NASA Contractor Report NASA/CR-2007-214732

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

Winifred Lambert
Mark Wheeler
Applied Meteorology Unit
Kennedy Space Center, Florida

July 2007

NASA STI Program ... in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at *http://www.sti.nasa.gov*

- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA STI Help Desk at (301) 621-0134

- Phone the NASA STI Help Desk at (301) 621-0390

- Write to:
  NASA STI Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

NASA Contractor Report NASA/CR-2007-214732

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

Winifred Lambert
Applied Meteorology Unit
Kennedy Space Center, Florida

July 2007

Acknowledgements

The authors thank Mr. William Roeder of the 45th Weather Squadron for lending his statistical expertise to this project, and Mr. Paul Wahner of Computer Sciences Raytheon for his assistance in incorporating the forecast tool into the Meteorological Interactive Data Display System (MIDDS).

Available from:

NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

This report is also available in electronic form at

http://science.ksc.nasa.gov/amu/

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 Kennedy Space Center (KSC) and Cape Canaveral Air Force Station (CCAFS). Forecasters at the Spaceflight Meteorology Group (SMG) also make thunderstorm forecasts during Shuttle flight operations. The lightning probability forecast was based on a subjective analysis of model and observational data and the output from an objective lightning forecast tool developed by the Applied Meteorology Unit (AMU) in Phase I. This tool was a set of five equations that provided a probability of lightning occurrence for the day on KSC/CCAFS during the warm season months of May – September. The forecasters accessed the equations by entering predictor values through a graphical user interface (GUI) developed within Microsoft® Excel©.

In the time since these equations were developed, new ideas regarding certain predictors were formulated and a desire to make the tool more automated was expressed by 45 WS forecasters. They anticipated that modifying the predictors would improve the performance of the equations, and automating the tool would reduce the time spent by forecasters in producing the daily lightning probability. Therefore, the AMU was tasked to re-examine and modify the calculation method of certain predictors and create an automated tool in the current operational weather display system for the 45 WS, the Meteorological Interactive Data Display System (MIDDS).

The 45 WS proposed five modifications to the data and predictors: 1) increase the period of record from 15 to 17 years, 2) modify the valid area to match the lighting warning areas, 3) add the 1000 UTC CCAFS sounding (XMR) to the other soundings used in determining the flow regime, 4) use a different smoothing function for the daily climatology, and 5) determine the optimal relative humidity (RH) layer to be used as a candidate predictor. The data sources were the same as for Phase I and included the Cloud-to-Ground Lightning Surveillance System (CGLSS), 1200 UTC Florida synoptic soundings, and the 1000 UTC XMR sounding. Data from CGLSS were used to determine lightning occurrence for each day. The 1200 UTC Florida and 1000 UTC XMR soundings were used to determine the flow regime for each day and the 1000 UTC XMR soundings were used to calculate local stability parameters. Each of the three datasets was processed and analyzed to create the predictand and candidate predictors needed for the statistical forecast equation development. The CGLSS data were used to create a binary predictand for lightning, where a '1' denoted that lightning occurred during the day and a '0' denoted that lightning did not occur. The flow regimes and stability parameters from the soundings were used to calculate the candidate predictors of lightning occurrence. This resulted in 14 candidate predictors for equation development.

The AMU stratified the data into two sub-sets: a development dataset consisting of 14 warm seasons from which the equations were developed, and an independent verification dataset of 3 warm seasons on which the equations were tested. One logistic regression equation was developed for each 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 five equations contained four to five predictors. The flow regime lightning probability was the second-most important predictor in all five equations. One-day persistence and Vertical Totals were in four of the five equations. Other predictors included the Thompson Index, 825–525 mb average RH, daily climatology, K-Index, and Total Totals.

The AMU then conducted five tests to determine equation performance. The results indicated that the Phase II equations showed an increase in skill over several standard forecasting methods and an 8% gain in skill over the Phase I equations. They also showed improved reliability and ability to distinguish between non-lightning and lightning days than the Phase I equations. Given the overall improved skill, the 45 WS requested that the Phase II equations be transitioned to operations and added to the current set of tools used to determine the daily lightning probability of occurrence.

An Excel© GUI was created in Phase I to facilitate forecaster access to the equations through user-friendly input and fast, easy-to-read output of the lightning probability for the day. This GUI was updated with the new equations developed in Phase II and transitioned to operations until a MIDDS GUI could be developed. The new MIDDS GUI gathers the data needed for the predictors and enters the appropriate values into the equations. The design of this GUI closely resembles the Excel GUI, making it easier for forecasters to transition between GUIs. Personnel from the 45 WS were involved in the MIDDS 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 3

List of Figures 5

List of Tables 7

1. Introduction 8
   - 1.1 Phase I 8
   - 1.2 Phase II 9
2. Data 10
   - 2.1 Cloud-to-Ground Lightning Surveillance System (CGLSS) 10
   - 2.2 Florida 1200 UTC Rawinsondes 11
   - 2.3 XMR 1000 UTC Rawinsonde 11
3. Modifications 12
   - 3.1 Increased POR 12
   - 3.2 New Valid Area 12
   - 3.3 Flow Regime Discriminator 14
   - 3.4 Smoother for Daily Climatology 16
   - 3.5 Optimal RH Layer 17
4. Equation Elements 19
   - 4.1 Binary Predictand 19
   - 4.2 Candidate Predictors 19
5. Equation Development and Testing 24
   - 5.1 Data Availability 24
   - 5.2 Equation Development 26
   - 5.3 Equation Performance 30
6. Graphical User Interface 38
   - 6.1 Microsoft Excel GUI 38
   - 6.2 MIDDS GUI 44
   - 6.3 Predictor Responses 47
7. Summary and Conclusions 52
   - 7.1 Equation Performance Review 52
   - 7.2 GUI Issues 52
   - 7.3 Future Work 54

References 55

List of Acronyms 56

List of Figures

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. 10

Figure 2. The red dots on the map show the locations of all soundings used in this task. 11

Figure 3. The 5 n mi lightning warning circles on KSC/CCAFS and Astrotech. The valid area for the Phase II work is within the four blue (KSC) and six red (CCAFS) circles with centers to the right of the vertical black line. 13

Figure 4. The CG flash density per km^2 per year over east central Florida. The area within the solid-outlined rectangle is analogous to the full area in Figure 1, and the dashed vertical line is analogous to the solid vertical line in Figure 1. This image was created by Mr. Geoffrey Stano for his graduate work at the Florida State University. 14

Figure 5. (a) The daily raw (thin blue curve), ±7-day smoothed (red curve), and ±14-day smoothed (thick blue curve) climatological probability values of lightning occurrence for the warm-season months in 1989–2005, and (b) The Gaussian weight values (W) used in the ±14-day smoothing equation. 17

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 [2006] Figure 6.12)..... 27

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 June development dataset. 28

Figure 8. Forecast probability distributions for lightning (red) and non-lightning (blue) days in the verification data. The solid lines represent the P-2 equations and the dashed lines represent the P-1 equations. The y-axis values are the frequency of occurrence of each probability value, and the x-axis values are the forecast probability values output by the equations. 32

Figure 9. Reliability diagram of the P-1 and P-2 probability forecasts for all months. The straight diagonal line represents perfect reliability, the blue curve represents the reliability of the P-1 equations, and the red curve represents the reliability of the P-2 equations. The histogram at the lower right shows the number of observations in each probability range for the old (blue) and new (red) forecast methods. 33

Figure 10. Graph showing the values in the four contingency table cells in Table 10 for the range of probability values 0–1 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 P-2 equation forecasts (eqn). The vertical 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. 36

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. 39

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. KI and V 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 6.1.4)..... 40

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 the sounding parameters TI, VT, and MRH. 40

Figure 14. Same as Figure 13 except for July, and with the sounding parameters TI and TT. 41

Figure 15. Same as Figure 13 except for August, and with the sounding parameters TI, MRH, and VT. 41

Figure 16. Same as Figure 12 except for September, and with the sounding parameters MRH and VT. 42

Figure 17. The equation output dialog box displaying the probability of lighting for the day based on the values input to the date and equation predictor dialog boxes. 43

Figure 18. The MIDDS Toolbar showing the 'FCST Tools' button drop-down menu with 'Lightning Forecast Tool' highlighted. 44

Figure 19. The error dialog box displayed when a 1000 UTC XMR sounding for the current date is not available. Clicking the 'OK' button closes the box. 44

Figure 20. Equation predictor dialog box for June in MIDDS. A tab for each month is at the top, followed by the date and sounding time, then the predictor values. Clicking the 'Dismiss' button closes the GUI, the 'Reset Parameters' button resets the sounding stability parameters to original values, and the 'Calculate Probability' button displays the probability output dialog box (Figure 22). 45

Figure 21. The error dialog box displayed when persistence is not chosen (left) or a flow regime is not chosen (right). Clicking the 'OK' button closes the box. 46

Figure 22. The dialog box displaying the probability of lightning occurrence for the day as calculated by the equation. Clicking the 'OK' button closes the box. 46

Figure 23. Equation response charts for May 15: (a) change in probability due to changes in VT and KI with flow regime = SW, persistence = Yes, KI = 17 when VT was varied from 10 to 50 (blue), and VT = 25 when KI was varied from -30 to 70 (red); (b) change in probability due to changes in flow regime and persistence with KI = 17 and VT = 25. The red bars represent persistence = Yes and the blue bars represent persistence = No. 47

Figure 24. Equation response charts for June 15: (a) change in probability due to changes in TI, VT, and MRH with flow regime = SW, persistence = Yes, VT = 25 and MRH = 59% when TI was varied from -20 to 60 (blue), TI = 30 and MRH = 59% when VT was varied from 0 to 45 (red), and TI = 30 and VT = 25 when MRH was varied from 0 to 100% (green); (b) changes in probability due to changes in flow regime and persistence with TI = 30, VT = 25, and MRH = 59%. The red bars represent persistence = Yes and the blue bars represent persistence = No. 48

Figure 25. Equation response charts for July 15: (a) change in probability due to changes in the values of TT and TI with flow regime = SW, persistence = Yes, TT = 44 when TI was varied from -20 to 70 (blue), and TI = 31 when TT was varied from 0 to 75 (red); (b) changes in probability due to changes in flow regime and persistence with TT = 44 and TI = 31. The red bars represent persistence = Yes and the blue bars represent persistence = No. 49

Figure 26. Equation response charts for August 15: (a) change in probability due to changes in the values of TI, MRH, and VT with flow regime = SW, persistence = Yes, MRH = 59% and VT = 34 when TI was varied from -20 to 70 (blue), TI = 31 and VT = 24 when MRH was varied from 0 to 100% (red), and TI = 31 and MRH = 59% when VT was varied from 0 to 50 (green); (b) changes in probability due to changes in flow regime with TI = 31, MRH = 59%, and VT = 24. 50

Figure 27. Equation response charts for September 15: (a) change in probability due to changes in the values of MRH and VT with flow regime = SW, persistence = Yes, VT = 24 when MRH was varied from 0 to 100% (blue), and MRH = 57% when VT was varied from 0 to 45 (red); (b) changes in probability due to changes in flow regime and persistence with MRH = 57% and VT = 24. The red bars represent persistence = Yes and the blue bars represent persistence = No. 51

List of Tables

Table 1. List of the flow regime names used in Phases I and II and the corresponding sectors showing the average 1000 – 700 mb wind directions at each of the stations. 15

Table 2. The number of days for each flow regime in the POR before and after replacing the synoptic regime with the local regime at XMR. The **black bold** values indicate more days and the **red bold** values indicate less days in the After column 16

Table 3. 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. 20

Table 4. 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. 21

Table 5. Summary of available data in the POR. The first column contains the names of the months 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 and 17 warm seasons. The three columns under the heading '# MISSING DAYS' show the number of unavailable days due to missing data from each dataset in the subheadings, and the number of days missing due to the combined missing data from both datasets. The value in parentheses in the third column is the number of days in which data were missing from both datasets. The final column shows the number of days with all data available. The percent of total possible days is given in parentheses. 25

Table 6. Summary of missing and available data for equation development and verification. 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 17 warm seasons, the number of those days for equation verification, 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 (Table 5), and the actual number of days in the verification and development datasets. 26

Table 7. The final predictors for each monthly equation, in rank order of their reduction in residual deviance. The predictors in red were in every equation, the predictors in blue were in four of the five equations, the predictors in green were in three of the five equations, and the predictors in black were in only one equation. 30

Table 8. The SS values that show the percent (%) improvement (degradation) in skill of the P-2 equations over the reference forecasts of persistence, daily and monthly climatologies, flow regime probabilities, and the P-1 equations developed in Lambert and Wheeler (2005). These scores were calculated using the verification data for each month and for the entire warm season (All). 31

Table 9. The values of the terms in Equation 12 for the P-1 and P-2 equations. The last row shows the percent change in value from the P-1 to the P-2 equations. 34

Table 10. Basic contingency table for calculating categorical accuracy measures and skill scores (Wilks 2006). The equations for the accuracy measures and skill scores are in the bottom row. 35

Table 11. The accuracy measures and skill scores for the P-2 equations with a cutoff probability of 0.47, the P-1 equations with a cutoff probability of 0.35, and the persistence forecasts associated with each equation set. 37

Table 12. Summary values for each of the predictors in the POR 1989–2005. The last two rows contain the upper and lower limits of the values allowed in the GUI. 43

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 to the 45 WS staff in the morning at 1100 UTC (0700 EDT) and released for use at 1130 UTC (0730 EDT). Forecasters at the Spaceflight Meteorology Group (SMG) also make thunderstorm forecasts during shuttle operations. The probability of lightning occurrence is used by personnel involved in determining the possibility of violating launch commit criteria, evaluating shuttle flight rules, and planning for daily ground operation activities on Kennedy Space Center (KSC) and Cape Canaveral Air Force Station (CCAFS).

