no Calm days. Ms. Stroupe assisted the AMU in determining possible reasons why the AMU algorithm calculated a few Calm days and the modified Lericos algorithm calculated none. The differences in calculated values were determined to be insignificant and were attributed to differences in software and computer platforms.

There are six defined flow regimes in this task, named according to the resulting flow over KSC/CCAFS. The Calm regime in Lericos was replaced with a northeast flow regime (NE) in which the wind direction at all three stations was in the northeast sector. Table 1 shows the flow regime names, a description of the flow, and the wind direction sectors at each station that define the regime. The wind directions must be within the sector defined at each station for a particular flow regime to be valid for the day. If one or more of the stations exhibited a wind direction in a sector other than those defined in Table 1, a seventh flow regime, Other, was assigned. All three soundings had to be available to classify the flow regime. If one or more soundings were missing on any day, no flow regime classification was made and data from that day were not used in the equation development. Note that the first four regimes in Table 1 depend on the position of the ridge that extends westward toward the Florida peninsula from the Bermuda High typically in place over the western Atlantic Ocean during the warm season.

Table 1. List of the flow regime names used in this study and the corresponding sectors showing the average 1000–700 mb wind directions at each of the stations.

Flow Regime Name and Description		***Rawinsonde Station***		
		MIA	**TBW**	**JAX**
SW-1	Subtropical ridge south of MIA Southwest flow over KSC/CCAFS	180°-270°	180°-270°	180°-270°
SW-2	Subtropical ridge north of MIA, south of TBW Southwest flow over KSC/CCAFS	90°-180°	180°-270°	180°-270°
SE-1	Subtropical ridge north of TBW, south of JAX Southeast flow over KSC/CCAFS	90°-180°	90°-180°	180°-270°
SE-2	Subtropical ridge north of JAX Southeast flow over KSC/CCAFS	90°-180°	90°-180°	90°-180°
NW	Northwest flow over KSC/CCAFS	270°-360°	270°-360°	270°-360°
NE	Northeast flow over KSC/CCAFS	0°-90°	0°-90°	0°-90°
Other	When the layer-averaged wind directions at the three stations did not fit in defined flow regime			
Missing One or more soundings missing				

3.2.2.3. Flow Regime Climatology

The frequency distributions of flow regimes for each month and for the entire warm season were created to determine if any particular regime(s) dominated in any month or the whole season. This may help forecasters determine which month(s) lightning is most likely to occur over KSC/CCAFS, should any of the flow regimes prove dominant in lightning occurrence. Figure 5a-f contains bar charts showing the number of days that each flow regime occurred in each individual month (Figure 5a-e) and for all warm-season months combined (Figure 5f) in the POR. The possible maximum number of days was 2295 for the entire POR, 465 for May, July, and August, and 450 for June and September. There were relatively few Missing days at 167 (7%) over the POR, with a relatively uniform distribution through the five individual months. This resulted in a total number of 2128 days (93%) in which it was possible to determine a flow regime.

One notable item in all the charts is that the Other category had the most occurrences by far. Lericos initially considered 14 flow regimes, but found several of the regimes had an insufficient number of occurrences to calculate meaningful statistics. Regimes were dropped or combined resulting in six regimes. It is possible that the Other category is made up of those regimes eliminated in Lericos, and possibly other regimes. In any case, the number of occurrences in this category was significant and was considered as a legitimate flow regime in the analysis.

The NW and NE regimes had the least number of occurrences in all months. Discounting the Other category, the SW-1, SW-2, SE-1, and SE-2 regimes dominated during the warm season. Since these regimes are based on the position of the subtropical ridge extending from the Bermuda High, this shows the strong influence that this ridge has on the flow patterns over the Florida peninsula during the warm season. In May (Figure 5a), SW-1 flow dominated with a somewhat even distribution of the other regimes. The SW-1, SW-2, and SE-1 regimes dominated in June and July (Figure 5b-c), indicating that the ridge tended to stay south, over, or just to the north of KSC/CCAFS (i.e. south of JAX). In August (Figure 5d) a transition to increased easterly flow over KSC/CCAFS began, shown by the increase in SE-2 regime events but with a uniform distribution between the four ridge regimes. By September the transition to more easterly flow events was complete with both SE regimes dominating over both SW regimes, and even an increase in the NE regime.

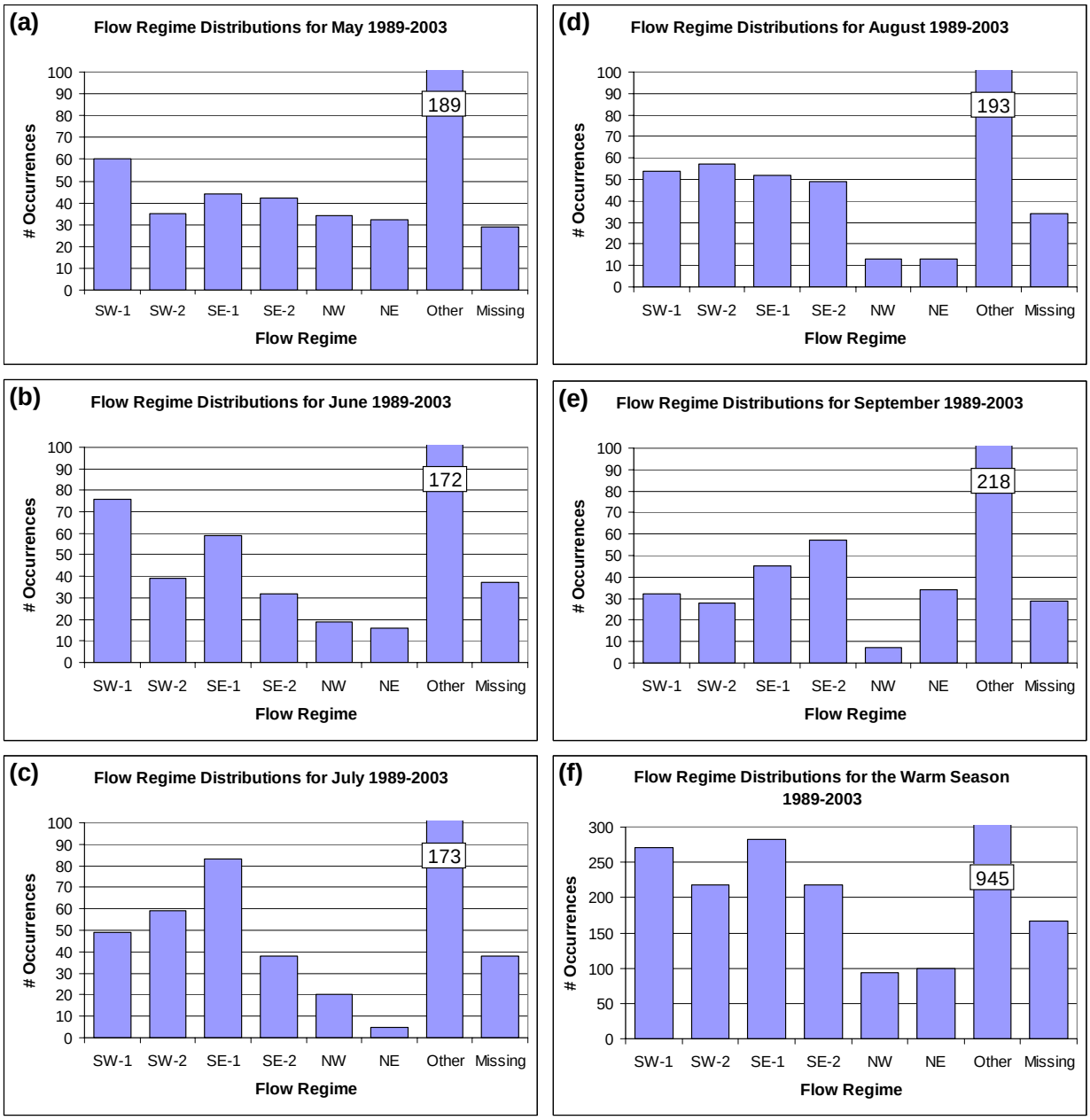

Figure 5. Bar charts showing the number of days each flow regime was observed in (a) May, (b) June, (c) July, (d) August, (e) September, and (f) all warm season months in the period of record 1989–2003. The values for the 'Other' category are shown on that bar. The y-axis was scaled to increase the resolution of all other regimes, thereby cutting off the 'Other' regime bar since had a large number of occurrences in each month.

3.2.2.4. Flow Regime Lightning Probabilities

The probabilities of lightning occurrence based on flow regime for each month and the entire warm season were calculated using the CGLSS binary predictand. The number of days that each regime occurred was checked against the CGLSS predictand to see how many of those days experienced lightning. The climatological probability was calculated simply by dividing the number of lightning days within a particular regime by the total number of days the regime occurred.

After calculating the probabilities by flow regime and examining their values, it was clear that they could be excellent predictors of lightning occurrence over KSC/CCAFS. The probability values were so definitive that the 45 WS requested tables of the probability values be created for each month and transitioned for immediate operational use until the final lightning probability tool was available. The tables were created through several iterations with Mr. Roeder and Mr. Weems of the 45 WS to ensure the tables would be useful to operations. Six tables were created: one for the entire warm season and one for each of the five months in the warm season. Each table has a descriptive caption at the top, six columns and a notes section at the bottom. Table 2 provides an example of the content found in all the tables. It contains the lightning statistics by flow regime for the entire warm season.

Table 2. Example of the tables containing the lightning probabilities based on flow regime. This table contains the probabilities for all the months in the warm season combined.

Flow Regime Lightning Statistics
Warm Season (May–September) 1989–2003

Probabilities of lightning occurring within a rectangle encompassing all 5 n mi warning rings based on flow regime are shown in the right-most column.

The strikes/day statistical values in the second column are based on lightning days only (fifth column). The median (M) value of strikes per day in each regime is shown with the 1st (Q1) and 3rd (Q3) quartiles in the order Q1, M, Q3. The mean and standard deviation of the strike numbers are shown in parentheses below Q1, M, Q3 (see explanation of M, Q1, and Q3 below).

Flow Regime	***Q1, M, Q3 of Strikes/Day (Mean, Stdev)***	***Total # Days (% of Total)***	***# Non Lightning Days***	***# Lightning Days***	***Probability of Lightning***
SW-1 Ridge S of MIA	68, 248, 507 (396, 496)	271 (12.7)	92	179	**66 %**
SW-2 Ridge between MIA/TBW	37, 169, 528 (357, 435)	218 (10.2)	60	158	**72 %**
SE-1 Ridge between TBW/JAX	4, 18, 110 (117, 223)	283 (13.3)	140	143	**51 %**
SE-2 Ridge N of JAX	3, 8, 41 (61,141)	218 (10.2)	133	85	**39 %**
NW	28, 179, 359 (342, 545)	93 (4.4)	53	40	**43 %**
NE	2, 14, 62 (68, 114)	100 (4.7)	82	18	**18 %**
Other (Regime Undefined)	9, 65, 265 (200, 325)	945 (44.4)	527	418	**44 %**
TOTALS	**10, 75, 324 (238, 381)**	**2128**	**1087**	**1041**	**49 %**

There is a 6% improvement in the forecast when using the individual flow regime probabilities over the seasonal climatological probability of 49%, and a 23% improvement over persistence. Forecast improvement was calculated using the Brier Skill Score.

The median is the strike-number value at which 50% of the cases had higher and 50% had lower strike numbers, i.e. the center of the strike-number distribution. It is *not* equal to the mean because the strike-number distributions are not symmetric. The 'middle' 50% of the cases are found between Q1 and Q3. For asymmetric distributions, like lightning strikes/day, the median and inter-quartile ranges are more representative of the data than the mean and standard deviation.

The first column (left-most) in Table 2 contains the names of the flow regimes as defined in Table 1. The second column contains statistical properties of the strike counts for days on which lightning occurred in each flow regime. The third column shows the number of days and the percentage of total days that each flow regime occurred during the period. The fourth column shows the subset of flow regime days on which lightning did not occur, and the fifth column shows the number of days on which lightning did occur. The value in the sixth (right-most) column contains the climatological probability of lightning occurrence based on flow regime. This is the value to be used by the forecasters, and was also a candidate predictor for the equations. The Totals row in Table 2 shows the values for all flow regimes combined. The value in the sixth column of this row contains the climatological lightning probability for the entire warm season. For each of the monthly tables, it is the monthly climatology.

There is further information found in the notes in the last row of Table 2. The first note describes the forecast performance of the flow regime probabilities when compared to that of climatology and persistence in terms of percent forecast improvement or degradation. The second note gives a brief description of the median and first and third quartiles of the daily strike numbers in the second column. The details of how these values were calculated and other aspects of the tables were written in an AMU Memorandum (Lambert 2004a). The tables and memorandum were distributed to the 45 WS, SMG, and the National Weather Service (NWS) at Melbourne, FL (MLB) for their use during the 2004 warm season.

The flow regime lightning probability values for the individual months were used as candidate predictors in the equation development and the overall monthly climatologies were used as forecast benchmarks in determining the skill of the equations. The values for these parameters are in the sixth column of the individual monthly tables in Lambert (2004a) and are shown in Table 3. The values for the SW-1 and SW-2 regimes were calculated separately for each month. However, the values were within 10% of each other. Therefore, the SW-1 and SW-2 days in each month were combined to increase the sample size and produce a more reliable probability value. The resulting combined SW1-2 values for June, July, and August were also within 10% of each other, and the days for these flow regimes and months were combined to create one SW value for the three months. Also for June–August, the SE-1 and SE-2 regimes were within 10% of each other within and between months. Their values were also combined to create one SE flow regime value for all three months. This was not the case for the SE flow regimes in May and September, therefore there are separate columns for SE-1 and SE-2 in Table 3. The parentheses around the SE-2 values for June–August indicate that it is a combined value and the same as SE-1.

Table 3. Monthly probabilities of lightning occurrence based on the flow regimes that were used as candidate predictors. The values in the far-right column are the monthly probabilities for all flow regimes combined, and were used as a forecast benchmark.

Month	SW1-2	SE-1	SE-2	NW	NE	Other	Monthly
May	31	36	10	24	6	23	23
June	79	51	(51)	58	19	51	57
July	79	51	(51)	60	40	58	63
August	79	51	(51)	69	23	54	59
September	72	60	35	0	24	38	43

3.2.3. Stability Index Predictors

The stability indices calculated from the 1000 UTC XMR sounding were those normally available to the forecasters through MIDDS. In order to calculate the same values that would be available to the forecasters, the same equations used in the MIDDS code were used. MIDDS uses the Man-computer Interactive Data Access System (McIDAS) software (Lazzara et al. 1999) for processing sounding data. Mr. Wahner of CSR provided copies of all the necessary McIDAS code.