Until completion of the Phase II work described in this report, the lightning probability forecast was based on a subjective analysis of model and observational data and the output from an objective lightning forecast tool developed by the Applied Meteorology Unit (AMU) in Phase I. This tool was a set of equations that provided a probability of lightning occurrence for the day on KSC/CCAFS during the warm season months of May – September. The forecasters accessed the equations through a Microsoft® Excel© graphical user interface (GUI) by entering predictor values of sounding parameters and making choices for the flow regime and one-day persistence (hereafter persistence). After they were developed, these equations showed an improvement in performance over other standard forecast methods in use and were transitioned to operations for the 2005 warm season.

In the time since these equations were developed, new ideas regarding certain predictors were formulated and a desire to make the tool more automated was expressed by 45 WS forecasters. They anticipated that modifying the predictors would improve the performance of the equations, and automating the tool would reduce the time spent by forecasters in producing the daily lightning probability. Phase II, therefore, had two parts: 1) to re-examine and modify the calculation method of certain predictors and to use the modified predictors to develop new monthly equations, and 2) to create an automated tool in the current operational weather display system for the 45 WS, the Meteorological Interactive Data Display System (MIDDS).

1.1 Phase I

The Phase I objective lightning probability tool was a set of five logistic regression equations, one for each month in the warm season, that calculated the probability of lightning occurrence for the day (Lambert and Wheeler 2005). They were developed using a 15-year (1989–2003) archive of Cloud-to-Ground Lightning Surveillance System (CGLSS) data, 1200 UTC Florida synoptic soundings, and the 1000 UTC CCAFS sounding (XMR). Each equation had five to six predictors that were chosen from a larger set of candidate predictors through an iterative statistical process. These equations outperformed the operational tools used by the 45 WS, in particular the Neumann-Pfeffer Thunderstorm Index (NPTI) (Neumann 1971) and persistence. They 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. Based on the test results, the equations were transitioned to operations in time for the 2005 warm season and replaced the NPTI as the official lightning forecast tool.

The 45 WS requested the AMU to create and deliver a GUI to facilitate user-friendly input to the equations and fast output. The AMU created this GUI using Microsoft® Excel© Visual Basic®. During development of the GUI, the 45 WS provided comments and suggestions on the design to ensure that the final product addressed their operational needs. The GUI has three dialog boxes. The first asks for the date to determine which monthly equation to use and the value of the daily climatology to use in the equation. The second dialog box is different for each month. It asks for the particular equation predictor values specific for that month. The third dialog box displays the resulting lightning probability for the day in a large font.

1.2 Phase II

There were two facets to the Phase II work. The first was to modify certain parameters and predictors to determine if these modifications would improve the equation performance. The second was to make the tool more automated, eliminating the need for the forecasters to gather information from one source and entering the values manually into the Excel GUI.

1.2.1 Modifications

In an effort to improve the lightning probability forecast, the 45 WS proposed five modifications to the Phase I tool:

1) Increase the period of record (POR) by adding data from the 2004–2005 warm seasons. The new 17-year POR would likely produce a more accurate daily lightning climatology and produce more robust statistics in the development of the equations.
2) Modify the valid area. The valid area for the lightning forecasts was reduced to include the 5 n mi warning circles on KSC and CCAFS only, eliminating the western portion of the area used in Phase I. This produced an accurate estimation of whether lightning occurred in the warning areas of responsibility for the 45 WS.
3) Modify the method used to determine the flow regime of the day. The method of determining the flow regime for the Phase I equations followed the procedure outlined in Lericos et al. (2002). It used the mean wind direction in the 1000–700 mb layer from the Jacksonville (JAX), Tampa (TBW), and Miami (MFL) 1200 UTC soundings. However, this method failed to classify the flow regime in 44% of the days in the 15-year POR. This method was modified in Phase II to include the 1000–700 mb mean wind direction in the 1000 UTC XMR sounding. This wind direction was used to determine the flow regime when it could not be classified by using the combined wind directions from the other three soundings.
4) Use a different smoothing function for the daily lightning climatology. A ±7-day Gaussian smoother with a scale factor of 3 days was used in Phase I to smooth the daily climatology curve, but it still showed some noisiness. A ±14-day smoother with a 7-day scale factor produced smooth results with no noisiness, which may be closer to the actual climatology.
5) Determine the optimal average relative humidity (RH) layer. The average RH in the 800–600 mb layer from the XMR 1000 UTC sounding was a predictor for the Phase I equations. This parameter was determined as valuable for forecasting convection in the KSC/CCAFS area over 30 years ago. It has been used in several studies since then, but no attempt has been made to verify whether 800–600 mb is the optimal layer.

After the AMU incorporated these changes into the predictor set, new monthly equations were developed and their performance compared to standard forecast benchmarks and the Phase-1 equations. These modifications improved the performance of the equations.

1.2.2 Automated Input

To use the Excel GUI, the forecasters gathered data from the XMR 1000 UTC sounding and other sources, then input that data manually. This increased the risk of a forecaster entering an incorrect value, resulting in the calculation of an erroneous probability value. It also increased the time a forecaster spent in preparing the daily and weekly forecasts. The 45 WS requested a tool be developed on MIDDS to retrieve the required parameter values automatically for the equations to calculate the probability of lightning for the day. This would reduce the possibility of human error and increase efficiency, allowing forecasters to do other duties.

Mr. Paul Wahner of Computer Sciences Raytheon (CSR) created a GUI in MIDDS that gathers the data needed for the predictors and enters the appropriate values into the equations. The MIDDS GUI design resembles the Excel GUI, making it easier for forecasters to transition from the Excel to the MIDDS GUI.

2. Data

The POR for the data used to develop the forecast equations was increased from 15 to 17 years by adding the data collected during the 2004 and 2005 warm seasons. The data sources include the

- CGLSS,
- 1200 UTC JAX, TBW, and MFL 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 1000 UTC XMR and 1200 UTC JAX, TBW, and MFL 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 2005a).

More details on each data type can be found in the Phase I final report (Lambert and Wheeler, 2005). Discussions for each data type used are included in this report for completeness, but they only contain information pertaining to Phase II for brevity.

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. Mr. Wahner of CSR provided the additional data for the 2004 and 2005 warm seasons. The CGLSS data were used to determine whether or not lightning occurred on each day in the POR. The primary purpose of the CGLSS data was to create the binary predictand for the equations. The data were also used to create the daily climatological lightning frequency and persistence forecasts that would be used as candidate predictors and forecast benchmarks against which to test the new equations.

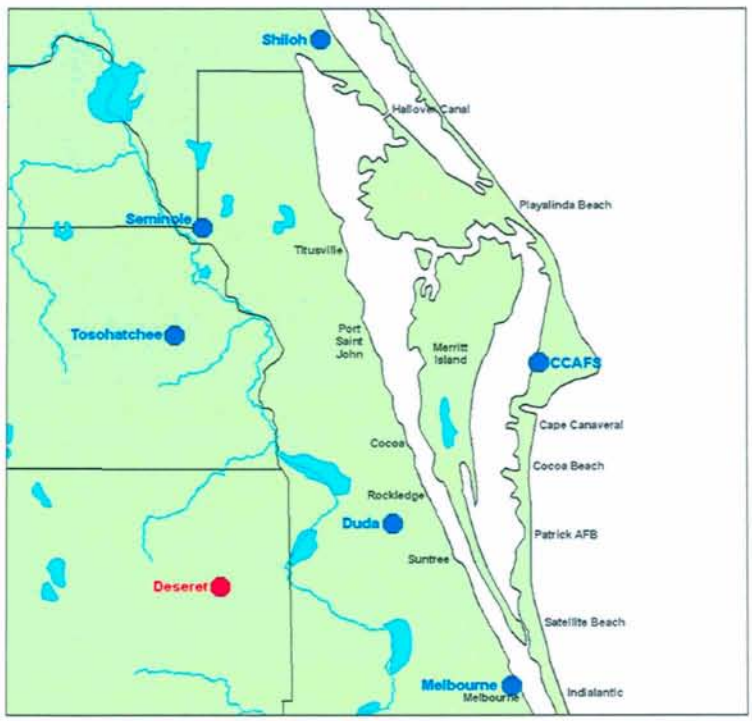

Figure 1. The locations of the six CGLSS sensors are indicated by the blue circles. The location names are next to the circles. The Duda sensor was moved to the Deseret site (red circle) in 2005.

2.2 Florida 1200 UTC Rawinsondes

These data were collected to determine the daily flow regimes using the procedure outlined in Lericos et al. (2002). The data from the 2004 and 2005 warm seasons were downloaded from the Global Systems Division (GSD) web site http://raob.fsl.noaa.gov/. As noted in Lericos, the current MFL 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 MFL and JAX, respectively, during the period 1989–1994. All future references to MFL and JAX include the 1989–1994 data from AYS and PBI. The map in Figure 2 shows the locations of all the 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 7.2.2.

2.3 XMR 1000 UTC Sounding

The XMR sounding location is shown in Figure 2. The 45 WS forecasters use data from the 1000 UTC sounding 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 for Phase I and Phase II. The parameters were used as candidate predictors in the equation development. In Phase II, they were also used to determine the flow regime of the day along with the 1200 UTC JAX. TBW, and MFL soundings. The procedure will be discussed in Section 3. Mr. Wahner of CSR supplied the 2004 and 2005 warm season data to the AMU.

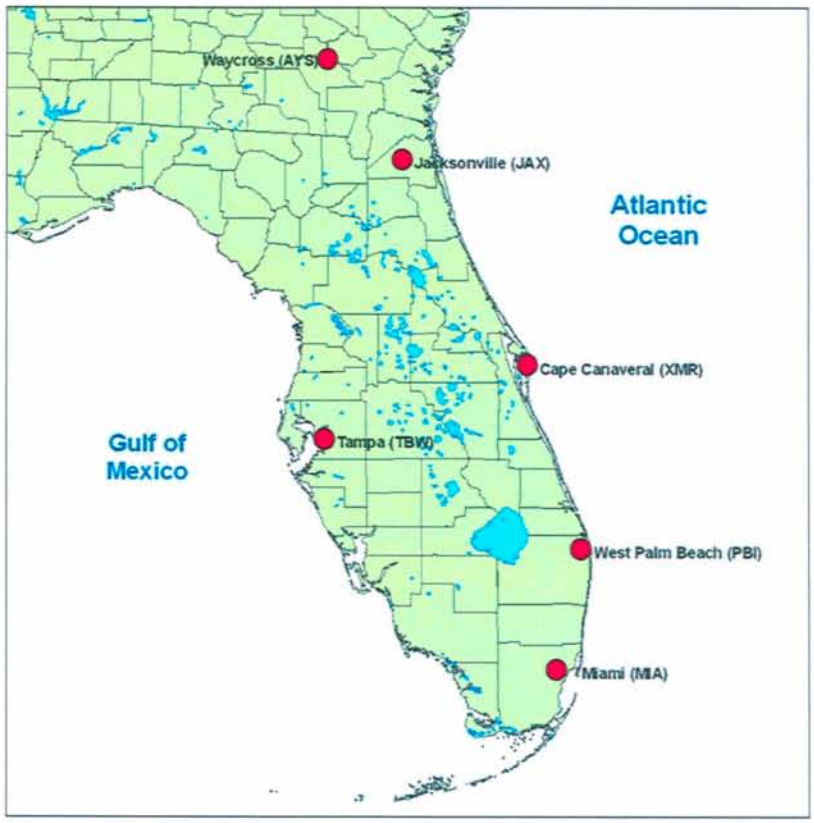

Figure 2. The red dots on the map show the locations of all soundings used in this task.

3. Modifications

As stated in Section 1, the 45 WS requested five modifications to the Phase I data and candidate predictors that could improve their performance. The five modifications were to increase the POR by two years, modify the valid area, use the XMR 1000 UTC sounding to help determine the flow regime of the day, use different smoothing values for the daily climatology, and determine an optimal layer for the average RH calculation.

3.1 Increased POR

Two more warm seasons occurred since the Phase I equations were developed, and the 45 WS requested that data from these two seasons be used in the development of the Phase II equations. The new POR now includes data from all the warm seasons in the years 1989–2005. This increased the POR from 15 to 17 years and could possibly produce a more accurate daily lightning climatology and produce more robust statistics in the development of the equations. Statistically, the standard error is inversely proportional to the square root of the sample size. The increase in the number of years from 15 to 17 will decrease the standard error by

$$100 \text{ X } \frac{\frac{1}{\sqrt{15}} - \frac{1}{\sqrt{17}}}{\frac{1}{\sqrt{15}}} = 6.07\%. \quad (1)$$

3.2 New Valid Area

The equations were meant to forecast lightning within 10 warning circles, each 5 n mi in diameter, surrounding specific asset locations (Figure 3). This is analogous to a 45 WS Phase II lightning warning in which lightning is imminent or occurring within one or more of the circles. The valid area for cloud-to-ground (CG) lightning occurrence in Phase I was the entire area shown in Figure 3, a rectangle surrounding all 5 n mi warning circles including Astrotech. The AMU considered it computationally simpler to use the area of a rectangle than to determine whether each strike detected by CGLSS occurred within one or more of the warning circles.

For Phase II, the 45 WS requested that the valid area be reduced to include only the 10 circles on KSC and CCAFS, those circles to the right of the vertical black line in Figure 3. While the 45 WS has a warning responsibility for Astrotech, this facility is outside the area covered by the daily 24-Hour Planning Forecast, which is the product supported by the equations. Also, upon further consideration, the AMU devised a simple mathematical algorithm that determines how far each strike occurred from the center of each of the 10 circles. The latitude/longitude (lat/lon) values from each CGLSS strike were used to calculate the distance between it and the center lat/lon of all 10 circles using the Great Circle Distance Formula (http://www.meridianworlddata.com/distance-calculation.asp):

$$D = 3437.75 * \arccos[\sin(lat1) * \sin(lat2) + \cos(lat1) * \cos(lat2) * \cos(lon2\text{-}lon1)], \quad (2)$$

where D is the distance between the strike and the circle in nautical miles, lat1/lon1 is the lat/lon of the circle center, lat2/lon2 is the lat/lon of the strike, and all lat/lon values are in radians. The strikes that were within 5 n mi of any circle ($D \leq 5$ n mi) were considered in the valid area, and the day on which those strikes occurred was considered a lightning day. As with Phase I, the number of strikes was not considered in the lightning occurrence probabilities.

The new valid area represents the actual lightning warning areas and is smaller than the area used in Phase I. Changing the valid area reduced both the number of strikes and the number of lightning days in the Phase II data base compared with the Phase I data base. Even with a 13% increase in the number of days due to the 17-year POR as compared to the 15-year POR, the number of strikes decreased by over 40% while the number of lightning days decreased by 6%.

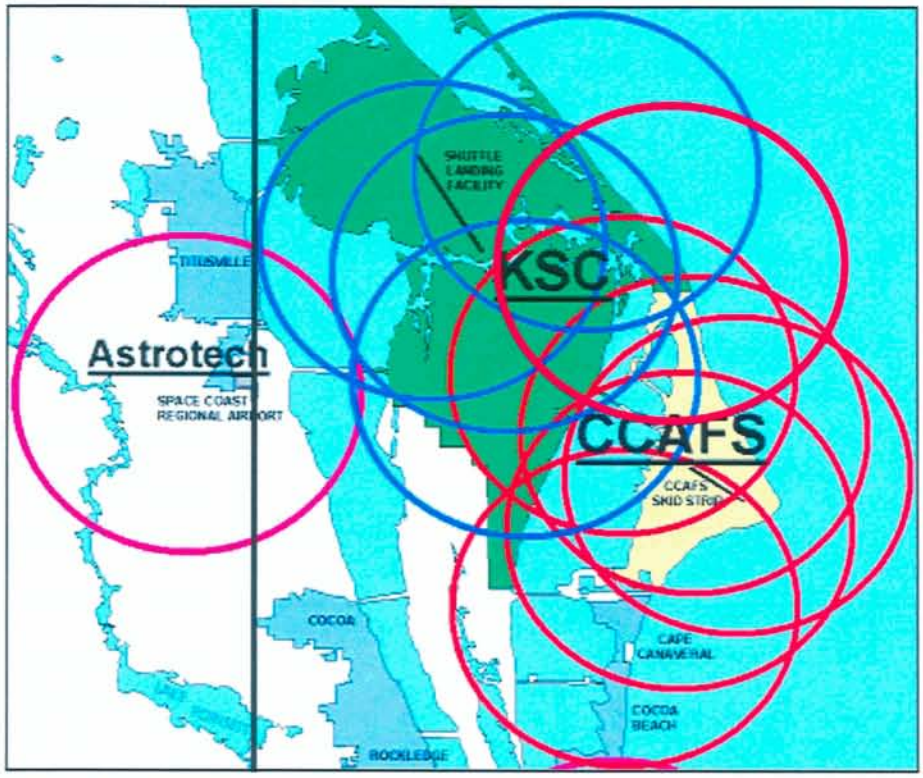

Figure 3. The 5 n mi lightning warning circles on KSC/CCAFS and Astrotech. The valid area for the Phase II work is within the four blue (KSC) and six red (CCAFS) circles with centers to the right of the vertical black line.

In addition to the reduction in the spatial area, the spatial CG strike-density climatology was also an important factor in the reduction of CG strikes and lightning days. Figure 4 shows the yearly climatology of the number of strikes per square kilometer using National Lightning Detection Network (NLDN) data collected in the period 1992–2004, a 13-year subset of the Phase II 17-year POR. The approximate outline of the Phase I valid area is drawn by a solid black rectangle in Figure 4, and a dashed vertical line shows the westernmost edge of the 5 n mi circles, analogous to the solid vertical line in Figure 3. Note the decrease in lightning activity going from the mainland to the coast (left to right). The area in Phase I encompassed areas with CG strike densities of > 10 per km^2 per year over the mainland. The new valid area contains a small area of 10–12 CG per km^2 per year near its western edge, with most of the area having a strike density of < 10 per km^2 per year and approaching 2 per km^2 per year at the coastline on the right side of the image. This is the likely cause for the large decrease in the number of CG strikes in the Phase II data set. The number of lightning days is not, however, related to the number of CG strikes in this study. It takes only one strike to define a lightning day. More or fewer strikes do not necessarily translate to more or fewer lightning days. This helps explain the disparity in the percent decrease between these two parameters: 40% for the number of strikes and only 6% for the number of days.

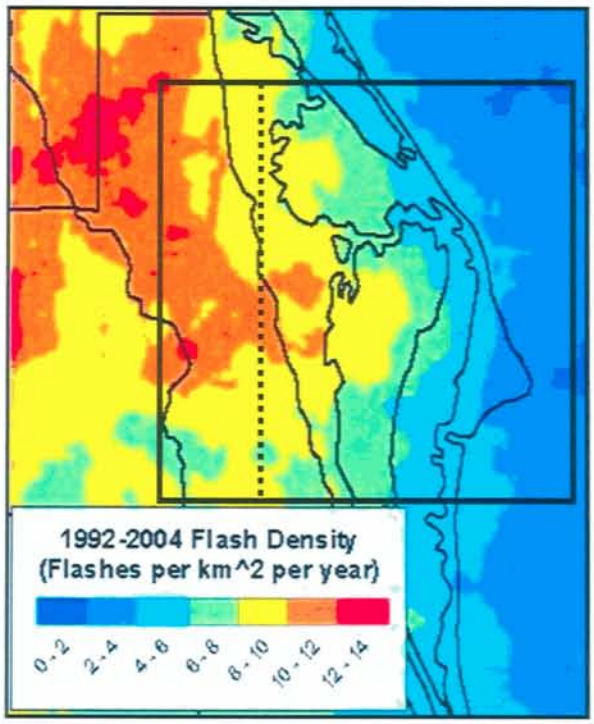

Figure 4. The CG flash density per km^2 per year over east central Florida. The area within the solid-outlined rectangle is analogous to the full area in Figure 1, and the dashed vertical line is analogous to the solid vertical line in Figure 1. This image was created by Mr. Geoffrey Stano for his graduate work at the Florida State University.

3.3 Flow Regime Discriminator

After stratifying the days by flow regime in the Phase I work, the AMU found that 44% of the days could not be categorized into any of the defined regimes. Given that lightning occurred on 45% of those days, they could not be discounted. Therefore, the AMU stratified them into a new flow regime category named 'Other'. The 45 WS suggested that perhaps the 1000–700 mb winds in the 1000 UTC XMR sounding could be used to determine a flow regime for the 'Other' days in the Phase II work. This would reduce the number of days in that category and increase the number of days in the defined categories such that more robust statistics could be calculated for them.

The first step in the procedure was to determine the 'synoptic' flow regime of the day by using a combination of the average 1000–700 mb wind directions from the 1200 UTC MFL, TBW, and JAX soundings, as outlined in Lericos et al. (2002) and done in Phase I. The specifics of the mathematical procedure used to calculate the average wind direction in the 1000–700 mb layer are given in Lambert and Wheeler (2005). The wind speeds and directions were decomposed into u- and v-components, then the average u- and v-winds in the layer were calculated using a depth-weighted average and recombined to get an average wind speed and direction. Table 1 contains the definitions for the flow regimes used in Phases I and II.

The next step was to calculate the average 1000–700 mb wind directions in the 1000 UTC XMR soundings, which were used to determine the 'local' flow regime of the day. The local flow regime was the discriminator in determining the final flow regime of the day when the synoptic regime was Other, Missing, SE-1, or SW-2. In the SE-1 and SW-2 regimes, the ridge axis from the high over the Atlantic Ocean was just north or south of TBW, respectively. Exactly where the ridge was located relative to KSC/CCAFS was unknown. The local flow regime would be used to determine whether the ridge was north, south, or over KSC/CCAFS. For example, it was possible that the average direction in the 1200 UTC soundings could determine that the ridge was north of TBW, but the flow at XMR indicated the ridge was actually south of the KSC/CCAFS area.

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

Flow Regime Name and Description		*Rawinsonde Station*		
		MFL	**TBW**	**JAX**
SW-1	Subtropical ridge south of MFL Southwest flow over KSC/CCAFS	180°-270°	180°-270°	180°-270°
SW-2	Subtropical ridge north of MFL, 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 Florida, likely from a stronger-than-average subtropical ridge south of MFL extending into Gulf of Mexico	270°-360°	270°-360°	270°-360°
NE	Northeast flow over Florida, likely from a stronger-than-average subtropical ridge north of JAX extending into southeast U.S., at times forming a closed high pressure center	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			

The Other and Missing synoptic regimes were replaced with the local flow regime when it was SW, SE, NW, or NE according to the definitions in Table 1 and the average wind speed was greater than 4 kt. Since there are two regimes each for SE and SW flow, the AMU consulted with Mr. Roeder of the 45 WS to determine which regime should be chosen when the XMR mean direction was from the SE or SW. They decided the default regimes would be SE-1 and SW-2 in which the ridge is south of JAX/north of TBW and south of TBW/north of MFL, respectively. Based on these criteria, the AMU developed an algorithm with the following logic:

- If the local flow regime was not missing and the speed was greater than 4 kt,
  - If synoptic regime was Other, replace with local regime.
  - If synoptic regime was Missing, replace with local regime.
  - If synoptic regime was SW-2, replace with local regime if it was SE-1.
  - If synoptic regime was SE-1, replace with local regime if it was SW-2.
- If the local flow regime was missing, the synoptic regime was not changed.

The last two if statements under 'If the local flow regime is not missing', added to account for times when the ridge was just north or south of the KSC/CCAFS area, were executed infrequently. There were 59 cases in which the synoptic SE-1 flow was changed to SW-2, and 18 cases in which the synoptic SW-2 flow was changed to SE-1.

The number of days for each flow regime in the POR before and after this algorithm was applied are shown in Table 2. The bold black numbers in the 'After' column show an increase in the number of days and the bold red numbers show a decrease. The algorithm increased the number of SW-2, SE-1, NW, and NE cases, and reduced the number of Other and Missing days by ~70%. The SW-1 and SE-2 regimes did not change due to the fact that SE flow at XMR was considered to be the SE-1 regime and SW flow was considered to be the SW-2 regime. The synoptic regimes could only be replaced by one of these two regimes.

Table 2. The number of days for each flow regime in the POR before and after replacing the synoptic regime with the local regime at XMR. The **black bold** values indicate more days and the **red bold** values indicate less days in the After column

Flow Regimes	*Before*	*After*
SW-1	301	301
SW-2	256	**606**
SE-1	318	**438**
SE-2	248	248
NW	100	**307**
NE	114	**317**
Missing	187	**58**
Other	1077	**326**

3.4 Smoother for Daily Climatology

The changes in the POR and valid area necessitated the recalculation of the daily climatological probability values of lightning occurrence. These values were used in Phase I as predictors in all five equations, and were also used as forecast benchmarks when testing the performance of the equations. The number of years that each day experienced lightning was determined first. Then, a raw climatology was calculated by dividing this number by 17, the number of years in the POR. This yielded a fractional value between 0 and 1 for each day. The thin blue jagged curve in Figure 5a is the raw 17-year climatology for each day in the warm season. The noisy appearance of this curve is likely due to the few number of years in the POR; 17 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 in order to smooth out the curve and infer what the long-term climatology would be if enough observations were available. While this results in a smoother, presumably more representative climatological curve, it does so at the cost of temporal resolution where valid small-scale variations are lost. In Phase II as in Phase I, a Gaussian center-weighting function was used to smooth the curve, defined by the equation

$$P = \frac{1}{N}\left\{\frac{\sum_{k=0}^{m}\left[W(F_{n-k} + F_{n+k})\right] + F_n}{\sum_{k=0}^{m}\left[W * 2\right] + 1}\right\} \text{ (Everitt 1999),} \qquad (3)$$

where W is the Gaussian weighting function

$$W = \exp\left[\frac{-\left(k^2\right)}{2 * \sigma^2}\right] \text{ (Wilks 2006),} \qquad (4)$$

P = climatological probability on the day of interest,
N = number of years in the POR (17),
n = day number of interest,
k = number of days distant from n,
m = maximum ± number of days distant from n,
F = raw probability on day of interest, and
σ = scale factor in units of days.