The stability index candidate predictors included the

- Total Totals (TT),
- Cross Totals (CT),
- K-Index (KI),
- Lifted Index (LI),
- Thompson Index (TI),
- Severe Weather ThrEAT Index (SWEAT),
- Showalter Stability Index (SSI),
- Temperature at 500 mb (T_{500}),
- Mean Relative Humidity in the 800–600 mb layer (RH),
- Precipitable water up to 500 mb (PW),
- Convective Inhibition (CIN),
- Convective Available Potential Energy (CAPE),
- CAPE based on the forecast maximum temperature, and
- CAPE based on the maximum θ_e below 300 mb.

3.2.3.1. Calculation of Stability Indices

The formulas in the McIDAS code used for the indices are standard and can be found in several sources (e.g. Peppler and Lamb 1989; Ohio State University Severe Weather Products web page at http://twister.sbs.ohio-state.edu). They will not be repeated here. Only two indices in the above list are not readily available to the forecasters: TI and RH. The TI is calculated easily with the equation TI = KI – LI. The RH was calculated using a depth-weighted average in the same form as that used for the average wind direction in the 1000–700 mb layer (Section 3.2.2.1, Equations 4 and 5).

Certain issues arose when calculating the level of free convection (LFC) with the McIDAS algorithms. The LFC was necessary to calculate the three CAPE values and CIN in the above list. In certain soundings, the LFC was calculated to be below the lifting condensation level (LCL), a physical impossibility. LFC values ranged from within a few millibars of the LCL to 6300 mb. There were also some negative LFC values down to -2500 mb. Even if the LFC value is unrealistic, the McIDAS code would still calculate a CAPE value. This CAPE value would be incorrect and misleading to forecasters. Iterations with the McIDAS developers revealed that this issue was caused by the equations and assumptions used in the algorithms rather than bad data quality. The McIDAS developers agreed to make certain changes to the code that would fix the issue. Those same changes were made to the code in this task and new CAPE values calculated. Details of all the issues found in the McIDAS code and their proposed solutions are in an AMU Memorandum that was distributed to the 45 WS, SMG, and NWS MLB (Lambert 2004b).

While working on the LFC issues with the McIDAS developers, Mr. Weems suggested using the CAPE algorithms in another program available on MIDDS called HUGE. The HUGE program was acquired several years ago from the National Severe Storms Laboratory. It analyzes sounding data and calculates many of the same stability parameters as McIDAS. Mr. Rick Kulow of CSR provided the code and a sample input file to test the program. While testing the algorithms and analyzing their output, several typographic and logic errors were noticed in the algorithms causing erroneous output. Since this is a common algorithm used by forecasters, the findings were summarized in an AMU Memorandum (Lambert 2004c) and distributed to CSR, 45 WS, SMG, and NWS MLB.

3.2.3.2. Analysis of Stability Indices

Before using the 14 candidate stability index predictors from the list above in the equation development, a test was done to ensure their validity as predictors. An index that did not pass the test would no longer be considered a candidate predictor. The indices were stratified first by month, and then stratified between lightning and non-lightning days. Mean values for each of the 14 stability indices were calculated separately for the lightning and non-lightning days in each month, then checked to see if there was a statistically significant difference between them.

The stability index means for the lightning and non-lightning days were always unequal, but that did not mean the differences were significant. To check whether the differences were statistically significant, a two-sample two-sided Student's t-test (Wilks 1995) in S-PLUS was applied. This form of the Student's t-test helps determine the probability that two sample means came from the same population. The two-sided test checks whether the means are different, not which one is larger or smaller. The null hypothesis in the test is that the two means are equal. The Student's t-test in S-PLUS produces a p-value that is used to determine the confidence level at which this null hypothesis can be rejected. The p-value represents the probability of error involved in accepting that the difference between the two means is significant (Statsoft, Inc. 2004), or the likelihood that the difference in the means is due to chance. The smaller the p-value, the less likely the difference is due to chance and the more probable that the difference is significant. The common convention is to use a p-value of 0.05 (5%) as the threshold value to accept or reject the null hypothesis. This is interpreted as having 95% confidence that the means are not equal.

This test was conducted on each set of means for each stability index in each individual month and for all months combined. With the exception of CIN and the three CAPE values, the null hypothesis for most stability parameters could be rejected at the 99+% confidence level, indicating that the differences in their means were statistically significant. Very little confidence could be placed in rejecting the null hypothesis for CIN and the three CAPE values. The p-values for these indices were between 0.5 – 0.9, indicating that any differences in mean values between lightning and non-lightning days was not statistically significant. This was not a surprise to local forecasters since anecdotal evidence suggests that there is often substantial CAPE on both lightning and non-lightning days as low level warm air and moisture are abundant in the Florida warm season.

Another difficulty in using CAPE and CIN as predictors was that a value was not able to be calculated for every sounding. The McIDAS code needs to calculate an LFC before calculating CAPE and CIN. If an LFC could not be found, the values were not calculated. This artifact of the code resulted in reducing the available dataset by another 10% beyond that accounted for by missing data. Given that the difference in CAPE and CIN means between lightning and non-lightning days was not statistically significant, it was not worth losing the extra data. Therefore, all the stability indices except CIN and the three CAPE values were used as candidate predictors.

3.2.4. Summary of Candidate Predictors

A summary of the candidate predictors is given here as a reference for the reader. As a result of the analyses presented in Sections 3.2.1 – 3.2.3, 13 candidate predictors were created for the equation development. They are

- Persistence,
- Daily climatological lightning frequency,
- Flow regime lightning probability,
- Total Totals (TT),
- Cross Totals (CT),
- K-Index (KI),
- Lifted Index (LI),
- Thompson Index (TI),
- Severe Weather ThrEAT (SWEAT) Index,
- Showalter Index (SSI),
- Temperature at 500 mb, (T_{500}),
- Mean Relative Humidity in the 800–600 mb layer (RH), and
- Precipitable water up to 500 mb (PW).

The values for these candidate predictors were used with the binary predictand in the development of the statistical lightning forecast equations.

4. Equation Development and Testing

There were three major steps in this portion of the task:

- Ascertain data availability,
- Develop the logistic regression equations, and
- Determine the equation performance.

The amount of data available for equation development was critical to the reliability of the new equations. After determining that an appropriate amount of data was available, a set of five equations was developed, one for each month in the warm season. The performance of the equations was assessed using several verification techniques appropriate for probability forecasts.

4.1. Data Availability

The amount of available data was determined before equation development began. This was important since the data had to be stratified into equation development and verification datasets followed by stratification into monthly datasets, thereby limiting the amount of data available for equation development. To ensure that the new equations would be reliable, ample data were required to create realistic relationships between the predictors and the predictand. The World Meteorological Organization (1992, hereafter WMO) states that there should be at least 250 events in the dataset in order to derive stable statistical relationships. This was the threshold in determining whether there were sufficient data in the POR.

4.1.1. Missing Data

There are 153 days in any given warm season, 1 May–30 September. This equates to 2295 days over the 15-year POR. Sounding data were not available for every day in the POR. Data were considered missing for a specific day if there was one or more Florida rawinsonde missing (MIA, TBW, or JAX), or when a 1000 UTC XMR sounding was missing. Table 4 shows a summary of how many days were in the POR, how many of those days had missing data, which type of data was missing, and the total number of days with available data. In most of the missing cases, data were not available from either the XMR or the Florida rawinsondes. There were few cases in which data were missing from both sources on the same day. The number in the third column under the heading "# Missing Obs" in Table 4 is less than the sum of the first two columns in every case except for September because there were a few "overlap" days in which data were missing from both sources. The numbers of overlap cases are shown in parentheses in the third column. The sum of the first two numbers in the Total row is 313 (167 + 146), but the total missing is 297. This says that data were missing from both sources on the same day only 16 times.

The final column in Table 4 shows that data availability ranged from 85–90% for each month, and 87% overall. Most important, though, was the actual number of available days per month, ranging from 389 – 420. This was promising in that it was still probable that there would be a sufficient number of events for the equation development, according to the WMO standard, after stratifying the full dataset into development and testing datasets.

Table 4. Summary of missing and available data in the POR. The first column contains the name of each month in the warm season, where Total is for the entire warm season. The two columns under the heading "# POSSIBLE DAYS" show the number of days in 1 warm season and 15 warm seasons. The three columns under the heading"# MISSING DAYS" show the number of unavailable days due to missing data from each source in the subheadings, and the number of days missing due to the combined contribution of missing data from both sources. The value in parentheses in the third column is the number of days in which data were missing from both sources. The final column shows the number of days with all data available. The percent of total possible days is given in parentheses.

Warm Season Months	# POSSIBLE DAYS		# MISSING DAYS			Total Available (% of # Possible)
	1 Year	15 Years	MIA TBW JAX	XMR	Total (Overlap)	
May	31	465	29	21	45 (5)	420 (90)
June	30	450	37	29	61 (5)	389 (86)
July	31	465	38	25	60 (3)	405 (87)
August	31	465	34	39	70 (3)	395 (85)
September	30	450	29	32	61 (0)	389 (86)
Total	**153**	**2295**	**167**	**146**	**297 (16)**	**1998 (87)**

4.1.2. Development and Verification Datasets

The development dataset required enough samples so that the resulting set of equations was stable, i.e. the equations would maintain consistent forecast accuracy on different datasets. A small dataset may not contain a representative set of events. The equations developed from such a small set may show wide variations in accuracy on different datasets causing forecasters to not have confidence in the results. The verification dataset was needed for equation testing in order to have a more realistic view of how the equations would perform in operations. It was expected that the equations would not perform as well on the verification data as they would on the data from which they were developed. However, if performance were a great deal worse with the verification data, this would indicate that either too many predictors were chosen and the equations were fit too strongly to the development data, or the development dataset was too small.

The dataset described in Section 4.1.1 was stratified into development and verification datasets. Care was taken to ensure there would be at least 250 events in the development dataset, while still having enough events in the verification dataset to make reasonable conclusions about equation performance. Of the 15 warm seasons in the POR, 13 were used for equation development and 2 were set aside for testing the equations. This ensured that each month in the warm season was equally represented in both datasets.

The stratification did not involve choosing individual warm season years for each dataset, but rather individual warm season days. Days for the verification dataset were chosen first. Given that there are 153 days in the warm season, the random number generator in Microsoft® Excel© was used to create two sets of 153 numbers representing the years between and including 1989 and 2003. The resulting two sets of years were assigned to each day in the warm season. Thus, each day in the warm season was represented by days from two random "years". For example, the verification dataset contains 1 May 1992 and 2000, 2 May 1998 and 1999, etc. All other dates were made part of the development dataset. This random method was chosen to reduce the likelihood that any unusual convective seasons would bias the results. Table 5 shows the possible number of events for the development and verification datasets and the actual number of events after accounting for missing data. Note the number of days in the development dataset for each month in the right-most column. All are well above the 250 events defined by the WMO needed to develop reliable equations.

Table 5. Summary of missing and available data for equation development and testing. The first column contains the name of each month in the warm season, where Total is for the entire warm season. The three columns under the heading "# POSSIBLE DAYS" show the number of days in 15 warm seasons, the number of those days for equation testing, and the number for equation development. The three columns under the heading"# AVAILABLE DAYS", show the number of days actually available in the POR due to missing data (from Table 4), and the actual number of days in the verification and development datasets.

Warm Season Months	# POSSIBLE DAYS			# AVAILABLE DAYS		
	Total	Verification	Development	Total	Verification	Development
May	465	62	403	420	56	364
June	450	60	390	389	51	338
July	465	62	403	405	51	354
August	465	62	403	395	51	344
September	450	60	390	389	48	341
Total	**2295**	**306**	**1989**	**1998**	**257**	**1741**

4.2. Equation Development

Similar to Everitt, five logistic regression equations were created, one for each month. In Everitt, predictors were chosen based on their relationship to the predictand over the whole warm season resulting in the same predictors being used in each month. However, the predictors were regressed against the predictand for each individual month. This created different values for the predictor constants in the individual monthly equations. In this task, predictor selection was conducted for each individual month due to the possibility that different variables may become more critical to convection formation as the warm season progresses.

4.2.1. Logistic Regression

Besides data availability, another important factor in creating a reliable probability forecast tool is the selection of the statistical regression method. According to Wilks (1995), logistic regression is the appropriate method when the predictand is binary. Everitt showed that logistic regression yielded 48% better skill over the linear regression equations in NPTI when using the same predictor variables and data. The gain in skill was solely due to use of the logistic regression method. Given a predictand, y, and a set of predictors x_1–x_k, where k is the total number of predictors, logistic regression is represented by the equation

$$y = \frac{e^{(b_0 + b_1 x_1 + \ldots + b_k x_k)}}{1 + e^{(b_0 + b_1 x_1 + \ldots + b_k x_k)}}, \tag{8}$$

where b_1–b_k are the coefficients for the corresponding predictors.

Although linear regression can be used to calculate probability forecasts, it has certain weaknesses. It can allow the calculation of values greater than 1 or less than 0, which are unrealistic. Linear regression also cannot account for a marked change in probability when a parameter passes beyond a threshold value or range of values, as often happens in the atmosphere. Output from a logistic regression equation is bounded between 0 and 1. It allows for marked changes in probability as predictor values exceed a threshold, or for nearly linear response to the predictor if that is appropriate.

Figure 6 illustrates the differences between linear and logistic regression using an idealized single-predictor example. Assuming the predictor values increase to the right, one can see that the probability of a predictand event occurring increases as the value of the predictor increases. The linear relationship between the predictand and predictor values is shown by the dashed line and the non-linear logistic relationship by the solid curve. For predictor values at the high and low ends of the x-axis, the linear regression predicts probabilities greater than 1 and less than 0, respectively. From Equation 8, the value of y approaches 1 as the value of $(b_0 + b_1x_1 + \ldots + b_kx_k)$ approaches $+\infty$, and approaches 0 as the value of $(b_0 + b_1x_1 + \ldots + b_kx_k)$ approaches $-\infty$. As a result, the logistic regression curve approaches 0 and 1 but can never go beyond those bounds.