This is a center-weighted function in which an equal number of points before and after n are used to create the smoothed value. The value of W is 1, the maximum, for the original time and decreases for the observed values further away in time, before and after n.

In Phase I, m = ±7 and σ = 3 days. Using these values for the 17-year POR resulted in the red curve in Figure 5a. It was smoothed considerably from the raw climatology, but still had a certain level of noise. The 45 WS suggested using m = ±14 and σ = 7 days. This created an even smoother curve through the warm season, represented by the thick blue curve in Figure 5a. The values of W for these parameters are shown in Figure 5b. The daily climatology using the ±14-day Gaussian smoother was used as a candidate predictor variable in developing the forecast equations and as a forecast benchmark when testing equation performance.

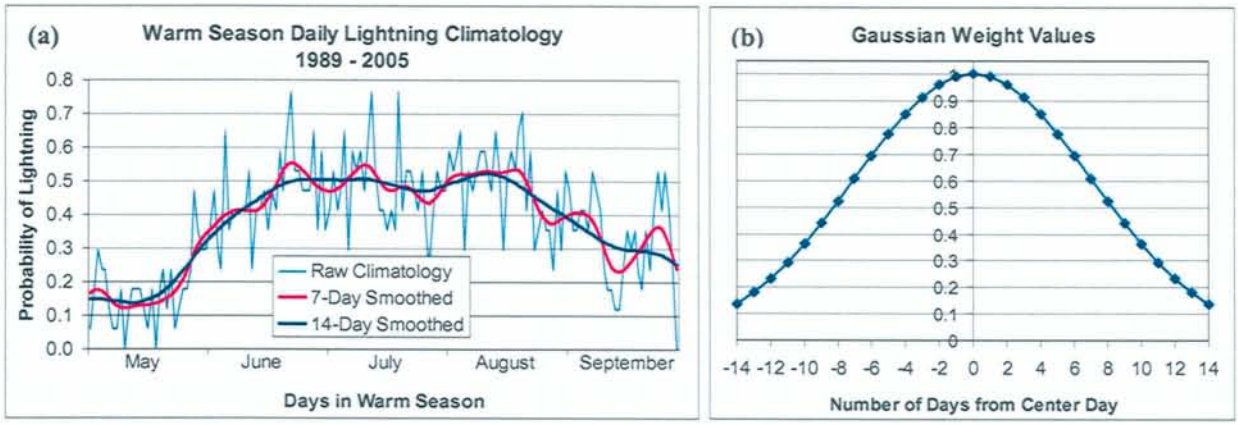

Figure 5. (a) The daily raw (thin blue curve), ±7-day smoothed (red curve), and ±14-day smoothed (thick blue curve) climatological probability values of lightning occurrence for the warm-season months in 1989–2005, and (b) The Gaussian weight values (W) used in the ±14-day smoothing equation.

3.5 Optimal RH Layer

The average RH in the 800–600 mb layer was an important predictor in four of the five equations developed in Phase I. This parameter was determined as valuable in the study that created the NPTI (Neumann 1971) over 30 years ago. It has been used in several studies since that time, but no rigorous attempts have been made to determine if 800–600 mb is truly the optimal layer for this predictor. In collaboration with Mr. Roeder of the 45 WS, the AMU employed an iterative technique to determine the optimal layer for the average RH calculation using the 1000 UTC XMR sounding.

The iterative technique began by calculating the average RH in all 200-mb layers between 950 mb as the lowest base and 400 mb as the highest top, incrementing the base and top of each layer by 25 mb. This resulted in 15 layers. The sounding data included the mandatory and significant levels. All levels were checked to determine if they contained all possible pressure levels divisible by 25 (e.g. 950, 925, 900, 875, …, 425, 400). If not, the levels were created and the RH at each calculated using a log(p)-based linear interpolation, where p is the pressure, between the existing levels above and below the new level. Then, the average RH in each of the 15 200-mb layers was calculated with a log(p)-weighted averaging method using all levels in the layer. The region of influence, D_i, for each RH observation in the sounding was defined as

$$D_i = \frac{[\log(p_{i-1}) - \log(p_{i+1})]}{2}, \tag{5}$$

where 'i' is the level number in the layer, p_{i-1} is the pressure at the observation directly below and p_{i+1} is the pressure at the observation directly above p_i. The average RH for each 200-mb layer was calculated with the equation

$$RH_{avg} = \frac{\sum(RH_i * D_i)}{\sum D_i}. \tag{6}$$

This was done for each warm-season month. The next step was to determine the layer with the highest linear correlation to lightning occurrence for each month. The centers of the five monthly layers were all within 50 mb of each other. Mr. Roeder of the 45 WS and the AMU consulted and determined that the layers for each month were similar enough to combine the data and determine one optimal RH layer for the entire warm season. Using the above procedure, the 200-mb layer average RH with the highest correlation to lightning occurrence was 775–575 mb, with a center at 675 mb.

The iterative technique began anew at the pressure of 675 mb by adding layers in 25 mb increments above and below this pressure level to find an optimal thickness. This procedure created 24 layers ranging in pressure-thickness from 25 to 400 mb. The correlation to lightning occurrence was calculated for each layer. This procedure yielded the average RH in the 825–525 mb layer as the most highly correlated to lightning occurrence in the warm season. This is close to the original 200-mb thick layer of 800–600 mb, which is centered at 700 mb. The new layer is 300 mb thick and centered at 675 mb, 25 mb higher than the former layer.

4. Equation Elements

The three datasets described in Section 2 were processed in the same way as in Phase I, except with the modifications described in Section 3, 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 CGLSS data provided the ground truth indicating whether or not lightning occurred and were used to create the predictand as well as the daily climatology, persistence, and flow regime lightning probability candidate predictors. The sounding datasets were used to calculate the stability index and flow regime lightning probability candidate predictors.

4.1 Binary Predictand

The CGLSS data were filtered spatially to include only strikes that occurred within the 10 5-n mi warning circles as described in Section 3.2 and shown in Figure 3. Then they were filtered temporally as they were in Phase I to include only lightning strikes recorded in the time period 0700–0000 EDT. The 45 WS morning forecast is created by 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). They consider times after midnight as Day 2. Since the goal of this task was to develop equations for Day 1 forecasts, lightning occurring between midnight and 0700 EDT were not considered.

Once the data were filtered, the value of the binary predictand 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 is 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.

4.2 Candidate Predictors

The list of candidate predictors for Phase II was the same as in Phase I. They 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 persistence and daily climatological lightning frequency calculated from the CGLSS binary predictand, the flow regime probabilities from the soundings and CGLSS binary predictand, and 10 stability parameters calculated from the XMR sounding.

4.2.1 CGLSS Predictors

The binary predictand discussed in Section 4.1 was used to create two candidate predictors: a binary persistence and the daily climatological probability of lightning occurrence described in Section 3.4. Calculation of the persistence predictor 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 30 April was used to create the persistence value for 1 May in each year. A persistence value was created for each individual day in the POR.

The values along the thick blue curve in Figure 5a, created using the ±14-day Gaussian smoother described in Section 3.4, were used for the daily climatological values of lightning probability. The new valid area had a significant effect on the daily climatology values. They were on the order of 10% lower than those in Phase I due to the large reduction in the valid area from Phase I to Phase II (see Figure 3), and the associated spatial gradient in the annual CG flash density (Figure 4).

4.2.2 Flow Regime Probabilities

The AMU calculated the 1000–700 mb layer-average winds and determined a flow regime for each day using the morning JAX, TBW, MFL, and XMR soundings as described in Section 3.3. Then, 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 compared to 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.

It was clear in Phase I that the flow regime lightning probabilities were good predictors of lightning occurrence over KSC/CCAFS when used alone. The same was true for the probabilities calculated in Phase II. As in Phase I, the new probabilities were transitioned for immediate operational use. The details of how these values were calculated and other aspects of the tables are contained in an AMU Memorandum (Lambert 2006). Six tables in the same format as those in Phase I 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 3 is an example of their content. It contains the lightning statistics by flow regime for the entire warm season.

Table 3. 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 – 2005

Probabilities of lightning occurring within 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 MFL	21, 117, 281 (226, 338)	301 (11.8)	113	188	**62 %**
SW-2 Ridge between MFL/TBW	13, 66, 252 (186, 294)	606 (23.8)	260	346	**57 %**
SE-1 Ridge between TBW/JAX	2, 9, 35 (51, 135)	438 (17.2)	299	139	**32 %**
SE-2 Ridge N of JAX	1, 6, 25 (33,87)	248 (9.8)	183	65	**26 %**
NW	13, 75, 277 (186, 257)	307 (12.1)	209	98	**32 %**
NE	3, 10, 38 (38, 83)	317 (12.5)	283	34	**11 %**
Other (Regime Undefined)	6, 24, 135 (100, 166)	326 (12.8)	213	113	**35 %**
TOTALS	**7, 38, 185 (150, 264)**	**2543**	**1560**	**983**	**39 %**

There is a 12% improvement in the forecast when using the individual flow regime probabilities over the seasonal climatological probability of 39%, and a 56% improvement over 1-day 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 the median and inter-quartile ranges are more representative of the data than the mean and standard deviation.

The first (left to right) column in Table 3 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 the total number of 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 used by the forecasters, and was also a candidate predictor for the equations. The TOTALS row in Table 3 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. In each of the monthly tables, this value is the monthly climatology. There is further information found in the notes in the last row of Table 3. 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 flow regime lightning probability values for the individual months were used as candidate predictors in the equation development and the 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 (2006) and are shown in Table 4. 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 SW-1/2 values for June, July, and August were also within 10% of each other, therefore 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 4. The parentheses around the SE-2 values for June–August indicate that it is a combined value and the same as SE-1.

Table 4. 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	SW-1/2	SE-1	SE-2	NW	NE	Other	Monthly
May	30	19	6	16	2	16	18
June	68	32	(32)	46	11	30	46
July	68	32	(32)	53	14	43	48
August	68	32	(32)	38	12	55	49
September	55	42	29	16	17	24	33

4.2.3 Stability Indices

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 for the Phase I task. All the routines that the AMU developed in Phase I to create the stability indices were used in Phase II.

The stability index candidate predictors included the

- Total Totals (TT),
- Cross Totals (CT),
- Vertical Totals (VT),
- 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 RH in the 825–525 mb layer (MRH), and
- Precipitable water (PW),

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). The formulations will not be shown here. Only three indices in the above list are not readily available to the forecasters: VT, TI and MRH. The TI is calculated easily with the equation TI = KI – LI, as is VT with T_{850} - T_{500}. The MRH was calculated using a log(p)-weighted average described in Section 3.5 (Equations 5 and 6).

4.2.3.1 Candidate Predictor Test

Before using the 11 candidate stability index predictors from the list above in the equation development, the AMU performed a test to ensure their validity as predictors. An index that did not pass the test would not be used as a candidate predictor. The indices were stratified by month, and then stratified between lightning and non-lightning days. Mean values for each of the 11 stability indices were calculated separately for the lightning and non-lightning days, 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. To check whether the differences were statistically significant, the AMU used a two-sample two-sided Student's t-test (Wilks 2006) in S-PLUS. This form of the Student's t-test determines 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 the 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 for each stability index in each individual month and for all months combined and produced p-values less than 0.01 for each stability parameter listed under Section 4.2.3. Therefore, the null hypothesis for all the stability parameters could be rejected at the 99+% confidence level, indicating that the differences in their means were statistically significant.

4.2.3.2 A Word about CAPE and CIN

Issues pertaining to using the Convective Available Potential Energy (CAPE) and Convective Inhibition (CIN) variables as candidate predictors are detailed in the Phase I final report. Even so, they were calculated and tested in the Phase II work in case more data helped make them useful predictors, since they are commonly used as predictors of convection in other areas of the U.S. However, similar results to those in Phase I were found. One main difficulty in using CAPE and CIN as predictors was that their values were not able to be calculated for every sounding. The McIDAS code needs to calculate a level of free convection (LFC) before calculating CAPE and CIN. If an LFC was not found, the values were not calculated. This artifact of the code resulted in reducing the available dataset by over 10% beyond that accounted for by missing data. Also, the p-values from the Student's t-test for these indices were between 0.5–0.9 (50–10% confidence level), 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. The availability of a low-level trigger is more important in the Florida warm season, such as the east and/or west coast sea breeze fronts that occur with the afternoon maximum in solar heating. The role of the flow regimes mentioned in Sections 3.3 and 4.2.2 is to parameterize the steering flow for the sea breeze fronts and the timing of their in-land positions.

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 caused by the code not being able to calculate an LFC. Therefore, all the stability indices except CIN and the three CAPE values were used as candidate predictors.

4.2.4 Summary of Candidate Predictors

A summary of the candidate predictors is given here as a reference for the reader. They are

- Persistence,
- Daily climatological lightning frequency,
- Flow regime lightning probability,
- Total Totals (TT),
- Cross Totals (CT),
- Vertical Totals (VT),
- K-Index (KI),
- Lifted Index (LI),
- Thompson Index (TI),
- Severe Weather ThrEAT (SWEAT) Index,
- Showalter Index (SSI),
- Temperature at 500 mb, (T_{500}),
- Mean RH in the 825–525 mb layer (MRH), and
- Precipitable water (PW).

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

5. 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.

5.1 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.

5.1.1 Missing Data

There are 153 days in the warm season, 1 May–30 September. This equates to 2601 days over the 17-year POR. Sounding data were not available for every day in the POR. Data were considered missing for a specific day if one or more of the 1200 UTC Florida synoptic soundings (MFL, TBW, JAX) and the 1000 UTC XMR sounding were missing to determine the flow regime, or when a 1000 UTC XMR sounding was missing to calculate the stability parameters. Table 5 shows a summary of how many days were in the POR, how many of those days had missing data, which dataset was considered missing, and the total number of days with available data. There were few cases in which data were missing from both datasets on the same day. The number in the third column under the heading '# Missing Obs' in Table 5 is less than the sum of the first two columns in every case because there were a few 'overlap' days in which data were missing from both datasets. 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 258 (58 + 200), but the total missing is 237. This says that data were missing from both datasets on the same day only 21 times.

The final column in Table 5 shows that data availability ranged from 89–94% for each month, and 91% overall. These percentages are higher than those from Phase I, which ranged from 85–90% and was 87% overall, most likely due to the fact that the XMR sounding was used to determine the flow regime if one or more of the Florida synoptic soundings were missing. Most important, though, was the actual number of available days per month, ranging from 452 to 494. 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 verification datasets.

Table 5. Summary of available data in the POR. The first column contains the names of the months 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 and 17 warm seasons. The three columns under the heading '# MISSING DAYS' show the number of unavailable days due to missing data from each dataset in the subheadings, and the number of days missing due to the combined missing data from both datasets. The value in parentheses in the third column is the number of days in which data were missing from both datasets. 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	17 Years	MFL TBW JAX XMR	XMR	Total (Overlap)	
May	31	527	9	29	33 (5)	494 (94)
June	30	510	16	32	43 (5)	467 (92)
July	31	527	14	38	47 (5)	480 (91)
August	31	527	10	50	56 (4)	471 (89)
September	30	510	9	51	58 (2)	452 (89)
Total	**153**	**2601**	**58**	**200**	**237 (21)**	**2364 (91)**

5.1.2 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 candidate predictors and predictand for each month were 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 17 warm seasons in the POR, 14 were used for equation development and 3 were set aside for equation verification. 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 Excel was used to create three sets of 153 numbers representing the years between and including 1989 and 2005. The resulting three sets of years were assigned to each day in the warm season. Thus, each day in the warm season was represented by days from three random years. For example, the verification dataset contains 1 May 1989/1999/2001, 2 May 1993/1998/2000, etc. Care was taken to ensure the there were no duplicate years for each day from the random number generator. 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 6 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 6. Summary of missing and available data for equation development and verification. 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 17 warm seasons, the number of those days for equation verification, 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 (Table 5), 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	527	93	434	494	90	404
June	510	90	420	467	81	386
July	527	93	434	480	84	396
August	527	93	434	471	84	387
September	510	90	420	452	84	368
Total	**2601**	**459**	**2142**	**2364**	**423**	**1941**

5.2 Equation Development

As in Phase I, five logistic regression equations were created, one for each month. 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.

5.2.1 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 (2006), logistic regression is the appropriate method when the predictand is binary. Logistic regression was chosen as the statistical method in Phase I due to the binary nature of the predictand and also due to results from a previous study. Everitt (1999) 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

$$y = \frac{e^{(b_0 + b_1x_1 + \ldots + b_kx_k)}}{1 + e^{(b_0 + b_1x_1 + \ldots + b_kx_k)}}, \tag{7}$$

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 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 7, 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.

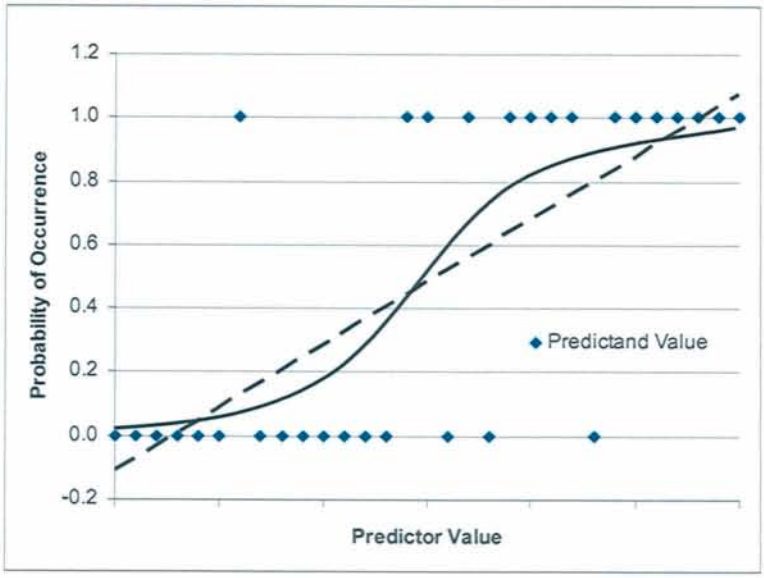

Figure 6. Illustration of linear (dashed line) vs. logistic (solid curve) regression probability forecasting for a binary predictand and one predictor. The blue diamonds represent the predictand values at certain predictor values. The forecast probability values are along the y-axis. The predictor values along the x-axis are assumed to increase monotonically to the right (similar to Wilks [2006] Figure 6.12).

5.2.2 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 2005b). 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 Phase I and II.

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 7. As proven in Phase I, the NULL equation predicts the monthly climatology as found in the development dataset. The residual deviance for the NULL equation is calculated with the general equation

$$\text{Residual Deviance} = -2*[\log(y)*(\#\text{yes}) + \log(1-y)*(\#\text{no})], \tag{8}$$

where y is the probability calculated by Equation 7, #yes is the number of days with lightning and #no is the number of days with no lightning. Equation 8 becomes more complex when other predictors are added. As each predictor is added, the residual deviance is reduced from the NULL value.

5.2.3 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.

5.2.3.1 Residual Deviance Check

The values for the predictor coefficients in a logistic regression equation (Equation 7) cannot be solved analytically, but must be estimated using computationally intensive iterative techniques (Wilks 2006) 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 (Insightful Corporation 2005a) 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 14 candidate predictors as a lone predictor in Equation 7, resulting in 14 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 13 candidate predictors were added individually with the first predictor creating a set of 13 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 12 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 14 predictors. Figure 7 shows the percent reduction in residual deviance from the NULL model for the first eight predictors added for the month of June. The TI reduced the residual deviance by the most (19%) and was, therefore, the first predictor in the June equation. The second predictor was the flow regime lightning probability (FRProb in Figure 7), which accounted for an additional 9% reduction in residual deviance. The third predictor was persistence (Pers in Figure 7), reducing the residual deviance by 1%, and so on.

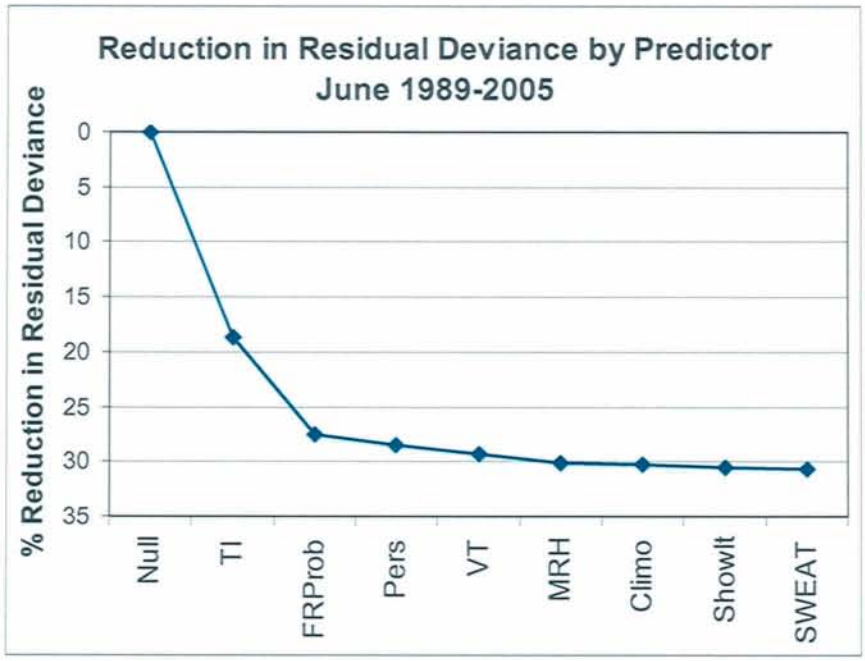

Figure 7. The total percent reduction in residual deviance from that of the NULL model as each predictor was added to the equation using the June development dataset.

All predictors were used in each equation to determine their rank order, but not all could be used in the final equations as that could lead to overfitting (Wilks 2006). When this happens, the equations will perform well with data from the development dataset, but will perform poorly on data not used to create the equations. Wilks (2006) suggests several 'stopping rules', the point at which no more predictors will be added to the equation. In Phase II, the AMU used a similar method to that in Phase I. Charts showing the reduction in residual deviance, like Figure 7, were created for each month (not shown). The AMU then examined the change in slope of the curves as each predictor was added. The change in the slope of the curve at MRH in Figure 7 is at the point where the residual deviance is reduced by less than 0.5%. Similar changes in slope were found in the other four months for the same cutoff value, so the stopping rule was that when a candidate predictor effected a < 0.5% change in the residual deviance, it was not added to the equation nor were the candidate predictors that came after it. In the above example, the first five parameters were selected (TI through MRH). The AMU tested equations with one more and one less predictor than the stopping rule indicated, only to conclude that the equation created with the < 0.5% stopping rule had superior performance.

5.2.3.2 Predictor Interaction

Another reason for using a manual process to choose predictors was to monitor the process to make sure that predictors that had a strong mutual correlation were not chosen for the same equation. This could create unrealistic results (Wilks 2006). Most of the stability candidate predictors had some level of correlation to each other since they were calculated from the same soundings on the same day using some of the same sounding variables. However, there were two sets of three candidate predictors that had a very close mathematical relationship and were watched closely to ensure they did not mix in one equation:

- TI, KI, and LI; and
- TT, CT, and VT.

The relationship between these predictors are TI = KI – LI and TT = CT + VT. For the first set, if TI was chosen as a predictor, as in the June equation (Figure 7), then KI and LI were no longer considered as candidate predictors for that equation. If KI was chosen, LI could still be chosen but TI could not; and if LI was chosen, KI could still be chosen but not TI. The second set of candidate predictors had similar rules: if TT was chosen as a predictor, then CT and VT were no longer considered as candidate predictors for that equation. If CT was chosen, VT could still be chosen but TT could not; and if VT was chosen, CT could still be chosen but not TT.

When one of the predictors in the TI/KI/LI set was chosen, the routines in S-PLUS indicated that the correlated predictor(s) did not reduce the residual deviance enough to be considered as a final predictor, so predictor interaction was not an issue for this set. It was an issue for the TT/CT/VT predictor set in the July equation, where TT and CT were chosen as predictors that reduced the residual deviance by > 0.5%. The first of the three to be chosen was TT, followed immediately by CT. Both predictors have a positive correlation with lightning occurrence in which the probability of lightning increases as their values increase. This was also found to be true with the data in the POR. It would follow that the coefficients determined by the logistic regression for each of these predictors should be positive, but the coefficient for CT in the July equation was negative. The equation was redeveloped by not considering CT as a candidate predictor and tested against the equation containing CT. The non-CT equation outperformed the CT equation for both the development and verification datasets, helping to prove that such closely related predictors should not be used in the same equation.

5.2.3.3 Final Set of Predictors

Table 7 shows the final predictors for each of the monthly equations in rank order of their reduction in residual deviance. The predictor names are color-coded according to the number of equations in which they appear. Red indicates that a predictor was chosen in every equation. There was only one: the probability of lightning occurrence based on the flow regime (Table 4). It was also ranked second in every month, underscoring its importance as a predictor in the KSC/CCAFS area. Blue identifies the two predictors, VT and persistence, that were chosen in four of the five equations. The July equation did not use VT and the August equation did not use persistence. The July equation did include VT indirectly, given that it is in the equation that calculates TT. The predictors in green were chosen for three of the equations, and they are the daily climatology, TI, and MRH. These are followed by the predictors in black, which were only used in one equation each: KI and TT. The June, July, and August equations did include KI indirectly in the equation for TI.