Figure 6 also shows a rather distinct change in the frequency of occurrence of a predictand event at the midpoint of the predictor values. The slope of the logistic regression curve increases at the midpoint, responding to the predictand event frequency change. The linear regression curve cannot change slope to respond to such changes. The result when using logistic regression tends to be more realistic, yielding more accurate probabilities of predictand event occurrence than linear regression in situations of sharp changes in predictand event frequencies.

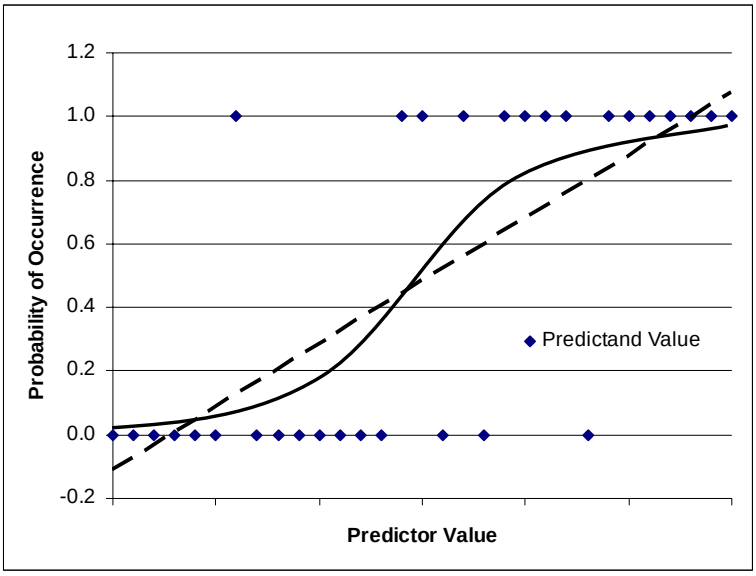

Figure 6. Illustration of linear (dashed line) vs. logistic (solid curve) regression probability forecasting for a binary predictand and one predictor. The blue diamonds represent the predictand values at certain predictor values. The forecast probability values are along the y-axis. The predictor values along the x-axis are assumed to increase monotonically to the right (similar to Wilks [1995] Figure 6.10).

4.2.2. Residual Deviance Calculation

Before discussing the specifics of predictor selection, the reader should have a general understanding of a parameter called residual deviance. The contribution of each candidate predictor to the reduction in variance was determined by this parameter. The residual deviance serves the same role in logistic regression as does the residual sum of squares in a linear regression (Insightful Corporation 2001b). Menard (2000) examined several methods that help determine the amount of predictand variance explained by predictors in logistic regression equations. The preferred method in that study was determining the percentage drop in the residual deviance when a new predictor was added. Therefore, it was the method employed in this study.

To obtain the residual deviance, each equation was input to the S-PLUS function ANOVA (analysis of variance), which output residual deviance for a NULL equation and for each of the predictors in the equation. A NULL model has only one predictor, x_0, whose value is 1. Assuming b_0 is equal to b_0x_0 in this case, this results in b_0 as the only term in the exponents of Equation 8. The NULL equation b_0 values ranged from -1.19 in May to 0.52 in July. Putting these values in Equation 8 results in y = 0.23 for May and y = 0.63 for July. These values are equal to the lightning climatology for these months (see Table 3, right-most column). In essence, the NULL equation predicts the monthly climatology as found in the dependent dataset. The residual deviance for the NULL equation is calculated with the general equation

$$\text{Residual Deviance} = -2*[\log(y)*(\#\text{yes}) + \log(1-y)*(\#\text{no})], \tag{9}$$

where y is the probability calculated by Equation 8, #yes is the number of days with lightning and #no is the number of days with no lightning. With the above values for b_0, the NULL residual deviance was 395.7 for May and 467.6 for July. Equation 9 becomes more complex when other predictors are added. As each predictor is added, the residual deviance is reduced from the NULL value.

4.2.3. Predictor Selection

As stated earlier, predictor selection was conducted for each individual month using the development dataset. The predictors were selected and equations developed using the S-PLUS software, which has functions specifically designed to create logistic regression equations and test how each individual predictor contributes to the reduction in variance of the predictand.

4.2.3.1. Residual Deviance Check

The values for the predictor coefficients in a logistic regression equation (Equation 8) cannot be solved analytically, but must be estimated using computationally intensive iterative techniques (Wilks 1995) that are much too cumbersome to be done manually. The procedure to develop a logistic regression equation outlined in the S-PLUS User's Manual was used to create the equations. The candidate predictors were added to a logistic regression equation one-by-one and their contribution to the reduction in residual deviance noted. While more automatic predictor selection methods in S-PLUS could have been employed, the manual process used here allowed for more control over understanding exactly how each individual predictor contributed to the reduction in residual deviance individually and in combination with other predictors. It was also facilitated by the relatively small number of candidate predictors available for selection.

Predictor selection began by using each of the 13 candidate predictors as a lone predictor in Equation 8, resulting in 13 single-predictor logistic regression equations. The reduction in residual deviance from each single predictor was measured from that of the NULL model. The candidate predictor that affected the largest reduction in the residual deviance was chosen as the first predictor in the equation. Next, the other 12 candidate predictors were added individually with the first predictor creating a set of 12 two-predictor equations. The second candidate predictor that reduced the residual deviance by the largest amount in combination with the first was chosen as the second predictor. The remaining 11 candidate predictors were added individually to the new two-predictor equation, and the predictor that reduced the remaining residual deviance by the most was chosen as the third predictor. This iterative process continued for all 13 predictors. Figure 7 shows the percent reduction in residual deviance from the NULL model as each predictor was added for the month of July. The TT reduced the residual deviance by the most (11%) and was, therefore, the first predictor in the July equation. The second predictor was persistence, which accounted for an additional 5% reduction in residual deviance. The RH was the third predictor, reducing the residual deviance by 2%, and so on.

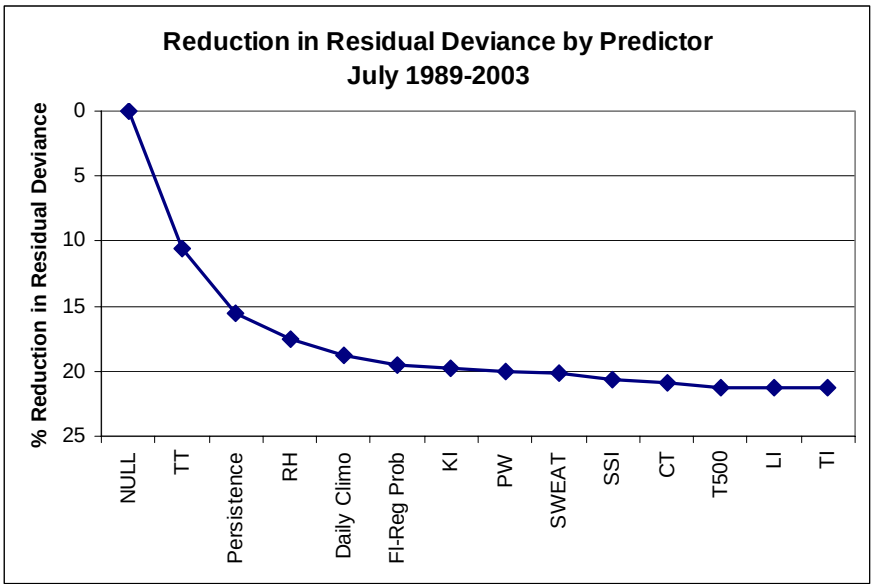

Figure 7. The total percent reduction in residual deviance from that of the NULL model as each predictor was added to the equation using the July development dataset.

Optimally, this procedure would have resulted in one 13-predictor equation. However, there were times that the residual deviance explained by one predictor was very similar or equal to that of another predictor in the same iteration. Two equations were created at that point, one for each of the predictors with similar reductions in residual deviance. Each equation continued through the iteration process from that point. Between 2 and 3 equations for each month were created in this manner. In every instance, the first three candidate predictors chosen in each month were large outliers in terms of their reduction of the residual deviance. Equation "splitting" did not take place until at least the fourth predictor was added. For July, predictor selection was decisive for each predictor up to the flow regime lightning probability (fifth iteration). At the sixth iteration, KI and SSI both reduced the residual deviance by 0.3%. Two equations were created: one containing KI as the sixth predictor and the other containing SSI as the sixth predictor. The example in Figure 7 shows the 13-predictor equation with KI as the sixth predictor.

Using all the predictors would likely result in over-fitting the regression equations such that they would perform well with the development data but no other datasets. Several equations were developed for each month and tested to determine the point at which adding another predictor would result in over-fitting. First, a threshold of 0.5% in the reduction of residual deviance was chosen as a predictor cutoff point, with the assumption that the predictors causing a residual deviance reduction ≥ 0.5% would be retained and those causing a reduction of < 0.5% would be rejected for the final equation. The 0.5% reduction threshold coincided close to where the slope of the residual deviance reduction curve began to flatten, as can be seen in Figure 7 in the vicinity of KI and the flow regime probability (Fl-Reg Prob).

4.2.3.2. Cutoff Threshold Check

Next, equations were created that included predictors with residual deviance reductions close to and surrounding the 0.5% threshold to determine where the most appropriate predictor cutoff existed. A base equation was designated that contained only predictors that reduced the residual deviance by ≥ 0.5%. For example, the base equation in July contained five predictors:

- TT,
- Persistence,
- RH,
- Daily lightning climatology (Daily Climo in Figure 7), and
- Flow regime probability (Fl-Reg Prob in Figure 7).

Equations were created with one, two, and three less predictors; and one, two, and three more predictors than the base equation according to residual deviance reduction rank. For July, the KI-SSI split occurred immediately below the 0.5% threshold. This resulted in two equations with one more predictor added to the base equation (base equation plus KI, base equation plus SSI), two equations with two more predictors, and two equations with three more predictors. The total number of equations tested for July was 12 including the base equation and the two 13-predictor equations. The number of equations tested for each month was between 8 and 15 depending on the number of splits, if any.

These equations were used to create probability forecasts for the development and verification datasets. For July, that meant 12 sets of probabilities from the development data and 12 sets from the verification data. The mean-squared error (MSE) for the probability forecasts from each equation and dataset was calculated to determine equation performance. The MSE is given by the equation

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 , \tag{9}$$

where n is the number of forecast/observation pairs, p_i is the probability calculated from the equation, and o_i is the corresponding binary lightning observation (Wilks 1995). The MSE is 0 for a perfect forecast. The dependent dataset was used to confirm that the equations would perform well on the dataset from which they were developed. The 13-predictor equations produced the lowest MSEs (0.16 – 0.23) with the development dataset as expected, indicating the best performance compared to equations with fewer predictors. The MSE values from the 13-predictor equations using the verification dataset were higher than those produced by the other equations, again as expected. This indicated that the 13-predictor equations were over-fitted, performing well only on the dataset from which they were developed. The equation producing the lowest MSE with the verification data was chosen as the final equation for each month. For July, the base equation defined in the previous paragraph produced the lowest MSE with the verification dataset. This was also the case for May and September. The June and August equations used the base equation plus the next predictor (1-more) below the 0.5% threshold.

Table 6 shows the final predictors for each of the monthly equations in order of their contribution to the reduction in residual deviance. Three predictors stood out in all five equations:

- Flow regime lightning probabilities,
- Smoothed daily lightning climatology, and
- Persistence.

The RH was the next most common predictor occurring in four of the five equations, indicating the importance of moisture. The next most common predictors were TT and LI, occurring in two equations each. Three other predictors, TI, T_{500}, and KI, occurred in only 1 equation each.

Table 6. The final predictors for each monthly equation, in order of their contribution to the reduction in residual deviance. The predictors in red font were chosen in every month, and the predictors in blue font were chosen in four months. The other predictors in black font were chosen only once or twice.

May	***June***	***July***	***August***	***September***
TI	800–600 mb RH	TT	KI	Persistence
Flow Regime	Persistence	Persistence	Flow Regime	Flow Regime
Persistence	LI	800–600 mb RH	TT	800–600 mb RH
Daily Climatology	Flow Regime	Daily Climatology	Daily Climatology	Daily Climatology
T_{500}	Daily Climatology	Flow Regime	800–600 mb RH	LI
			Persistence	

4.3. Equation Performance

The predictors in the verification dataset were used in the equations to produce 'forecast' probabilities. These probabilities were compared with the binary lightning observations in the verification dataset using four tests that measured different aspects of forecast performance. They were the

- Brier Skill Score, which is a measure of equation performance versus other standard forecast methods,
- Distributions of the probability forecasts for days with and without lightning,
- Reliability diagram of the observed lightning frequency as a function of the forecast probability, and
- Categorical contingency table statistics.

The Brier Skill Scores were calculated for each individual month to show how each equation performs against corresponding standard forecast methods. The number of available days in each month of the verification data ranged from 48–56 (Table 5). The individual monthly samples were small, but large enough to provide a reasonable estimate of relative skill with the Brier Skill Score. The other three procedures required more data, so the available days in all months were combined into one dataset to increase the sample size.

4.3.1. Brier Skill Score

The first test of the equations was whether or not they showed an improvement in skill over other forecast methods. This involved calculation of the Brier Skill Score (SS) as

$$SS = \left(\frac{MSE_{eqn} - MSE_{ref}}{MSE_{perfect} - MSE_{ref}} \right) * 100 \text{ (Wilks 1995)}, \quad (10)$$

where MSE_{eqn} is the MSE of the equation being tested, MSE_{ref} is that for the reference forecast method against which the equation is being tested, and $MSE_{perfect}$ is the MSE of a perfect forecast, which is always 0. The SS represents a percent improvement (degradation) in skill of the equation over the reference forecast when it is positive (negative). Four methods were used for the reference forecasts:

- Smoothed daily lightning climatology (Figure 4a),
- Monthly lightning climatology (Table 3),
- Flow regime lightning probabilities (Table 3), and
- Persistence.

The SS values for each of the monthly equations are shown in Table 7. All SS values are positive, indicating that the equations produced an increase in skill over all four reference forecasts in all months. The percent improvement over persistence was the largest for all reference forecasts except for May, in which it was the second largest. This improvement is significant in that persistence is well-known to outperform NPTI. It should follow that the equations would outperform NPTI as well. The smallest percent improvements were over the probabilities based on flow regime, except for May in which it was the largest.

Table 7. The percent improvement in skill of the logistic regression equation forecasts over the reference forecasts of persistence, daily and monthly lightning climatologies, and flow regime probabilities. These results were calculated using the verification data.