The most important predictors in the May through August equations, KI and TI, account for instability and moisture in the profile, which are both necessary ingredients for thunderstorm formation. September has MRH as the most important predictor, which only accounts for the mid-level moisture. The fourth predictor, VT, accounts for mid-level instability, but it has a much smaller influence on the probability in September due to its rank. The flow regime probability as the second predictor accounts for the lifting mechanism, or lack thereof, as the low-level flow interacts with the sea breeze that occurs almost daily in the warm season.

Table 7. The final predictors for each monthly equation, in rank order of their reduction in residual deviance. The predictors in red were in every equation, the predictors in blue were in four of the five equations, the predictors in green were in three of the five equations, and the predictors in black were in only one equation.

May	*June*	*July*	*August*	*September*
K-Index	Thompson Index	Thompson Index	Thompson Index	825–525 mb MRH
Flow Regime	Flow Regime	Flow Regime	Flow Regime	Flow Regime
Vertical Totals	Persistence	Total Totals	Daily Climatology	Persistence
Daily Climatology	Vertical Totals	Persistence	825–525 mb MRH	Vertical Totals
Persistence	825–525 mb MRH		Vertical Totals	Daily Climatology

5.3 Equation Performance

The predictors from the three-warm-season verification dataset were used in the equations to produce 'forecast' probabilities. Using the verification dataset provided an independent assessment of equation performance that could be used to conclude how the equations will perform in future operations. The forecast 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 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 81–90 (Table 6). 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.

In several of the tests, the AMU used the probabilities produced by the equations developed in Phase I, hereafter designated as the P-1 equations, as a forecast benchmark to determine if the new equations create an improved forecast. The new equations, hereafter designated as the P-2 equations, were created with dataset that had undergone the five modifications described in Section 3. This made it difficult to calculate direct and fair differences in skill between the P-2 and P-1 equations. The AMU had several discussions with Mr. Roeder of the 45 WS to determine the best approach to ensure a fair comparison in skill between the equation sets. They decided that the input parameters for the P-1 equations should include data from the new POR and reflect the new area, since the new area represents the warning areas exactly. This included using the predictand, persistence, daily climatology, and the flow regime lightning probabilities calculated for the new area in Phase II. However, they decided to use flow regime lightning probabilities calculated using the three-sounding procedure in the P-1 equations since that was the procedure employed in the development of these equations.

5.3.1 Brier Skill Score

The first test was to determine if the P-2 equations showed improvement in skill over five forecast benchmarks:

- Persistence,
- Daily climatology (Figure 5a),
- Flow regime probabilities (Table 4),
- Monthly climatology (Table 4), and
- P-1 equation probabilities.

The AMU began the skill test by first calculating the mean squared error (MSE) between the forecasts and observations for all forecast methods. The MSE was calculated using the equation

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 \text{ (Wilks 2006),} \tag{9}$$

where n is the number of forecast/observation pairs, p_i is the probability associated with the forecast method, and o_i is the corresponding binary lightning observation. The skill of the P-2 equations over the five forecast benchmarks was calculated using the equation for the Brier Skill Score (SS):

$$\mathrm{SS} = \left(\frac{\mathrm{MSE}_{eqn} - \mathrm{MSE}_{ref}}{\mathrm{MSE}_{perfect} - \mathrm{MSE}_{ref}}\right) * 100 \text{ (Wilks 2006),} \tag{10}$$

where MSE_{eqn} was the MSE of the P-2 equations, MSE_{ref} was the forecast benchmark against which the new equations were tested, and $\mathrm{MSE}_{perfect}$ was the MSE of a perfect forecast, which is always 0. The SS represents a percent improvement or degradation in skill of the equation over the reference forecast when it is positive or negative, respectively.

The SS values for each of the monthly P-2 equations and a composite result for the entire warm season are shown in Table 8. The P-2 equations show a double-digit improvement in skill for the first four benchmarks in the table, similar to the results for the P-1 equations in Lambert and Wheeler (2005). Of the first four benchmarks, the smallest percent improvements were over the probabilities based on flow regime. The P-2 equations also show an 8% improvement in skill over the P-1 equations for the entire warm season. For the individual months, the P-2 equations show an improvement in skill over the P-1 equations for June, July, and September. The values of 0.2% for May and -0.8% for August are very small and indicate similar skill between the two equation sets.

Table 8. The SS values that show the percent (%) improvement (degradation) in skill of the P-2 equations over the reference forecasts of persistence, daily and monthly climatologies, flow regime probabilities, and the P-1 equations developed in Lambert and Wheeler (2005). These scores were calculated using the verification data for each month and for the entire warm season (All).

Forecast Method	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*All*
Persistence	28	41	37	47	41	40
Daily Climatology	23	25	24	24	26	25
Monthly Climatology	29	27	34	30	25	29
Flow Regime	16	12	11	18	18	15
P-1 Equations	0.2	5	19	(-0.8)	12	8

5.3.2 Probability Distributions

The AMU stratified the P-1 and P-2 probability forecast sets by lightning and non-lightning days and created a probability distribution for each. These distributions showed the percent occurrence of each probability value for the lightning and non-lightning days. Such distributions show how well the equations distinguished between lightning and non-lightning days in the verification data set. Figure 8 shows the probability distributions for lightning days, represented by the two red curves, and non-lightning days, represented by the two blue curves. For good performance, one would expect the blue curves to have a maximum in the lower probability values decreasing to a minimum at higher probability values, and the red curves to have a minimum in the lower probability values increasing to a maximum at the higher values.

Both blue curves for the non-lightning days in Figure 8 peak at a probability of 0.2, decrease rapidly through 0.4, and then decrease more slowly toward 1. This indicates good performance for both equation sets. However, the P-1 equations distinguished non-lightning days with a bit more accuracy as evidenced by the higher peak of 59% versus 50% at 0.2 probability and the larger drop off to 21% versus 24% to 0.4. The percent occurrence for the P-1 equations remained ~1% below those of the new equations from 0.4 to 1. In Phase I, this curve had a secondary maximum at 0.8, indicating a possibility of increased false alarms. That secondary maximum is no longer evident in either the P-1 or P-2 non-lightning day forecasts. This improvement could be due to the five modifications described in Section 3.

The red curve for the P-2 equations indicates that they distinguished lightning days more accurately than the P-1 equations. The percent occurrences of the P-2 equation probabilities were lower than those for the P-1 equations for all probability values less than 0.7, and higher for all probabilities greater than 0.7. The peak percent occurrence for the P-2 equations was 36% at 0.8 probability, while the peak for the P-1 equations was 30% at 0.6 probability.

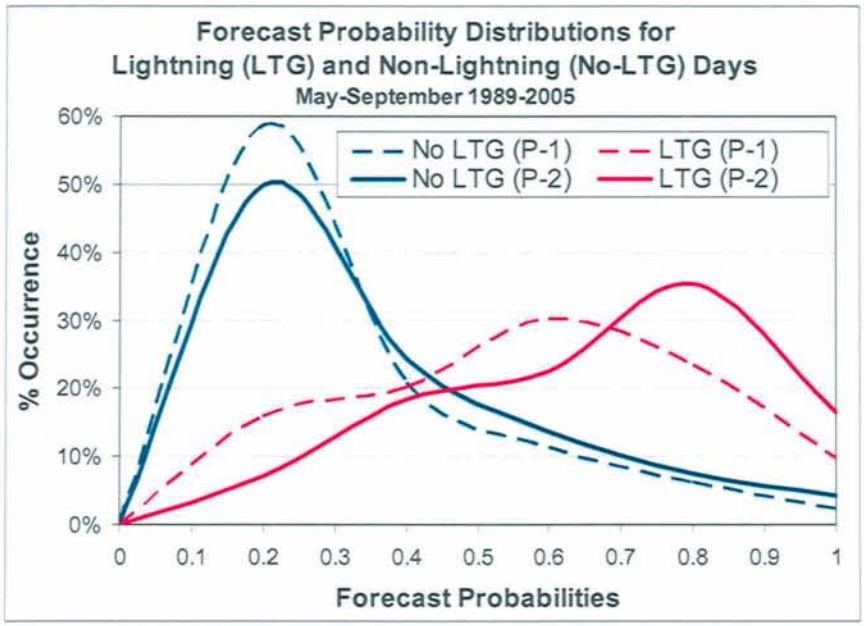

Figure 8. Forecast probability distributions for lightning (red) and non-lightning (blue) days in the verification data. The solid lines represent the P-2 equations and the dashed lines represent the P-1 equations. The y-axis values are the frequency of occurrence of each probability value, and the x-axis values are the forecast probability values output by the equations.

5.3.3 Reliability Diagram

The reliability diagram shows the distribution of forecasts and their associated observations. It indicates how the equations perform in terms of under- or over-forecasting lightning occurrence at discrete probability values from 0 to 1 in increments of 0.1. However, the equations do not output probability values in such discrete intervals. Therefore, the probability values were organized into bins according to their rounded value. For example, a

probability value of 0.23 would be added to the 0.2 probability bin, as would a probability of 0.17. The probability values themselves were not changed. The number of 'yes' lightning observations within each bin were divided by the total number of forecast/observation pairs in each bin to get a reliability value for that bin. For example, if there were 10 forecast/observation pairs assigned to the 0.1 bin and one of the observations was 'yes' for lightning, the reliability would be [1 'yes' observation] / [10 forecast/observation pairs] or 0.1. The forecast is said to exhibit perfect reliability when the calculated reliability is equal to the bin value.

The reliability diagrams for the P-1 and P-2 equations are shown in Figure 9. The black diagonal line represents perfect reliability, and the histogram in the lower right shows the number of observations in each probability bin for each method. Where the curves are below the black line, the equations over-forecasted lightning occurrence, and where the curves are above the line, the equations under-forecasted lightning occurrence. Both curves are mostly above the black line, indicating a tendency to under-forecast lightning occurrence. For example, when the P-1 equations calculated a 0.4, or 40%, probability, lightning occurred 60% of the time. The red curve for the P-2 equations was closer to the perfect reliability diagonal than the blue curve for the P-1 equations for all probabilities except for 0.6 and 0.7. However, the frequency values for each forecast method at these probabilities were within 10% of each other. The number of samples used to calculate the reliability decreased with increasing forecast probability, which may account for the 20% over-forecast at 0.9 forecast probability. Overall, these curves demonstrate that the P-2 equations have better reliability than the P-1 equations.

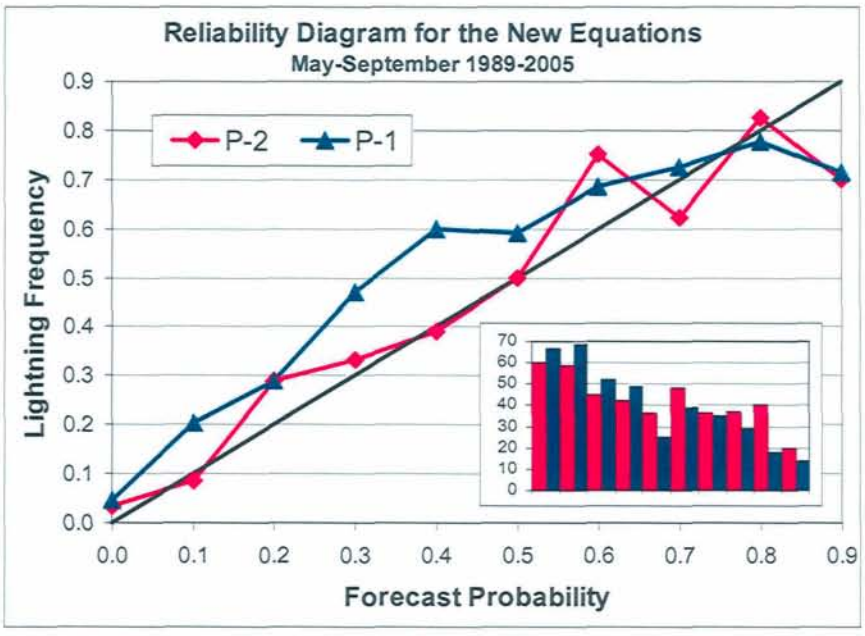

Figure 9. Reliability diagram of the P-1 and P-2 probability forecasts for all months. The straight diagonal line represents perfect reliability, the blue curve represents the reliability of the P-1 equations, and the red curve represents the reliability of the P-2 equations. The histogram at the lower right shows the number of observations in each probability range for the old (blue) and new (red) forecast methods.

The reliability diagram in Figure 9 shows a tendency for both sets of equations to under-forecast lightning, but a visual inspection indicates that the magnitude of the under-forecasting was less for the P-2 than the P-1 equations. The AMU quantified the extent of the under-forecasting by calculating the bias for each equation set. The forecasts and observation pairs for the entire warm season were used in the bias calculation. The average bias in percent was calculated using the equation

$$B = \frac{\sum_{i=1}^{i=N}(p_i - o_i)}{N} * 100, \qquad (11)$$

where B is the bias in percent, o_i is the binary lightning observation, p_i is the associated probability forecast, and N is the number of observation/forecast pairs. A negative value would show a tendency to under-forecast. The bias was -5.9% for the P-1 equations and -0.4% for the P-2 equations. The P-2 equations reduced the bias by 4.5%.

5.3.4 MSE Decomposition

Wilks (2006) describes a decomposition of the Brier Score, his nomenclature for the MSE in Equation 9 (Section 5.3.1). The term MSE will be used here for consistency in this report. Without going through the derivation, which can be found in Wilks (2006), the MSE can be defined as the sum of three terms:

$$\mathrm{MSE} = \left[\frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \bar{o}_i)^2\right] - \left[\frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2\right] + \left[\bar{o}(1-\bar{o})\right], \tag{12}$$

where

n	=	total number of observation/forecast pairs,
I	=	the number of discrete forecast probability values,
N	=	the number of forecasts in each discrete i probability bin,
p_i	=	the discrete probability value of each discrete i probability bin,
$\bar{o}_i$	=	the average observation value in each discrete i probability bin, and
$\bar{o}$	=	the average value for the total n observations.

The first term in brackets on the right hand side of Equation 12 is called the reliability, the second is the resolution, and the third is the uncertainty. Since a smaller MSE is indicative of a more accurate forecast, the reliability term should be as small as possible and the resolution term as large as possible. The uncertainty term reflects the climatology of the observations and is unaffected by the forecasts. The values calculated for the reliability diagram in Section 5.3.3 were used in Equation 12. There were I=11 probability values, p_i, from 0 to 1 in increments of 0.1. The actual probability values were not used since the formulation of Equation 12 does not account for probabilities other than the discrete values. The values of N_i for the P-1 and P-2 equations are those used to create the histogram inset in Figure 9.

The resulting reliability, resolution, uncertainty, and MSE values using Equation 12 for the P-1 and P-2 equation sets are shown in Table 9. There was only a small difference in the total number of forecast/observation pairs between the P-1 and P-2 verification data sets. This allowed the AMU to conclude that differences in values were most likely due to differences in forecast performance rather than a large discrepancy in the number of samples. The difference in the uncertainty terms between the P-1 and P-2 equations was also negligible, meaning the differences in MSE values would be due differences in reliability and resolution, both related to equation performance. The last row in Table 9 shows the percent change in the values from the P-1 to the P-2 equations. The P-2 reliability decreased by 50% and the resolution increased by 18% over the same values for the P-1 equations. Given that a smaller reliability term and a larger resolution term indicate better performance, this indicates an improvement in the forecast from the P-2 equations.

Table 9. The values of the terms in Equation 12 for the P-1 and P-2 equations. The last row shows the percent change in value from the P-1 to the P-2 equations.

Equation	*n*	*Reliability*	*Resolution*	*Uncertainty*	*MSE*
P-1	397	0.012	0.061	0.242	0.192
P-2	423	0.006	0.072	0.239	0.174
% Change	**6.5**	**- 50**	**18**	**- 0.6**	**- 9.4**

5.3.5 Contingency Table Statistics

The final test was to create a contingency table and calculate several accuracy measures and skill scores that would give further indication of equation performance. Forecast verification using a contingency table is most appropriate for categorical 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 2006). 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'.

The proper threshold, or cutoff, value depends on the forecast decision issue to which the user will apply the forecast (Wilks 2006). The original goal of Phase I was to create equations that perform better than persistence. The goal of Phase II is to improve the performance of the equations further. But to test that, the AMU had to determine the optimal threshold value for both the P-1and P-2 equations. In order to find the value for each equation set, the AMU used the condition from Phase I that the probability value chosen must outperform the persistence forecast for all of the contingency table values. Everitt (1999) produced graphs of the 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. As in Phase I, Everitt's procedure was followed here.

Table 10 shows an example of the contingency table with equations for the accuracy measures and skill scores (Wilks 2006). 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 hit rate (HR) is the percentage of correct yes or no forecasts, and the probability of detection (POD) is the percentage of 'yes' forecasts in the number of 'yes' observations. The false alarm ratio (FAR) is the percentage of 'no' observations in the number of 'yes' forecasts. The critical success index (CSI) is the percentage of correct 'yes' forecasts in the sum of all 'yes' forecasts and observations. The Heidke and Kuipers skill scores (HSS and KSS, respectively) 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.

Table 10. Basic contingency table for calculating categorical accuracy measures and skill scores (Wilks 2006). 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**

$n = a + b + c + d$

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

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

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

Figure 10 shows the contingency table values for persistence and the P-2 equation output probability values from 0–1 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 0.46–0.5, which is outlined by the vertical black lines in Figure 10. A similar table was made for the P-1 equations (not shown). The cutoff range for the P-1 equations was 0.35–0.4.

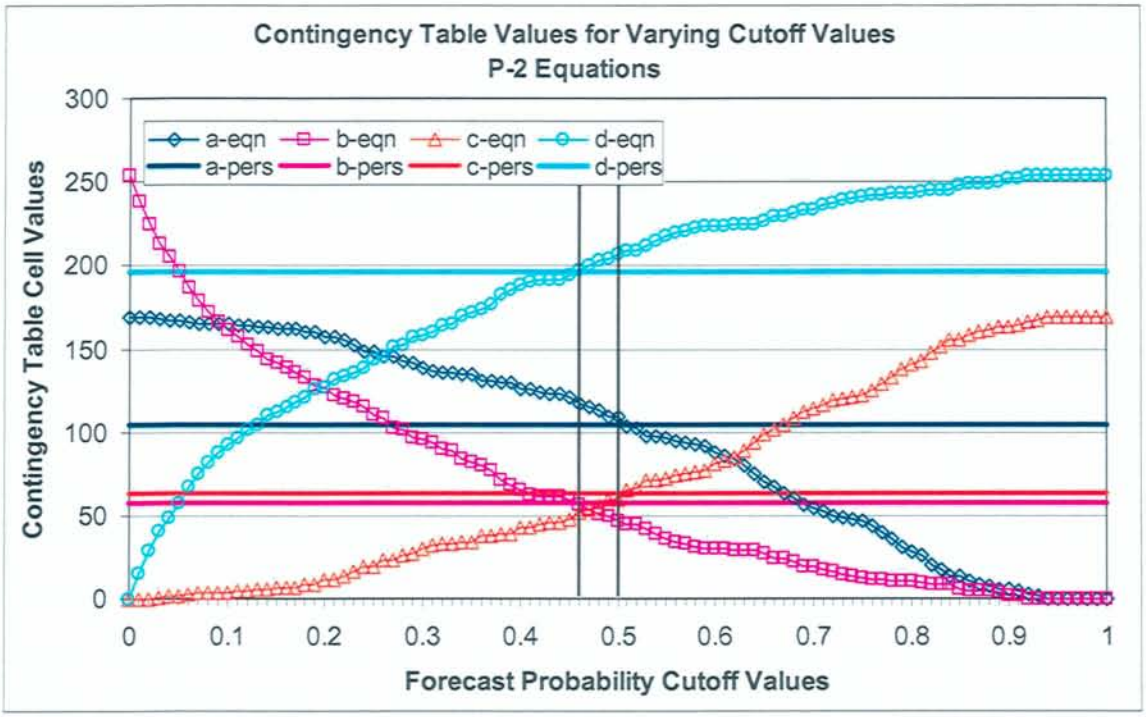

Figure 10. Graph showing the values in the four contingency table cells in Table 10 for the range of probability values 0–1 in increments of 0.01. Dark blue represents values in cell a, purple represents values in cell b, orange represents values in cell c, and cyan represents values in cell d. The horizontal straight lines represent the persistence forecast (pers) and the curves with symbols represent the P-2 equation forecasts (eqn). The vertical 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 AMU calculated the accuracy measures and skill scores shown in Table 10 for each probability in the range 0.46–0.5 for the P-2 equations and 0.35–0.4 for the P-1 equations to assist in determining which value should be the cutoff for each equation set. The HR was maximized and the bias ratio was exactly 1 at 0.47 for the P-2 equations, and the HR was maximized and bias ratio closest to 1 at 0.35 for the P-1 equations. A bias ratio of 1 means that an event was forecast as many times as it occurred (Wilks 2006). Therefore, all probabilities at or above 0.47 and 0.35 were considered 'yes' forecasts for the P-2 and P-1 equations, respectively. Probabilities less than these values were considered 'no' forecasts for the contingency table. Table 11 contains the accuracy measures and skill scores for each set of equations using the aforementioned cutoff values and those of their associated persistence forecasts. For every statistic in Table 11 except for FAR, a higher value indicates better forecast accuracy and skill. The P-2 equation set exhibited better accuracy and skill than its associated persistence and the P-1 equations. A perfect probability forecast technique would have 0.5 as the cutoff if it was unbiased. At 0.47, the cutoff for the P-2 equations is closer to 0.5 than the cutoff for the P-1 equations at 0.35. This also confirms that the P-2 equations were less biased than the P-1 equations.

Table 11. The accuracy measures and skill scores for the P-2 equations with a cutoff probability of 0.47, the P-1 equations with a cutoff probability of 0.35, and the persistence forecasts associated with each equation set.

Statistic	*P-2 (0.47)*	*Persistence (P-2)*	*P-1 (0.35)*	*Persistence (P-1)*
POD	0.68	0.62	0.66	0.63
FAR	0.21	0.23	0.23	0.24
HR	0.74	0.71	0.73	0.71
CSI	0.52	0.46	0.50	0.47
HSS	0.47	0.40	0.44	0.40
KSS	0.47	0.39	0.44	0.40

5.3.6 Equation Performance Summary

The main goal for this task was to create new lightning probability forecast equations that would outperform the P-1 equations currently used in operations. The new P-2 equations did outperform the P-1 equations as evidenced by the five tests described in Sections 5.3.1–5.3.5. The SS values indicated that the equations showed an increase in skill over daily and monthly lightning climatology, persistence, and the flow regime lightning probabilities. Three of the five P-2 equations showed a definite increase in skill over P-1 equations with August and May as the exceptions. The P-2 equation set performance in those two months was comparable to that of the P-1 set (Table 8). For the entire warm season, the P-2 equations showed an 8% increase in skill over the P-1 equations. Both sets of equations were able to distinguish between lightning and non-lightning days. The P-1 equations were slightly better at distinguishing non-lightning days, but the P-2 equations were better at distinguishing lightning days (Figure 8). The P-2 equations demonstrated an improved reliability over the P-1 equations, and reduced the overall negative bias by almost 5% (Figure 9). The MSE decomposition showed that the P-2 equations had improved reliability and resolution over the P-1 equations (Table 9). Finally, the P-2 equations had the best accuracy measures and skill scores compared it their associated persistence forecasts and the P-1 equations (Table 11).

Given that most of the tests indicated that the P-2 equations exhibited superior performance over the P-1 equations, they will replace those in current use before the start of the 2007 lightning season.

6. Graphical User Interface

In Phase I, the AMU created a GUI in Excel to facilitate user-friendly input to the equations and fast, easy-to-read output. 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 AMU updated the Excel GUI with the Phase II results and delivered it to the forecasters. The issue with the Excel GUI was that the forecasters had to gather the predictor values from one system and enter them in the GUI on a separate computer. This step used time that could be spent doing other required duties and increased the risk of entering an incorrect value, resulting in an erroneous probability value. Therefore, as part of Phase II, the AMU, assisted by Mr. Wahner of CSR, developed a similar GUI in MIDDS that gathers the required predictor values from the sounding automatically.

6.1 Excel GUI

The AMU updated the existing Excel GUI prior to development of a MIDDS tool. This got the new equations to the forecasters quickly so they could use them at the beginning of the 2007 warm season. This GUI was built within an Excel workbook using Visual Basic® for Applications. It accesses data in specific worksheets based on user input. The GUI has three dialog boxes: the first asks for the date, the second asks for equation predictor values, and the third displays the equation output.

6.1.1 Excel Workbook

The Excel workbook that contains the GUI has six worksheets. The first worksheet contains brief instructions on how to start and use the GUI. The AMU recommends first-time users to 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 coefficients,
- Flow regime names and their probabilities of lightning occurrence,
- Climatological lightning probability for each day,
- Minimum, maximum, median, mean, and first and third quartiles of the 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 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.

6.1.2 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.

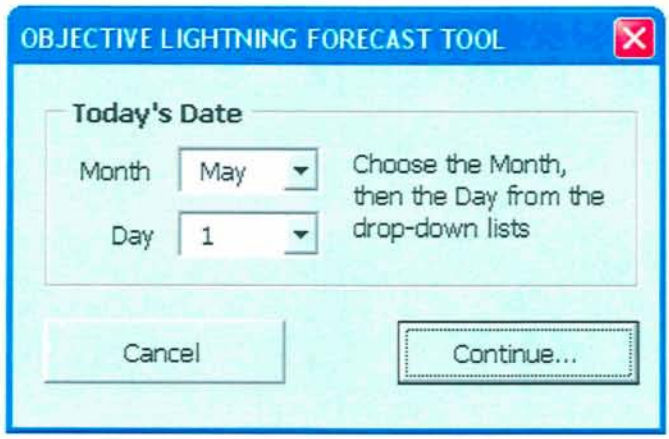

Figure 11. The first dialog box in the GUI queries the user for the Month and Day values. Month and Day are chosen by clicking on the down arrows next to each and choosing from the drop-down lists. The Cancel button exits from the GUI, the Continue button brings up the next dialog box.

6.1.3 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 equation predictor dialog boxes, one for each month since each has a different equation. The dialog boxes for each month are shown 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.