Forecast Method	*May*	*Jun*	*Jul*	*Aug*	*Sep*
Persistence	31	53	38	42	43
Daily Climatology	27	18	27	12	21
Monthly Climatology	34	20	27	16	22
Flow Regime	34	13	20	8	21

4.3.2. Probability Distributions

In the next test, the equation probability forecasts from all days in the verification dataset were stratified by lightning and non-lightning days. The distribution of the probability values was calculated for each stratification. Figure 8 shows the two probability distributions for lightning days, represented by the pink curve, and non-lightning days, represented by the blue curve. For good performance, one would expect the blue curve to have a maximum in the lower probability values decreasing to a minimum at higher probability values, and the pink curve to have a minimum in the lower probability values increasing to a maximum at the higher values.

The blue curve for non-lightning days has a peak near 40% at probability values of 0.2 then decreases to almost 10% at 0.6, followed by a small rise to 20% at 0.8, then a decrease to just below 10% at 1. It shows a high percentage occurrence of non-lightning events at the lower probabilities and decreasing toward the higher probabilities as one would expect for good performance. However, the secondary maximum at 0.8 suggests an increased possibility of false alarm forecasts. This secondary maximum could be caused by the fact that the equations do not take all factors into account that influence thunderstorm development. It is also possible that lightning could have occurred in the vicinity on those days but not within the spatial area as defined in Figure 2.

The pink curve for lightning days shows low frequencies slowly increasing to 10% up to a probability value of 0.5, then quickly increasing to just less than 40% at 0.8 and staying at that level through a probability of 1. This indicates that the equations perform well for lightning days. This curve also increases above the blue curve at 0.55 probability. This would show that probability forecasts above 0.55 are more likely to be calculated on lightning days as opposed to non-lightning days.

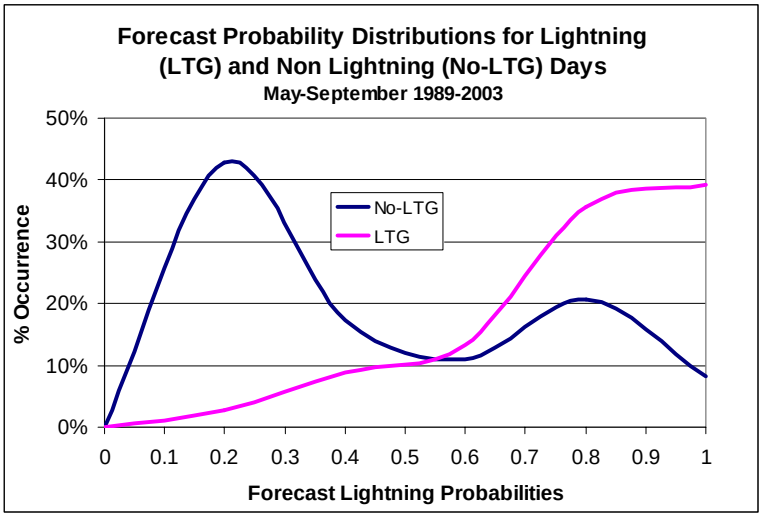

Figure 8. The forecast probability distributions for lightning (pink) and non-lightning (blue) days in the verification data. The y-axis values represent the frequency of occurrence of each probability value, and the values on the x-axis represent the forecast probability values output by the equations.

4.3.3. Reliability Diagram

Figure 9 shows the reliability diagram for probability forecasts using the verification dataset. Where the blue curve is below the pink curve, the equations were over-forecasting lightning occurrence, and where the blue curve is above the pink curve, the equations were under-forecasting lightning occurrence. Most blue curve values are below the pink curve but within 10% of the pink curve values, indicating only slight over-forecasting. The exceptions are at 0.4, 0.5, and 0.8 forecast probabilities. Lightning occurred 55% of the time when a probability of 0.4 was forecast, indicating an under-forecast of 15%. When a probability of 0.5 was forecast, lightning occurred only 38% of the time, indicating an over-forecast of 12%. A large over-forecast existed for a probability forecast of 0.8, for which lightning occurred only 50% of the time. A detailed examination of the data revealed no clear pattern of why there was such a discrepancy at this value. It could be an artifact of the dataset; a larger dataset may not exhibit such behavior. It may also be indicative of the false alarm rate at the higher forecast probabilities, although the blue curve values for all other probabilities above 0.5 were well within 10% of the corresponding pink curve values.

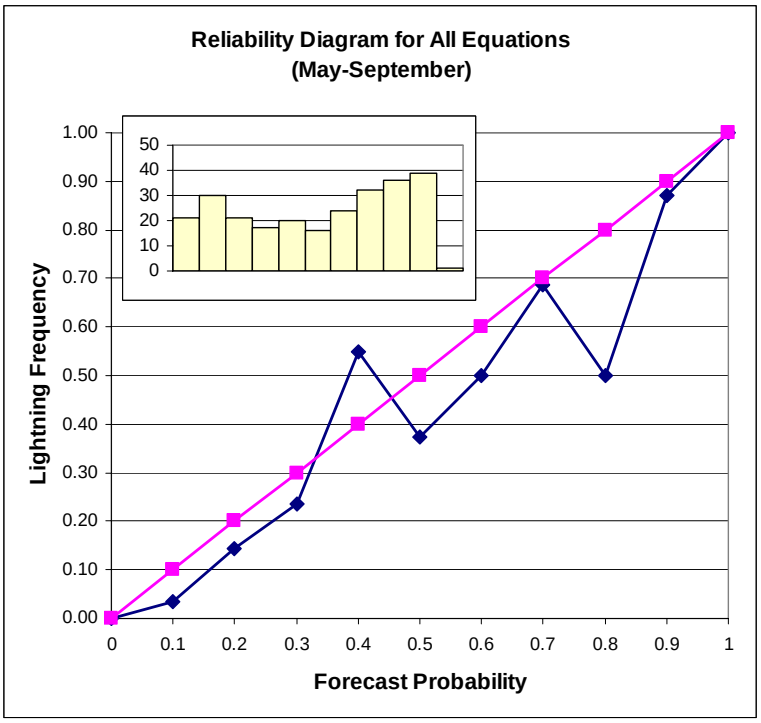

Figure 9. The reliability diagram of the probability forecasts for all months. The pink curve represents perfect reliability and the blue curve represents the probability forecast reliability. The inset rectangle is the histogram showing the number of observations in each probability range.

4.3.4. Contingency Table Statistics

The final test was to create a contingency table and calculate the probability of detection (POD), false alarm ratio (FAR), hit rate (HR), critical success index (CSI), and the Heidke and Kuipers skill scores (HSS and KSS, respectively). Table 8 shows an example of the contingency table with equations for the accuracy measures and skill scores (Wilks 1995). An event is counted in

- Cell a if it is forecast and observed (a forecast hit),
- Cell b if it is forecast and not observed (a false alarm forecast),
- Cell c if it is not forecast but observed (a forecast miss), and
- Cell d if it is not forecast and not observed (a forecast hit).

The HR is the percentage of correct yes or no forecasts, and the POD is the percentage of 'yes' forecasts in the number of 'yes' observations. The FAR is the percentage of 'no' observations in the number of 'yes' forecasts. The CSI is the percentage of correct 'yes' forecasts in the sum of all 'yes' forecasts and observations. The HSS and KSS values represent the forecast performance compared to a reference random forecast, the difference being that in the KSS the random forecast is constrained to be unbiased.

This type of forecast verification is most appropriate for categorical, or binary, forecasts in which a phenomenon is forecast to occur or not. It is a less appropriate method for probability forecasts that express levels of uncertainty in which no probability value in the range 0 – 1 is necessarily wrong or right (Wilks 1995). Nonetheless, it is a familiar and easily understood method that can shed light on forecast performance provided an appropriate probability threshold value is defined above which the forecast will be considered 'yes' and below which the forecast will be considered 'no'.

Table 8. Basic contingency table for calculating categorical accuracy measures and skill scores (Wilks 1995). The equations for the accuracy measures and skill scores are in the bottom row.

		Observation	
		Yes	No
Probability Forecast	Yes	**a**	**b**
	No	**c**	**d**

$n = a + b + c + d$

$POD = a/(a+c)$ $FAR = b/(a+b)$ $HR = (a+d)/n$

$CSI = a/(a+b+c)$ $KSS = (ad - bc)/[(a+b)(b+d)]$

$HSS = 2(ad - bc)/[(a+c)(c+d) + (a+b)(b+d)]$

The proper threshold value depends on the forecast decision issue to which the user will apply the forecast (Wilks 1995). The goal of this task was to create a system of equations that outperforms persistence, which has been shown to outperform NPTI. Everitt produced graphs of contingency table values versus equation probability cutoff values along with the contingency table cell values for persistence in order to determine an optimum cutoff value at which the accuracy measures and skill scores indicated better forecast skill than persistence. Everitt's procedure was followed here.

Figure 10 shows the contingency table values for persistence and a range of equation output probability values from 0.1–0.9 in increments of 0.01. The persistence forecast was purely categorical in that it was a binary forecast for a binary predictand, so it had only one set of contingency table values. They are designated by the horizontal straight lines in Figure 10. Contingency table values for each of the probability values were determined by assuming all probabilities at or above a specific cutoff value were 'yes' forecasts, and all values below were 'no' forecasts. The contingency table values at each probability cutoff value are shown by the curves with symbols in the graph, color-matched to the same contingency table cell for the persistence forecast. A range of probability cutoff values were then isolated such that all four cell values were optimized to be better than persistence. The objective was to have more forecast hits and fewer false alarms and misses than persistence. This resulted in a probability cutoff range of 0.59–0.63, which is outlined by the vertical black lines in Figure 10.

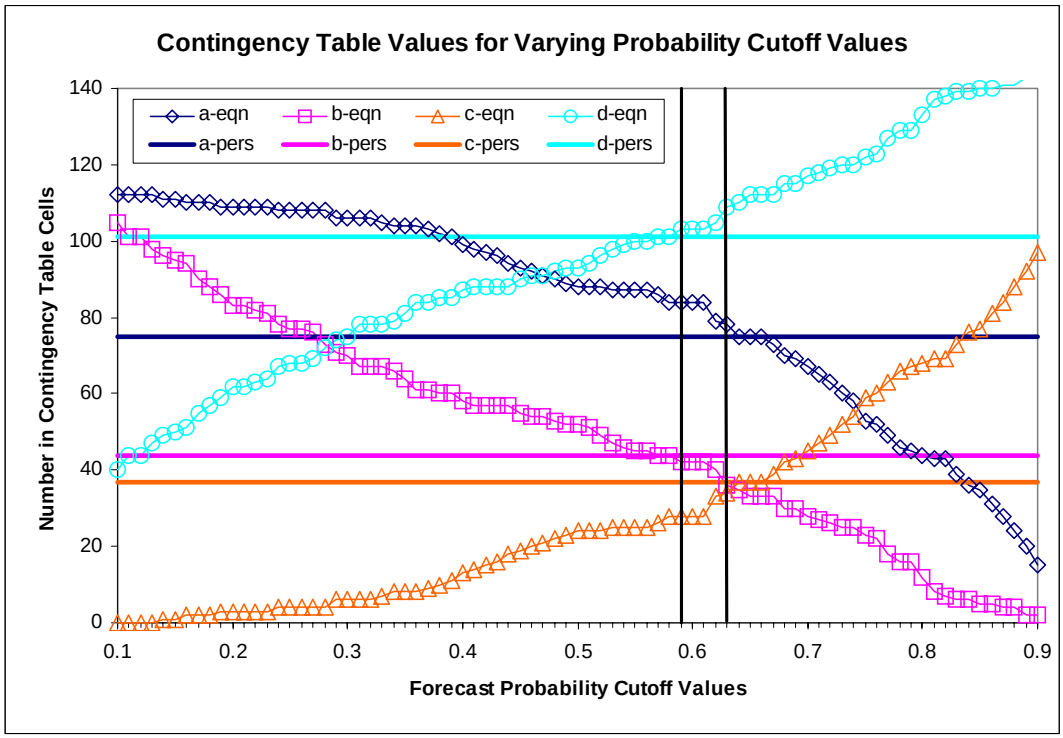

Figure 10. Graph showing the values in the four contingency table cells in Table 8 for the range of probability values 0.1–0.9 in increments of 0.01. Dark blue represents values in cell a, purple represents values in cell b, orange represents values in cell c, and cyan represents values in cell d. The horizontal straight lines represent the persistence forecast (pers) and the curves with symbols represent the equation forecasts (eqn) as shown in the legend. The vertical black lines show upper and lower bounds of the probability range of where all cell values are maximized or minimized such that the accuracy measures and skill scores will show better performance than persistence.

The accuracy measures and skill scores were calculated for each probability in the range 0.59–0.63 to assist in determining which value should be the cutoff. Because it had the maximum HR and was in the middle of the range of values, 0.61 was chosen as the cutoff value. All probabilities at or above 0.61 were considered 'yes' forecasts and all those below 0.61 were considered 'no' forecasts for the contingency table.

The contingency table cell, accuracy measure, and skill score values for the probability cutoff of 0.61 and persistence are shown in Table 9. The HR, POD, and CSI are 100% for a perfect forecast and 0% for the worst possible forecast, and vice versa for FAR. The HSS and KSS are 1 for perfect forecasts, 0 for performance equivalent to a random forecast, and < 0 for performance worse than that of a random forecast. The HR and POD values were relatively high at 73% and 75%, respectively, for the equations. The FAR was relatively low at 33%, but still high enough to be considered as a factor when using the equations for forecasting lightning occurrence. The CSI value is better than 0.5, but not necessarily an indicator of good performance. The HSS and KSS values are not high, but are positive and indicate performance better than that of random forecasts. Comparing the equation measures and scores to those of persistence, it can be seen that the equations outperform persistence in every value.

Table 9. Contingency table for the cutoff probability value of 0.61. Probability values ≥ 0.61 were considered a 'yes' forecast, and values < 0.61 were considered a 'no' forecast for lightning occurrence. Accuracy measure and skill score values for the equations and persistence are shown beneath the contingency table.

		Observation	
		Yes	No
Probability Forecast (0.61)	Yes	**84**	**42**
	No	**28**	**103**

Equations:

POD = **75%** FAR = **33%** HR = **73%**

CSI = **0.55** HSS = **0.45** KSS = **0.46**

Persistence:

POD = **67%** FAR = **37%** HR = **68%**

CSI = **0.48** HSS = **0.36** KSS = **0.34**

4.3.5. Equation Performance Summary

All four equation performance measures indicated that the equations showed an increase in skill over daily and monthly lightning climatology, persistence, and the flow regime lightning probabilities. The equations also demonstrated good reliability, an ability to distinguish between non-lightning and lightning days, and improved standard categorical accuracy measures and skill scores over persistence. The increase in skill over persistence seen in Table 7 is important since this method has been shown to outperform the NPTI, the current objective tool used for daily thunderstorm forecasting.

Three of the tests, however, showed a tendency for the equations to over-forecast the probability of lightning occurrence, i.e. high probability values were calculated when no lightning was observed by CGLSS in the area of interest on a considerable number of days. The explanation for the equations' tendency to over-forecast as seen in Figures 8 and 9 and the FAR in Table 9 was not fully explored. It could be that cloud-to-ground lightning occurred near but not within the area of interest on those days. It is also possible that LDAR signals existed over the area with no CGLSS signatures. As stated in Section 2, LDAR data were not used in this study due to the considerable size of the datasets and the shorter POR available. Finally, it is possible that certain atmospheric parameters acting to suppress convection on those days are not represented by the predictors in the equations. Forecasters should keep in mind the equations' slight tendency to over-forecast lightning when using the tool.

5. Graphical User Interface

Results from the equation testing in Section 4.3 indicate that the equations developed in this task outperform persistence and, therefore, the NPTI. Based on these results, the 45 WS requested that the equations be transitioned into operations, and that a GUI be developed to facilitate user-friendly input and fast output. A GUI was developed using Microsoft® Excel© Visual Basic®. The 45 WS was involved in the GUI development by providing comments and suggestions on the design to ensure that the final product addressed their operational needs.

The GUI was built within an Excel workbook. It accesses data in specific worksheets based on user input. The GUI itself has three basic dialog boxes. The first asks for the date, the second asks for equation predictor values, and the third displays the equation output. The workbook values and GUI code are all password-protected and cannot be changed by the user.

5.1. Excel Workbook

The Excel workbook in which the GUI resides contains six worksheets. The first worksheet contains brief instructions on how to start and use the GUI. It is recommended that first-time users read these instructions in their entirety before using the GUI. The other five worksheets contain information for each individual month. The information on these sheets includes the

- Predictor names and their coefficients in the equations,
- Flow regime names and their probabilities of lightning occurrence,
- Climatological lightning probability for each day,
- Minimum, maximum, median, mean, and first and third quartile values of the observed sounding stability indices,
- Range of valid values in the GUI for the stability indices, and
- Stability index values associated with convection.

The first worksheet, named Introduction, is displayed automatically upon opening the Excel file. There are three ways to initiate the GUI, all explained at the beginning of the instructions in the Introduction worksheet. When the GUI is initiated, the first dialog box requesting the date is displayed. After choosing a month and day in this dialog box and continuing, the worksheet corresponding to the chosen month is displayed along with the second dialog box. This allows the user to view all the possible parameter values as described in the above list for use in a particular month's equation. When the user is finished and exits out of all the dialog boxes, the Introduction worksheet will be displayed again before closing the file.

5.2. Current Date Dialog Box

When the user initiates the GUI, a dialog box is displayed that queries the user for the current month and day, shown in Figure 11. A drop-down list is shown for each parameter by clicking on the down-arrow to the right of the text boxes containing the Month and Day values. Choosing the month determines which equation will be used, and choosing the day determines which daily lightning climatology value will be used as a predictor in the equation. The user must choose a value from the Month drop-down list, but has the option of entering a Day value manually or through the Day drop-down list. The Day drop-down list will only have as many choices as there are days in the month. If a user inputs a day value manually that does not exist in a particular month, e.g. 31 for June, an error message will be displayed. It is important to choose the correct month and day as these values are used to determine what daily lightning climatology value will be used in the equations.

Choosing the "Continue…" button causes the equation parameter dialog box and the worksheet for the chosen month to be displayed. Choosing the "Cancel" button will close the GUI and return the worksheet display to the Introduction worksheet.

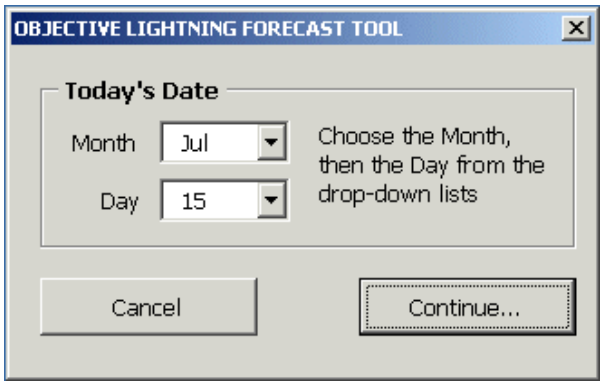

Figure 11. The first dialog box in the GUI queries the user for the Month and Day values. Month and Day are chosen by clicking on the down arrows next to each and choosing from the drop-down lists. The Cancel button exits from the GUI, the Continue button brings up the next dialog box.

5.3. Equation Predictor Dialog Boxes

After clicking the 'Continue…' button in the current date dialog box, an equation predictor dialog box is displayed in which predictor values can be chosen. There are five distinct equation predictor dialog boxes, one for each month since each has a different equation. The dialog boxes for each month are displayed in Figures 12 – 16. Each dialog box contains elements that must be changed by the user, either by making a choice between two or more elements or entering a value. All choices must be made and values entered before a probability can be calculated. Choosing the "Calculate Probability…" button will cause calculation of the equation using the choices and values input by the user, and output from the equation will be displayed in the equation output dialog box. Choosing the "New Date" button will close the equation predictor dialog box and return control to the date dialog box.

5.3.1. Persistence and Flow Regime

There are two features common to all five equation predictor dialog boxes: one frame titled Persistence and another frame titled Flow Regime. The first is the choice for persistence, whether or not lightning occurred in the area the previous day. The user will choose 'Yes' or 'No' by clicking in the white circle next to the choice. The default choice is 'Yes'. The second is the choice for the flow regime of the day. The user determines the flow regime for the day, then clicks in the white circle next to the appropriate choice. The default choice is for southwest (SW) flow. Note that for May and September, there are two southeast (SE) flow regimes (Figures 12 and 16), while for June, July, and August there is only one SE flow regime (Figures 13 – 15). The climatological characteristics of the SE flow regimes in the latter group were sufficiently similar that the two regimes were combined into one (Lambert 2004a). The user can choose only one item under Persistence and only one item under Flow Regime.

As stated in Section 2, the 1200 UTC soundings at MIA, TBW, and JAX were used to determine the flow regime. These soundings cannot be used to determine the flow regime of the day in real-time operations since the morning briefing occurs at 1100 UTC. It is not recommended that the forecasters use the 0000 UTC soundings from the previous evening since the large-scale low level flow may be disrupted by afternoon convective circulations. There are several data sources to help forecasters determine the flow regime of the day before the briefing, including surface observations and model output.

5.3.2. Sounding Parameters

The other predictors in the equation parameter dialog boxes are values taken from the 1000 UTC XMR sounding. Their initial values are set to the climatological medians for each month in an effort to minimize forecaster effort in changing the value. The forecaster will initially see a -999 for each sounding parameter value as a signal that a value for that parameter has not yet been input. If the user forgets to input values and clicks the "Calculate Probability…" button, an error message will be triggered that tells the user to input an appropriate value for each parameter. Values for the sounding parameters come from the MIDDS Skew-T program. There are a total

of six parameters in different combinations for each month: TI, T_{500}, RH, LI, TT, and KI. All are directly available from the Skew-T program except TI and RH. The TI is calculated with the equation TI = KI – LI. The RH should be calculated using a depth-weighted average of observed relative humidities in the 800-600 mb layer of the sounding (see Equations 4 and 5). Once the values are obtained from the sounding, the user can input the values manually in the appropriate text box or use the up/down arrows to make the choice.

There are also upper and lower limits on the parameter values to ensure realistic values are entered. These limits are shown in the cells of the worksheet that is displayed (not shown) along with the equation predictor dialog box. If a value is entered that is beyond the upper or lower limit, an error message will be triggered that tells the user to input an appropriate value. The upper and lower limits along with the summary values of mean, median, minimum, maximum, and first and third quartiles for each parameter in each month are shown in Table 10. The summary values were calculated from the entire dataset in the POR 1989–2003.

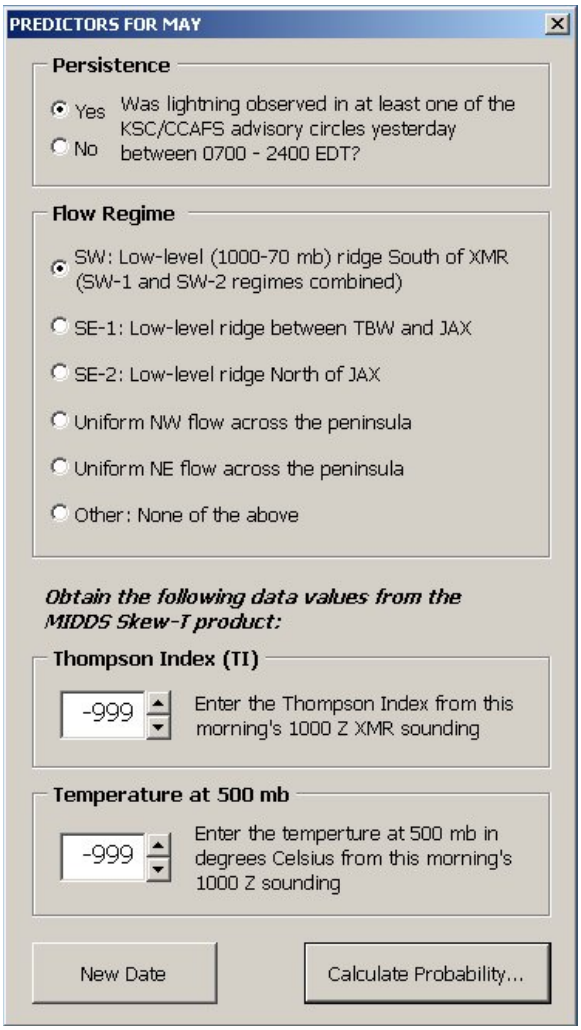

Figure 12. This dialog box contains choices for the predictors in the May equation. Persistence and Flow Regime are chosen by clicking one of the option buttons in each section. TI and T_{500} are chosen by entering their values manually or using the up/down arrows to the right of the text boxes. The "New Date" button closes this dialog box and returns control to the current date dialog box (Figure 11). The "Calculate Probability…" button displays the equation output dialog box (Section 5.4).

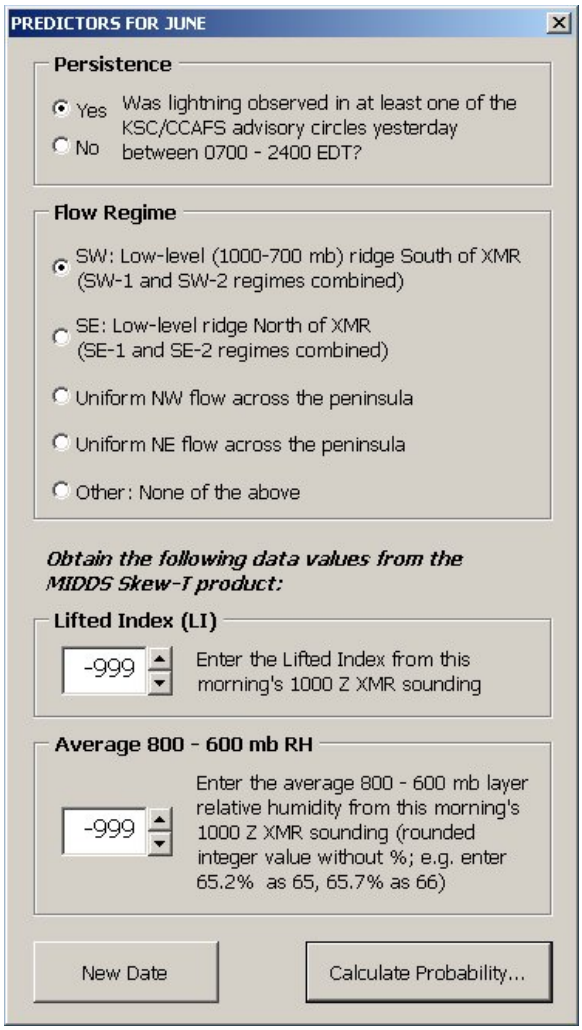

Figure 13. Same as Figure 12 except for June, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with sounding parameters LI and RH.

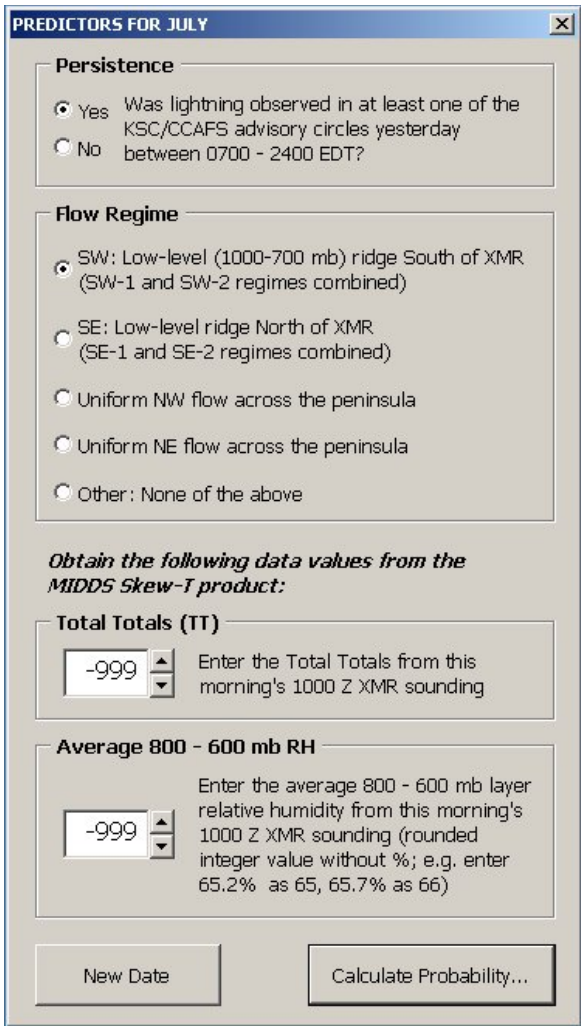

Figure 14. Same as Figure 12 except for July, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with sounding parameters TT and RH.

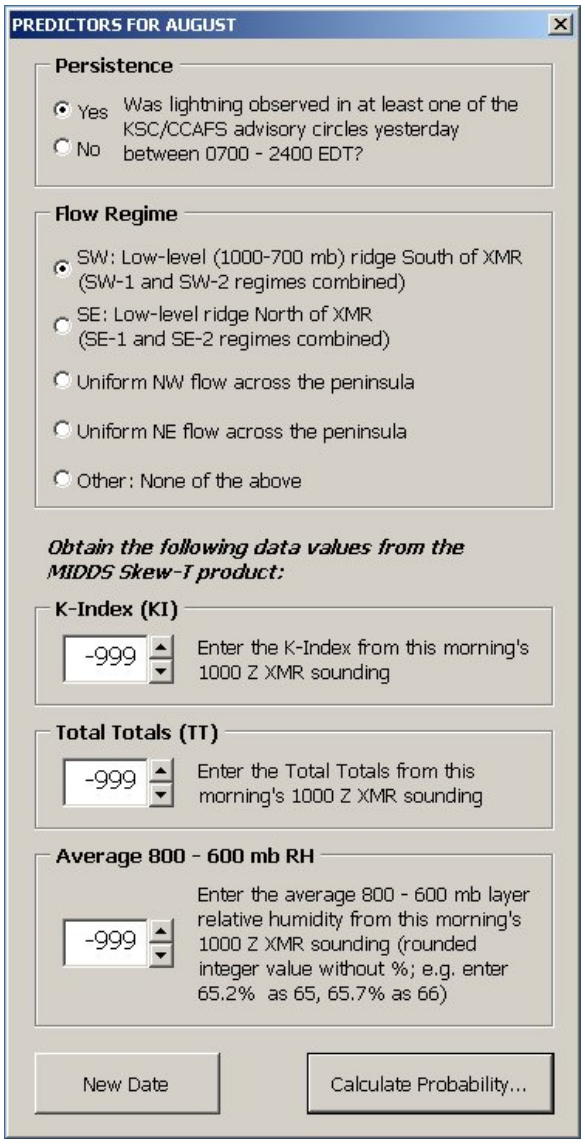

Figure 15. Same as Figure 12 except for August, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with sounding parameters KI, TT, and RH.

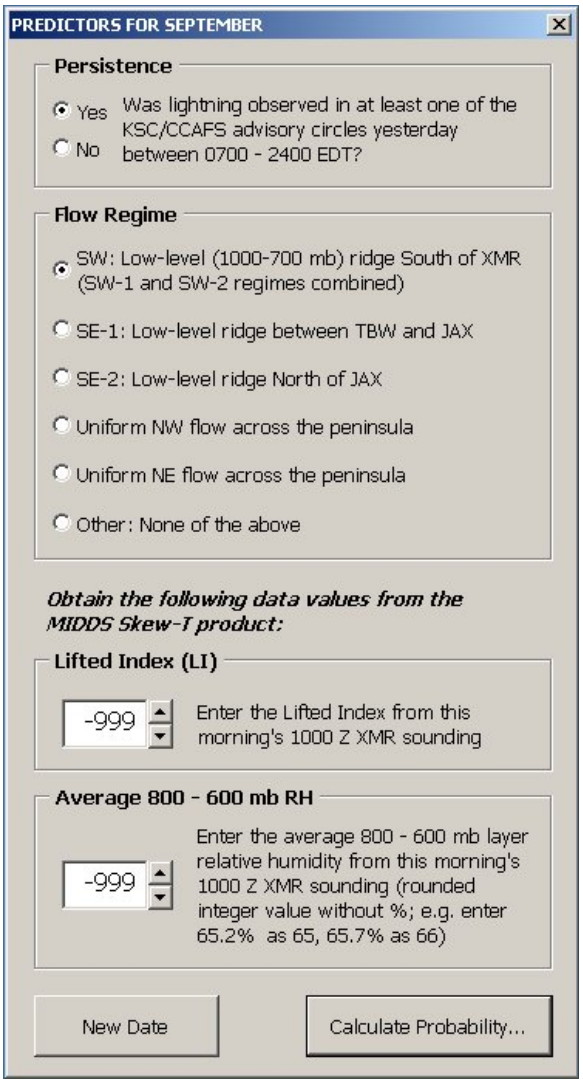

Figure 16. Same as Figure 12 except for September, and with sounding parameters LI and RH.

Table 10. Summary values for each of the predictors in the POR 1989–2003. The last two rows contain the upper and lower limits of the values allowed in the GUI.

Observed Data Summary	*May*		*June*		*July*		*August*			*September*	
	TI	**T_{500}**	**LI**	**RH**	**TT**	**RH**	**KI**	**TT**	**RH**	**LI**	**RH**
Minimum	-38	-17	-7	15	33	18	-10	26	19	-6	15
1st Quartile	8	-11	-4	45	43	46	26	42	44	-4	46
Median	17	-10	-2	62	45	62	31	44	60	-2	62
Mean	17	-10	-2	60	45	60	29	44	58	-2	60
3rd Quartile	30	-8	-1	76	47	72	34	46	73	-1	75
Maximum	44	-5	10	98	53	95	43	54	91	9	96
Data Value Range Allowed in GUI											
Minimum	-70	-30	-20	0	0	0	-50	0	0	-20	0
Maximum	70	20	30	100	70	100	60	70	100	30	100

5.4. Equation Output Dialog Box

After making all choices and entering all values in the equation predictor dialog box, the user should click on the "Calculate Probability…" button. This executes the equation and displays the third and final equation output dialog box (Figure 17). The lightning probability for the day as determined by the equation is displayed as a percentage value. When the user clicks the "Calculate Another Probability" button at the bottom, this dialog box is closed and control is returned to the equation predictor dialog box. The user can make new choices for the predictors and calculate a new probability, or click the "New Date" button and return control to the first dialog box.

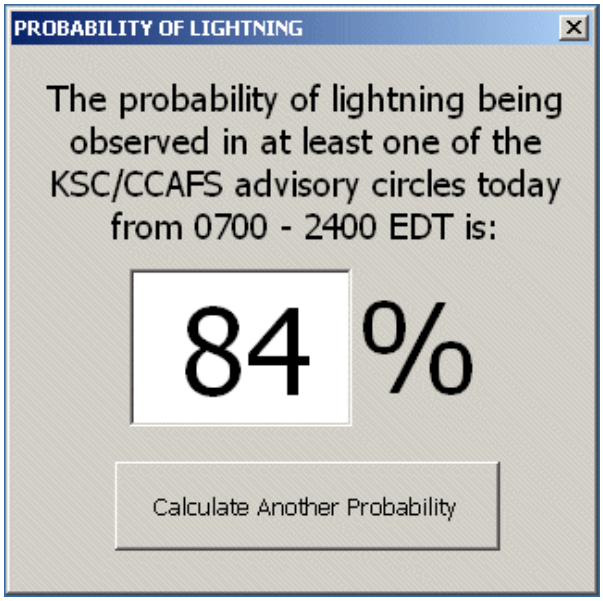

Figure 17. The equation output dialog box in the GUI containing the probability of lighting for the day based on the inputs from the date and equation predictor dialog boxes.

5.5. Predictor Responses

As an independent test, Mr. Roeder of the 45 WS generated curve and bar charts for each month to determine the response of the calculated lightning probability to changes in predictor values while holding all other predictor values constant. This was done to test the GUI for calculation errors and to determine how changes in the individual predictor values affect the output probability values. In order to use a constant daily lightning climatology value, the same day of the month was used in each monthly test. For consistency, the 15th of the month was used for all 5 months.

5.5.1. May

The response charts for May 15 are shown in Figure 18. The probability response curves due to changes in the predictors TI and T_{500} are given in Figure 18a. The flow regime and persistence values were held constant at SW and Yes, respectively. The predictor value ranges in Figure 18a covered their observed ranges in the POR (Table 10) for May. As TI was varied from -20 to 50, T_{500} was held constant at its observed May median value of -10 °C. Conversely, as T_{500} was varied from -20 to 0 °C, TI was held at its median value of 17. The curves are non-linear and shaped similarly to the logistic regression curve in Figure 6. It is also apparent that the probabilities were more sensitive to changes in T_{500} than in TI. In Table 10, the minimum observed value for TI in May was -38, but the minimum TI value in the chart is -20. The x-axis range did not go below -20 because it was obvious from the slope of the curve that the probability values would change little beyond that point.

The bar chart in Figure 18b shows the alternate case of varying flow regime and persistence with TI and T_{500} held constant at their median values. The SE-1 flow regime produced the highest probability, and the probabilities were higher for every flow regime when persistence = Yes. The probability values are quite low for all flow regimes and both persistence categories, ranging from 4% (NE, No) to 41% (SE-1, Yes). This is likely an artifact of the lightning climatology for May. Lightning occurred on only 97 of the 420 available days in May, yielding ~23% for the monthly climatology of lightning occurrence. The median value for TI is below 20, which is the threshold above which thunderstorm formation becomes probable. Even when TI = 20 in Figure 18a, the probability is still only 37%. Even though the median value for T_{500} is conducive for thunderstorm formation, this predictor contributed least to the reduction in residual deviance and has a relatively small effect on the probability outcome. In Figure 18a, T_{500} = -10 °C yielded ~30% probability of lightning occurrence. It follows that for days that exhibit values typical of May climatology, the calculated probabilities will tend to be low.

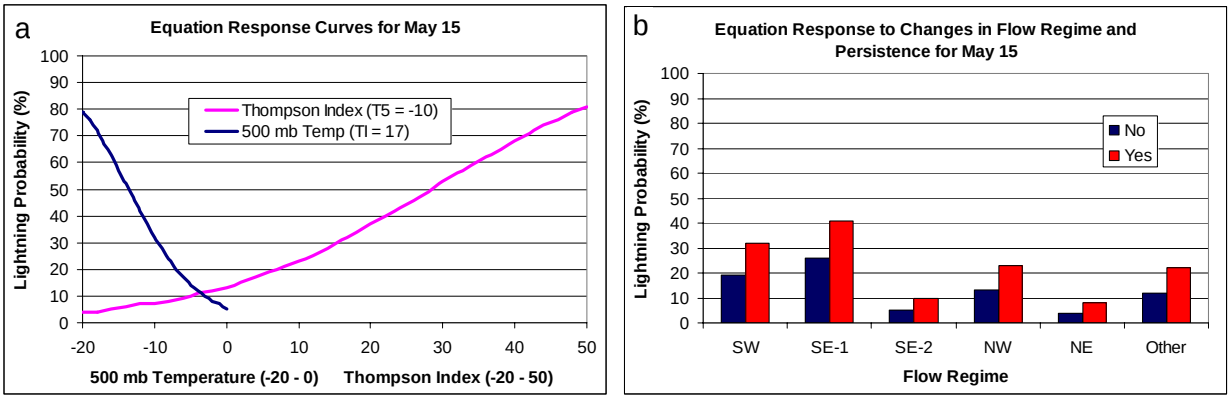

Figure 18. Equation response charts for May 15: (a) change in probability due to changes in values of TI and T_{500} with flow regime = SW, persistence = Yes, TI = 17 when T_{500} was varied from -20 to 0 °C (blue curve), and T_{500} = -10 °C when TI was varied from -20 to 50 (purple curve); (b) changes in probability due to changes in flow regime and persistence with TI = 17 and T_{500} = -10 °C. The red bars are for persistence = Yes and the blue bars for persistence = No.

5.5.2. June

The response charts for June 15 are shown in Figure 19. The probability response curves due to changes in the predictors LI and RH are given in Figure 19a. The flow regime and persistence values were held constant at SW and Yes, respectively. The predictor value ranges in Figure 19a covered their observed ranges in the POR for June (Table 10). As LI was varied from -10 to 10, RH was held constant at its observed June median value of 62%. Conversely, as RH was varied from 15 to 100%, LI was held at its median value of -2. The probabilities appear more sensitive to changes in LI than RH. The curves are non-linear, but do not approach the lower probability values asymptotically as do the curves for May. The lowest probability values are slightly greater than 20% for both predictors causing the curves to be truncated before reaching probabilities closer to 0.

The bar chart in Figure 19b shows the alternate case of varying flow regime and persistence with LI and RH held constant at their median values. The SW flow regime produced the highest probabilities. The other flow regimes were similar to each other except for NE, which had the lowest probabilities for both persistence categories. The probabilities were noticeably higher for every flow regime when persistence = Yes. Persistence ranked second in its reduction of residual deviance and, as such, had a large effect on the calculated probability. Overall, the probability values are much higher than the corresponding values for May. Unlike the low occurrence of convection in May, the monthly lightning climatology for June was 57%. The flow regime ranked fourth in the equation, which would indicate a minimal effect on the calculated probability. However, the predictor value for the NE flow regime is sufficiently small as to significantly decrease the probability of lightning occurrence when present.

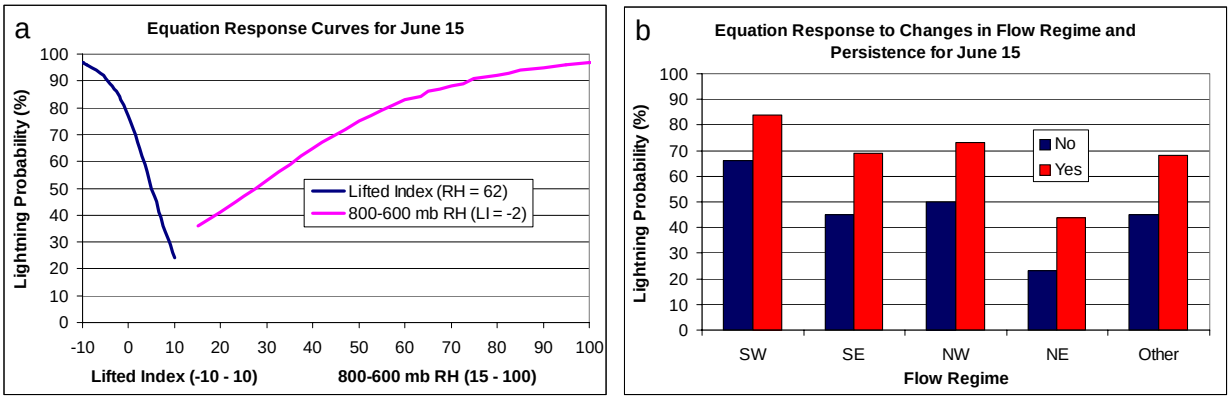

Figure 19. Equation response charts for June 15: (a) change in probability due to changes in the values of LI and RH with flow regime = SW, persistence = Yes, LI = -2 when RH was varied from 15 to 100% (purple curve), and RH = 62% when LI was varied from -10 to 10 (blue curve); (b) changes in probability due to changes in flow regime and persistence with LI = -2 and RH = 62%. The red bars are for persistence = Yes and the blue bars for persistence = No.

5.5.3. July

The response charts for July 15 are shown in Figure 20. The probability response curves due to changes in the predictors TT and RH are given in Figure 20a. The flow regime and persistence values were held constant at SW and Yes, respectively. The predictor value ranges in Figure 20a covered their observed ranges in the POR for July (Table 10). As TT was varied from 30 – 55, RH was held constant at its observed July median value of 62%. Conversely, as RH was varied from 15 to 100%, TT was held at its median value of 45. The probabilities are more sensitive to changes in TT than RH at the values used for these charts. The TT curve exhibits the same truncation issue as seen in Figure 19a for June, but the RH curve approaches 1 slowly beginning at the lowest probability of 58% for 0% humidity. The values along the RH curve also seem to indicate that changes in RH would have a small effect on the calculated probability. It ranked third among the predictors for July whereas TT ranked first in its reduction of the residual deviance.

The bar chart in Figure 20b shows the alternate case of varying flow regime and persistence with TT and RH held constant at their median values. The SW flow regime produced the highest probabilities, and the probabilities were higher for every flow regime when persistence = Yes. Overall, the probability values are quite high for each flow regime ranging from 49% (NE, No) to 84% (SW, Yes). The flow regime ranked last in terms of reduction in residual deviance, and had only a small effect on the calculated probability. The daily lightning climatology value for July 15 is 66%. Although the daily lightning climatology ranked fourth just ahead of flow regime in its reduction of residual deviance, it still shows that July was an active lightning month and calculated probability values will tend to be high.

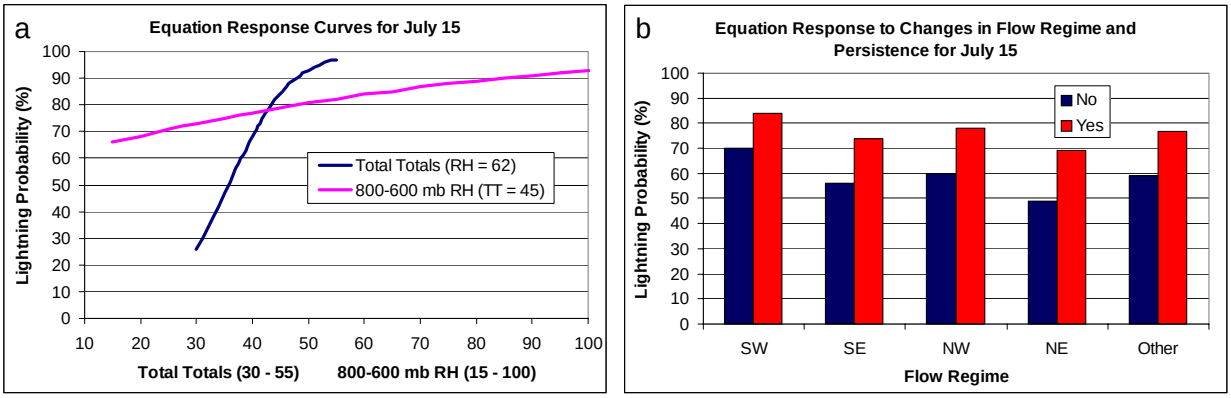

Figure 20. Equation response charts for July 15: (a) change in probability due to changes in the values of TT and RH with flow regime = SW, persistence = Yes, TT = 45 when RH was varied from 15 to 100% (purple curve), and RH = 62% when TT was varied from 30 to 55 (blue curve); (b) changes in probability due to changes in flow regime and persistence with TT = 45 and RH = 62%. The red bars are for persistence = Yes and the blue bars for persistence = No.

5.5.4. August

The response charts for August 15 are shown in Figure 21. The probability response curves due to changes in the predictors TT, KI, and RH are given in Figure 21a. The flow regime and persistence values were held constant at SW and Yes, respectively. The predictor value ranges in Figure 21a covered their observed ranges in the POR for August (Table 10). As TT was varied from 25 to 55, KI and RH were held constant at their median values of 31 and 60%, respectively. As KI was varied from -10 to 50, TT and RH were held at their median values of 44 and 60%, respectively. Finally, as RH was varied from 15 to 100%, TT and KI were held at their median values of 44 and 31, respectively. The probabilities are least sensitive to changes in RH and most sensitive to changes in TT. The TT and KI curves exhibit the same truncation issue described earlier, and the RH curve approaches 1 asymptotically beginning at the lowest probability of 66% for 0% humidity. The values along the RH curve also seem to indicate that changes in RH would have a small effect on the calculated probability. It ranked fifth among the six predictors for August whereas KI ranked first in its reduction of the residual deviance.

The bar chart in Figure 21b shows the alternate case of varying flow regime and persistence with TT, KI, and RH held constant at their median values. The SW and NW flow regimes produced the highest probabilities with SW having the largest values. The probabilities were higher for every flow regime when persistence = Yes, but only by a small amount. Persistence ranked last in its reduction of the residual deviance for August lightning occurrence and had only a small effect on the calculated probability. The NE flow regime produced the lowest probabilities by far. The flow regime lightning probability for NE flow in August is 23%. Since the flow regime probability ranked second in the equation, it had a large effect on the probability values resulting in a low value for the NE regime. Overall, the probability values exhibit a large range from 25% (NE, No) to 87% (SW, Yes).

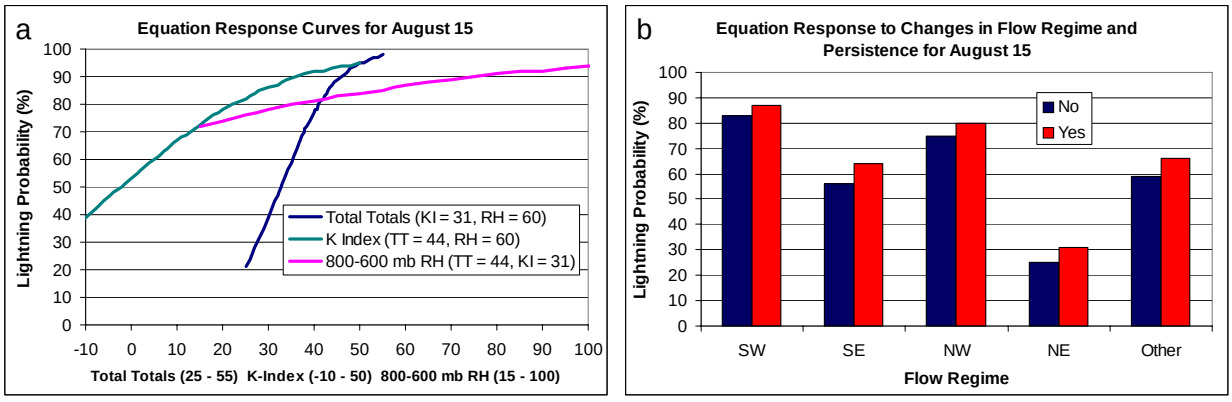

Figure 21. Equation response charts for August 15: (a) change in probability due to changes in the values of TT, KI, and RH with flow regime = SW, persistence = Yes, TT = 44 and KI = 31 when RH was varied from 15 to 100% (purple curve), TT = 44 and RH = 60% when KI was varied from -10 to 50 (green curve), and KI = 31 and RH = 60% when TT was varied from 25 to 55 (blue curve); (b) changes in probability due to changes in flow regime and persistence with TT = 44, KI = 31, and RH = 60%. The red bars are for persistence = Yes and the blue bars for persistence = No.

5.5.5. September

The response charts for September 15 are shown in Figure 22. The probability response curves due to changes in the predictors LI and RH are given in Figure 22a. The flow regime and persistence values were held constant at SW and Yes, respectively. The predictor value ranges in Figure 22a covered their observed ranges in the POR for September (Table 10). As LI was varied from -10 to 10, RH was held constant at its median value of 62%. As RH was varied from 15 to 100%, LI was held at its median value of -2. The probabilities are least sensitive to changes in RH and most sensitive to changes in LI. The curves exhibit the same truncation issue and have similar values to those in June.

The bar chart in Figure 22b shows the alternate case of varying flow regime and persistence with LI and RH held constant at their median values. The SW flow regime produced the highest probability and SE-1 the second highest, and the probabilities were higher for every flow regime when persistence = Yes. The percent increase in probability from a No to a Yes category in persistence is large for each flow regime: over 100% for SE-2, NW, NE, and Other, 50% for SW, and 70% for SE-1 flow regimes. Persistence ranked first among all predictors in its reduction of residual deviance and has a large effect on the calculated probability. The NW flow regime produced the lowest probabilities; however, there were only seven days with this flow regime in the POR for September and lightning did not occur on any of those days. There is also a large difference in probability between the flow regimes ranging from 3% (NW, No) to 75% (SW, Yes). The flow regime probability ranked second in its reduction of residual deviance in the equation. As with persistence, it follows that flow regime also had a large influence on the calculated probability. While the climatological median values of the stability parameters are at least minimally conducive for lightning occurrence, the probability values in Figure 22 are very dependent on the choice for persistence and flow regime.

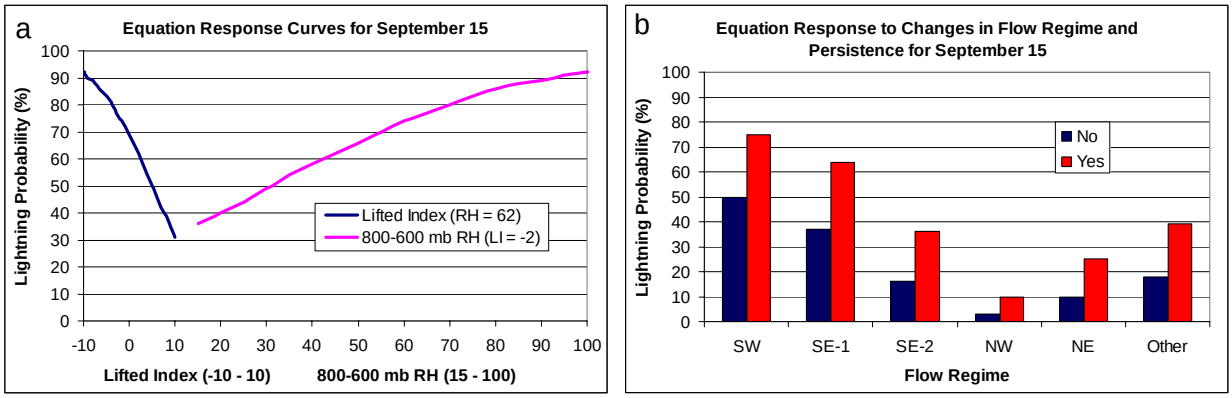

Figure 22. Equation response charts for September 15: (a) change in probability due to changes in the values of LI and RH with flow regime = SW, persistence = Yes, LI = -2 when RH was varied from 15 to 100% (purple curve), and RH = 62% when LI was varied from -10 to 10 (blue curve); (b) changes in probability due to changes in flow regime and persistence with LI = -2 and RH = 62%. The red bars are for persistence = Yes and the blue bars for persistence = No.

6. Summary and Conclusions

Five logistic regression equations were created that predict the probability of cloud-to-ground lightning occurrence for the day in the KSC/CCAFS area for each month in the warm season. These equations integrated the results from several studies over recent years to improve thunderstorm forecasting at KSC/CCAFS. All of the equations outperform persistence, which is known to outperform NPTI, the current objective tool used in 45 WS lightning forecasting operations. The equations also performed well in other tests. As a result, the new equations will be added to the current set of tools used by the 45 WS to determine the probability of lightning for their daily planning forecast.

The results from these equations are meant to be used as first-guess guidance when developing the lightning probability forecast for the day. They provide an objective base from which forecasters can use other observations, model data, consultation with other forecasters, and their own experience to create the final lightning probability for the 1100 UTC briefing.

6.1. Equation Performance Review

All four equation performance tests described in Section 4.3 showed an increase in skill over several standard forecast methods, good reliability, an ability to distinguish between non-lightning and lightning days, and improved accuracy measures and skill scores over those for persistence. Of particular interest was the increase in skill over persistence since this method has been shown to outperform the NPTI.

Three of the tests, however, showed a tendency of the equations to over-forecast the probability of lightning occurrence, i.e. high probability values were calculated when no lightning was observed by CGLSS. The explanation for the tendency to over-forecast was not fully explored. It could be that lightning occurred near the area of interest but not in it on those days. It is also possible that LDAR signals existed over the area with no CGLSS signatures. Finally, it is possible that certain atmospheric parameters acting to suppress convection on those days were not represented by the predictors in the equation.

At the request of Mr. Roeder, Mr. Castor Mendez-Vigo, a Project Emeritus Program volunteer for the 45 WS, conducted a comparison of performance between the equation output and lightning probability forecasts issued by 45 WS forecasts in the 2004 warm season. Data from September were not used due to the large number of days in which operations was suspended due to hurricane evacuations. He calculated the MSE values (Equation 9) for each forecast method, then calculated the SS (Equation 10) to determine the percent improvement or degradation of the equation forecasts compared to the 45 WS forecasts. The SS values for the equation probabilities were

- Overall: -20% (degradation compared to 45 WS forecast),
- May: -8% (degradation compared to 45 WS forecast),
- June: -28% (degradation compared to 45 WS forecast),
- July: -41% (degradation compared to 45 WS forecast), and
- August: 17% (improvement compared to 45 WS forecast).

The forecasters outperformed the equations in three of the four months, and for the whole season combined. August was the only month in which the equations performed better than the forecasters. These overall results were expected. Forecasters have other model and observational data available to make the probability forecast, as well as their own experience and the experience of other forecasters on the team. Therefore, they are able to fine tune the forecast to a probability that is more realistic than that produced by an equation that considers only five or six parameters.

The results from the equation tests underscore the importance of using this tool as a first-guess only, not as the sole source of determining the lightning probability for the day. Forecasters should keep in mind the equations' slight tendency to over-forecast lightning when using the tool. However, considering that the equations outperform NPTI and persistence, this first-guess probability will likely provide better guidance than either of these two forecast methods.

6.2. Graphical User Interface Issues

The GUI described in Section 5 interfaces with the equations and facilitates user-friendly input and fast output of the lightning probability for the day. Most of the input values are readily available to the user through the MIDDS Skew-T program. Three values are not readily available and must be determined by the user prior to using the GUI.

One of the three values is the flow regime for the day. The very first step forecasters should take before attempting to determine the flow regime is to refer to Lambert (2004a) and Section 3.2.2 of this report to understand how a flow regime was determined in this work. Since 1200 UTC soundings were used to create the flow regime climatologies but the forecast is issued by 1100 UTC, the forecasters are presented with a dilemma on what data source to use. It is not recommended that forecasters use data from the 0000 UTC soundings taken the previous evening as the larger-scale low-level flow pattern may be obscured by afternoon convection. There are several sources forecasters can used to discern the flow regime for the day:

- Pressure and wind field output from the most recent initializations of the
  - Rapid Update Cycle (RUC),
  - North American Mesoscale (NAM, formerly Eta), and
  - Global Forecast Systems (GFS) models.
- Area Forecast Discussion on the NWS MLB web site at http://www.srh.noaa.gov/mlb/forecast.html almost always discusses the position of the ridge and the low level flow for the day during the warm season, and
- Hourly surface observations of wind direction.

The surface wind directions should be used with caution as winds could be light and variable in the early morning hours. They should be used only in combination with one of the other data types in the above list. Most of the identifiable flow regimes in the warm season are due to the position of the ridge extending westward from the high pressure center over the Atlantic Ocean. The morning NWS MLB Area Forecast Discussion also offers an excellent discussion of other factors influencing the formation of convection for the day.

The other two values that must be determined are TI and RH. The TI is calculated simply with the equation

$$TI = KI - LI.$$

The values in the right-hand side of the equation, KI and LI, are readily available from the sounding on MIDDS. This variable is only used in the May equation, but was the most important in terms of reduction in residual deviance. The RH should be calculated using the same depth-weighted average used in this work, however that process could be too time-consuming in an operational setting. Forecasters can estimate this value over KSC/CCAFS from the most recent run of the Advanced Regional Prediction System (ARPS) Data Analysis System (ADAS). A contour plot of the 850–650 mb RH generated by ADAS is posted on the NWS MLB web site at

http://www.srh.noaa.gov/mlb/ldis/4km/layer_avg_850-650.gif.

This plot shows the layer-averaged RH field over the Florida peninsula and adjacent waters. Although the 850–650 mb layer is not the same as that used in this study, it is only offset by 50 mb and likely similar in value to that of the 800–600 mb RH. Forecasters could also calculate a straight average of all the relative humidity values in the 800–600 mb layer from the XMR sounding. No tests were conducted to determine how different this value would be from a depth-weighted value. It is also possible to build a routine in MIDDS that could calculate this value automatically from the 1000 UTC XMR sounding. This option should be strongly considered as a solution as it would give the forecasters the ability to determine the RH from the sounding quickly.

6.3. Future Work

At the most recent AMU Tasking Meeting in February 2005, a task to make some changes to the equations and tool was approved. The dataset described in this report will be used to make three predictor modifications in an attempt to improve equation performance. The first will be to use a new Gaussian filter developed by Mr. Roeder of the 45 WS that produces a smoother curve of the warm season daily lightning climatology. This will create new values for this predictor. Secondly, the 1000–700 mb average wind direction in the 1000 UTC XMR sounding will be calculated and used to confirm the flow regime of the day, especially in situations where the subtropical ridge is just north or south of the area (SW-2 and SE-1, respectively). New flow regime climatologies will be calculated based on the new information. For the third modification an iterative technique will be used to determine an optimal layer for the average relative humidity value. The 800–600 mb layer was used for NPTI and perpetuated in following studies with no attempt to test other layers. If another layer is found, it will be used to calculate RH for the equations. These modifications are likely to create new predictor values and necessitate re-development of the equations. At the very least, the predictors will have new constants associated with them in the equations. It is also possible that different predictors would be chosen for the equations.

Once the above predictor modifications are made, a MIDDS tool to access the equations will be developed to replace the GUI described in this report. The 45 WS forecasters already use MIDDS routinely to view data, so making the equations available in MIDDS would reduce the number of different computer platforms the forecasters must access to gather information. Most importantly, the MIDDS tool will be able to automatically retrieve all data within the MIDDS that are needed by the equations, reducing time spent by the forecaster retrieving and entering values. One exception is that the forecaster will have to manually enter the flow regime of the day and determine its value from other sources since the 1200 UTC soundings will not be available (see Section 6.2). Work on developing the new equations and the MIDDS tool will be complete in time for the 2006 warm season.

There are nine other improvements desired by the 45 WS but not yet tasked to the AMU:

1) Automate the daily Weekly Planning Forecast in addition to the 24-Hour Planning Forecast,
2) Investigate if a bias correction technique to correct for over forecasting can improve performance,
3) Investigate the secondary maximum in lightning non-occurrence when forecast probabilities are high, as seen in Figure 8, to determine the cause of the maximum and any possible corrective action,
4) Create an equation using data from June, July, and August, the months with the highest occurrence of lighting, to determine if the increased sample size produces an equation that can outperform the individual month equations,
5) Once enough data are collected, create equations for the first and second halves of May and September and determine if they outperform the full-month equations. Such equations would account for the climatological changes within each month as the convective season spins up in the second half of May and spins down in the second half of September,
6) Determine if the high percentage of Other flow regimes can be reduced by identifying new regimes or modifying the wind direction thresholds of the current flow regime definitions,
7) Change the lightning verification area to include just the KSC/CCAFS advisory areas, excluding the Astrotech advisory circle seen in Figure 2,
8) Use the available 45 WS Phase-II advisory archive as ground truth proxy for lightning occurrence in addition to CGLSS data to better match the 45 WS lightning advisory procedure, and
9) Investigate the role of persistence when the flow regime changes from the previous day.

The next step should be to collect additional data with which to develop more robust statistical relationships in the equations. More data are also needed for the verification dataset to determine the actual extent of over-forecasting. More data for equation development may help alleviate this issue. New techniques may be available over the next few years that could also help improve equation performance. Evaluation of equation performance should be done continuously to determine the tool's strengths and weaknesses, which can be used to guide future modifications.

References

Everitt, J. A., 1999: An Improved Thunderstorm Forecast Index for Cape Canaveral, Florida. M.S. Thesis, AFIT/GM/ENP/99M-06, Department of Engineering Physics, Air Force Institute of Technology, 98 pp. [Available from the Air Force Institute of Technology, Wright-Patterson Air Force Base, OH 45433].

Harms, D. E., B. F. Boyd, R. M. Lucci, and M. W. Maier, 1998: Weather Systems Supporting Launch Operations at the Eastern Range. AIAA 36th Aerospace Sciences Meeting and Exhibit, Reno, NV, 12-15 January 1998, Paper 98-0744, 11 pp.

Howell, C. L., 1998: Nowcasting Thunderstorms At Cape Canaveral, Florida Using An Improved Neumann-Pfeffer Thunderstorm Index, M.S. Thesis, Air Force Institute of Technology, AFIT/GM/ENP/98M-05, Mar 98, 93 pp. [Available from the Air Force Institute of Technology, Wright-Patterson Air Force Base, OH 45433].

Insightful Corporation, 2001a: *S-PLUS 6 for Windows User's Guide*, Insightful Corp., Seattle, WA, 699 pp.

Insightful Corporation, 2001b: *S-PLUS 6 for Windows Guide to Statistics, Volume 1*, Insightful Corp., Seattle, WA, 731 pp.

Lambert, W., 2004a: Lightning Probabilities Based on Flow Regime. NASA Applied Meteorology Unit Memorandum, 4 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL 32931].

Lambert, W., 2004b: Changes to Sounding Analysis Algorithm in McIDAS. NASA Applied Meteorology Unit Memorandum, 2 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL 32931].

Lambert, W., 2004c: Errors in the HUGE Program. NASA Applied Meteorology Unit Memorandum, 5 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL 32931].

Lazzara, M. A., J. M. Benson, R. J. Fox, D. J. Laitsch, J. P. Rueden, D. A. Santek, D. M. Wade, T. M. Whittaker, and J. T. Young, 1999: The Man computer Interactive Data Access System (McIDAS): 25 Years of Interactive Processing. *Bull. Amer. Meteor. Soc.*, **80**, 271 – 284.

Lericos, T. P., H. E. Fuelberg, A. I. Watson, and R. L. Holle, 2002: Warm season lightning distributions over the Florida Peninsula as related to synoptic patterns. *Wea. Forecasting*, **17**, 83 – 98.

Maier, M., C. Lennon, T. Britt, and S. Schaefer, 1995: Lightning Detection and Ranging (LDAR) system performance analysis. Preprints, *6th Conference on Aviation Weather Systems*, Dallas, TX, Amer. Meteor. Soc.

Menard, S., 2000: Coefficients of determination for multiple logistic regression analysis. *American Statistician*, **54**, 17 – 24.

NCDC, 1996: *Radiosonde Data of North America 1946 – 1996*. CD-ROM. [Available from the NCDC Online Store at http://www.ncdc.noaa.gov.]

NCDC, 1997: *Radiosonde Data of North America 1994 – 1997*. CD-ROM. [Available from the NCDC Online Store at http://www.ncdc.noaa.gov.]

Neumann, C. J., 1971: Thunderstorm forecasting at Cape Kennedy, Florida, utilizing multiple regression techniques. NOAA Technical Memorandum NWS SOS-8.

Ohio State University http://twister.sbs.ohio-state.edu/severe.html.

Peppler, R. A. and P. J. Lamb, 1989: Tropospheric static stability and Central North American growing season rainfall. *Mon. Wea. Rev.*, **117**, 1156 – 1180.

Pfeffer, G. C., 1967: Objective thunderstorm forecasting technique for Patrick AFB and Cape Kennedy AFS.

Roeder, W. P., 1998: Bias Correction to the Neumann-Pfeffer Thunderstorm Index, 45th Weather Squadron internal document, 45 WS/SYR, 1201 Edward H. White II St., MS 7302, Patrick AFB, FL 32925-3238, 21 Aug 98, 3 pp.

StatSoft, Inc., 2004: Electronic Statistics Textbook. Tulsa, OK, StatSoft. On the Web: http://www.statsoft.com/textbook/stathome.html.

Stroupe, J. R., 2003: Warm Season Lightning Distributions over the Northern Gulf of Mexico Coast and Their Relation to the Mesoscale and Synoptic Scale Environments. M.S. Thesis, Department of Meteorology, Florida State University, 69 pp. [Available from the Florida State University, 404 Love Building/Meteorology – 4520, Tallahassee, FL 32306-4520].

Wheeler, M., 2001: Stratified Logistic Thunderstorm Index. NASA Applied Meteorology Unit Memorandum 5 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL 32931].

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, Inc., San Diego, CA, 467 pp.

Wohlwend, C. S., 1998: Improving Cape Canaveral's Day-2 Thunderstorm Forecasting Using Meso-Eta Numerical Model Output, M. S. Thesis, Air Force Institute of Technology, AFIT/GM/ENP/98M-12, Mar 98, 148 pp. [Available from the Air Force Institute of Technology, Wright-Patterson Air Force Base, OH 45433].

World Meteorological Association (WMO), 1992: *Methods of Interpreting Numerical Weather Prediction Output for Aeronautical Meteorology.* Technical Note No. 195, ISBN 92-63-10770-X, 89 pp.

List of Acronyms

45 WS	45th Weather Squadron
AMU	Applied Meteorology Unit
AYS	Waycross, GA 3-letter Identifier
CAPE	Convective Available Potential Energy
CCAFS	Cape Canaveral Air Force Station
CGLSS	Cloud-to-Ground Lightning Surveillance System
CIN	Convective Inhibition
CSI	Critical Success Index
CSR	Computer Sciences Raytheon
CT	Cross Totals
EDT	Eastern Daylight Savings Time
FAR	False Alarm Rate
GUI	Graphical User Interface
HR	Hit Rate
HSS	Heidke Skill Score
JAX	Jacksonville, FL 3-letter Identifier
KI	K-Index
KSC	Kennedy Space Center
KSS	Kuipers Skill Score
LCL	Lifting Condensation Level
LDAR	Lightning Detection And Ranging
LFC	Level of Free Convection
LI	Lifted Index
LNPTI	Logistic Neumann-Pfeffer Thunderstorm Index
McIDAS	Man-computer Interactive Data Access System
MIA	Miami, FL 3-letter Identifier
MIDDS	Meteorological Interactive Data Display System
MSE	Mean Squared Error
NCDC	National Climatic Data Center
NE	Northeast Flow Regime
NLDN	National Lightning Detection Network
NPTI	Neumann-Pfeffer Thunderstorm Index
NW	Northwest Flow Regime
NWS MLB	National Weather Service in Melbourne, FL
PBI	West Palm Beach, FL 3-letter Identifier
PC	Personal Computer
POD	Probability of Detection
POR	Period of Record
PW	Precipitable Water
RH	800–600 mb Average Relative Humidity
SE-1	Southeast-1 Flow Regime
SE-2	Southeast-2 Flow Regime
SLTI	Stratified Logistic Thunderstorm Index
SMG	Spaceflight Meteorology Group
SS	Brier Skill Score
SSI	Showalter Stability Index
SW-1	Southwest-1 Flow Regime
SW-2	Southwest-2 Flow Regime
SWEAT	Severe Weather ThrEAT Index
T_{500}	Temperature at 500 mb
TBW	Tampa, FL 3-letter Identifier
TI	Thompson Index
TT	Total Totals
TTS	Shuttle Landing Facility 3-letter Identifier
UTC	Universal Coordinated Time
WMO	World Meteorological Organization
WOC	Weather Operations Center
XMR	CCAFS Balloon Facility 3-letter Identifier

NOTICE

Mention of a copyrighted, trademarked or proprietary product, service, or document does not constitute endorsement thereof by the author, ENSCO, Inc., the AMU, the National Aeronautics and Space Administration, or the United States Government. Any such mention is solely to inform the reader of the resources used to conduct the work reported herein.