6.1.3.1 Flow Regime and Persistence

There is one feature common to all five equation predictor dialog boxes: a frame titled Flow Regime. 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. The user can choose only one item in the Flow Regime frame. Four of the dialog boxes contain a frame titled Persistence that allows the forecaster to make a choice for persistence, whether or not lightning occurred in the area the previous day. August is the exception. The user will choose 'Yes' or 'No' by clicking in the white circle next to the choice. The default choice is 'Yes'.

6.1.3.2 Sounding Parameters

The other predictors in the monthly 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 five parameters in different combinations for each month: TI, MRH, VT, TT, and KI. The first three are not output by the Skew-T program and must be calculated according to the equations

- $TI = KI - LI$,
- $VT = T_{850} - T_{500}$, and
- MRH – see Equations 5 and 6 in Section 3.5.

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 worksheet that is displayed 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 12. The summary values were calculated form the entire dataset in the POR 1989–2005.

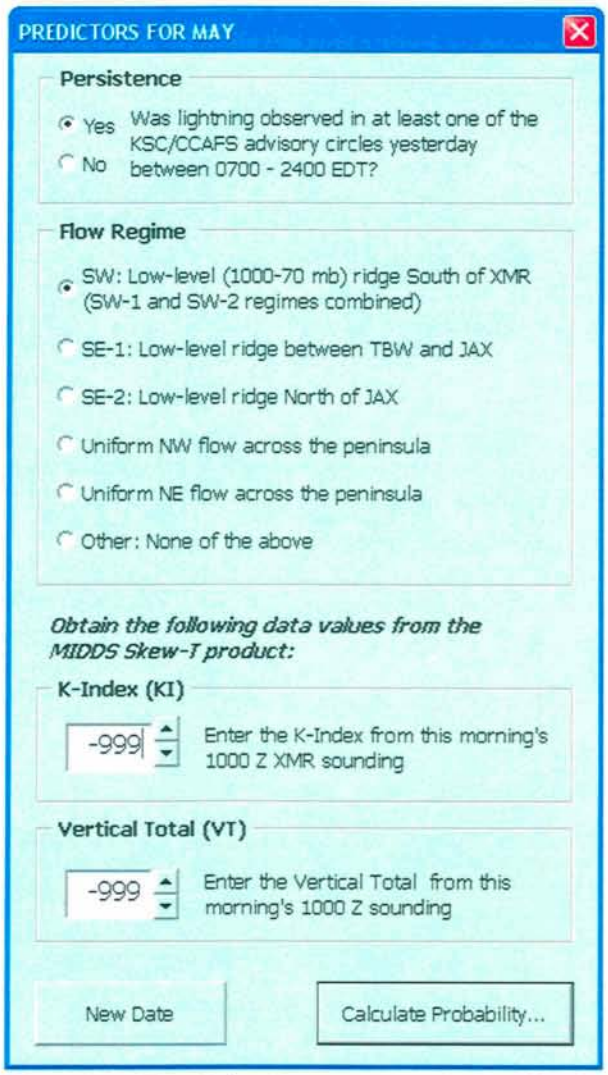

Figure 12. This dialog box contains choices for the predictors in the May equation. Persistence and Flow Regime are chosen by clicking one of the option buttons in each section. KI and V are chosen by entering their values manually or using the up/down arrows to the right of the text boxes. The 'New 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 6.1.4).

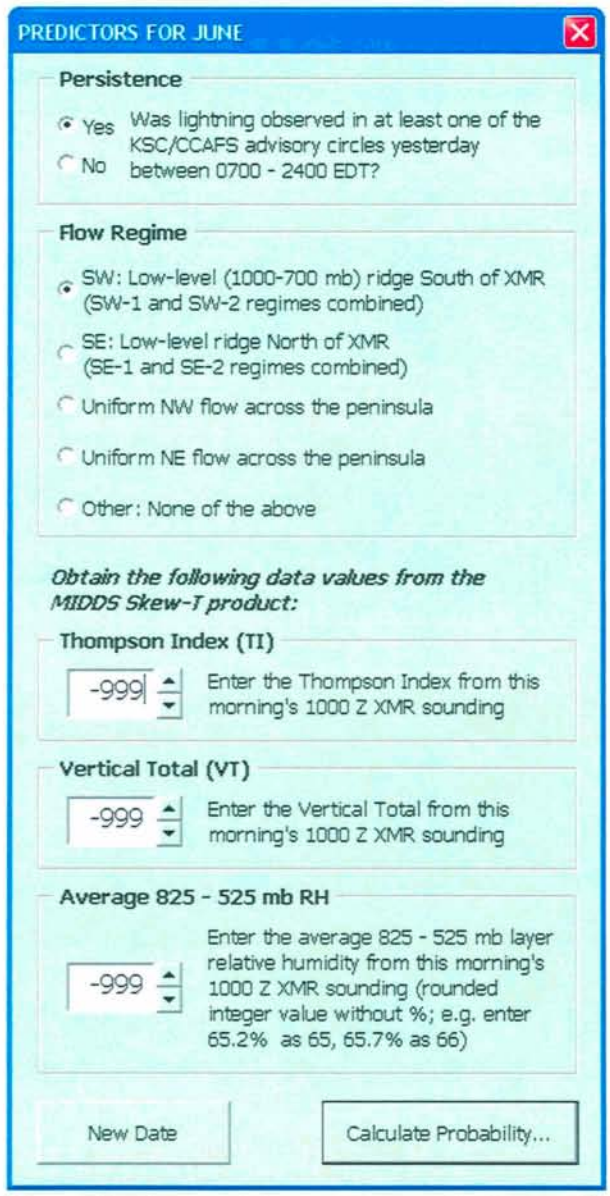

Figure 13. Same as Figure 12 except for June, with the SE-1 and SE-2 flow regimes combined into one SE flow regime, and with the sounding parameters TI, VT, and MRH.

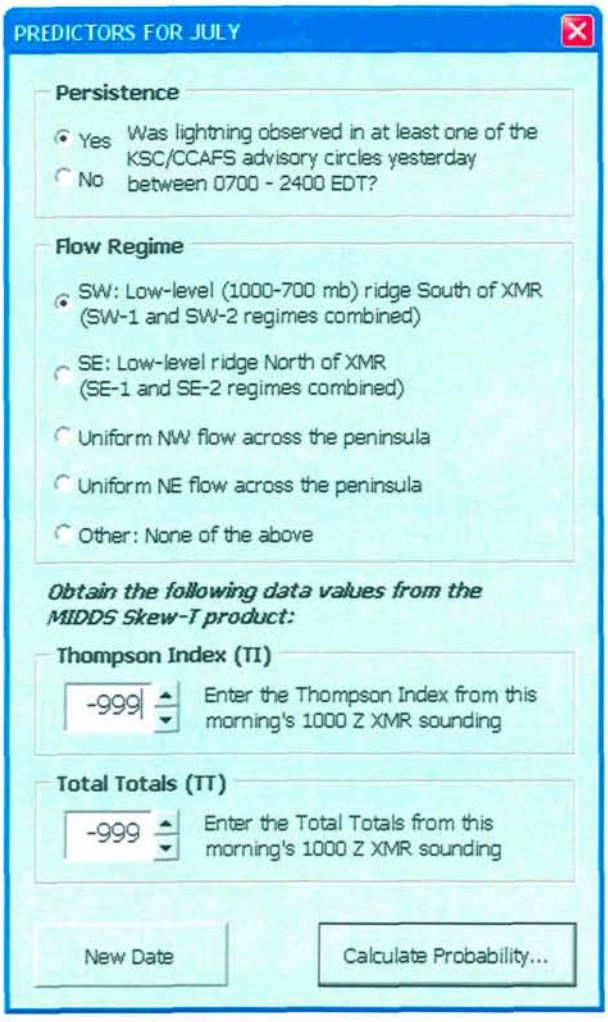

Figure 14. Same as Figure 13 except for July, and with the sounding parameters TI and TT.

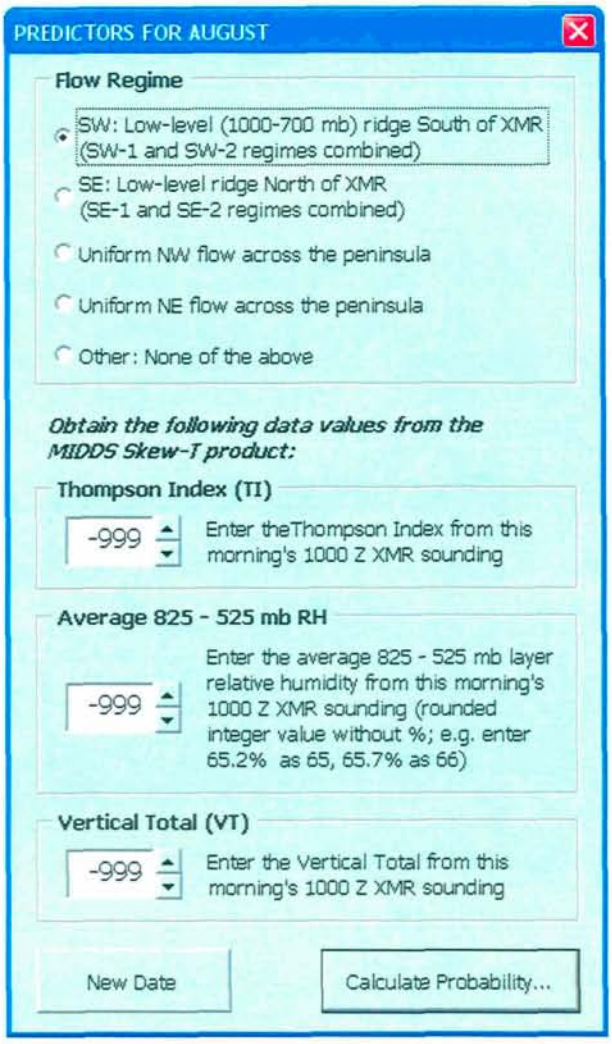

Figure 15. Same as Figure 13 except for August, and with the sounding parameters TI, MRH, and VT.

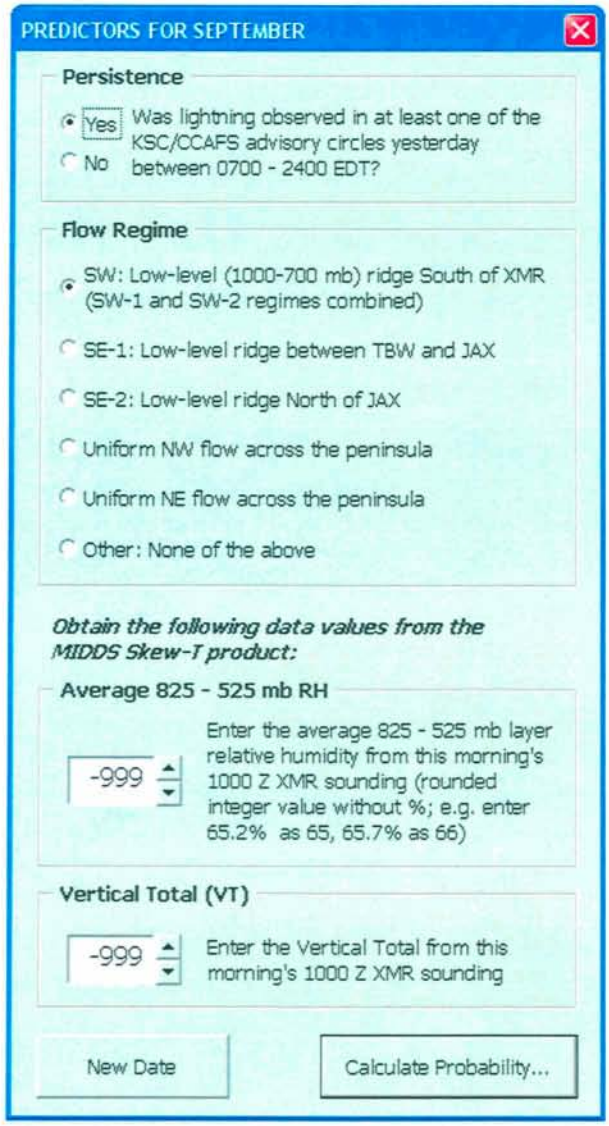

Figure 16. Same as Figure 12 except for September, and with the sounding parameters MRH and VT.

Table 12. Summary values for each of the predictors in the POR 1989–2005. 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*	
	KI	VT	TI	VT	MRH	TI	TT	TI	MRH	VT	MRH	VT
Minimum	-34	19	-10	17	15	1	23	-12	11	19	8	18
1st Quartile	9	23	26	24	45	28	43	28	47	24	42	23
Median	18	25	33	25	62	33	45	34	61	24	59	24
Mean	17	25	30	25	60	31	44	31	59	24	57	24
3rd Quartile	28	26	37	26	76	36	47	37	73	25	71	25
Maximum	40	33	48	31	98	49	53	50	94	30	96	27
Data Value Range Allowed in GUI												
Minimum	-70	-20	-30	-20	0	-30	0	-30	0	-20	0	-20
Maximum	70	70	70	70	100	70	80	70	100	70	100	70

6.1.4 Equation Output Dialog Box

After making all choices and entering all values in the equation predictor dialog box, the user will 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.

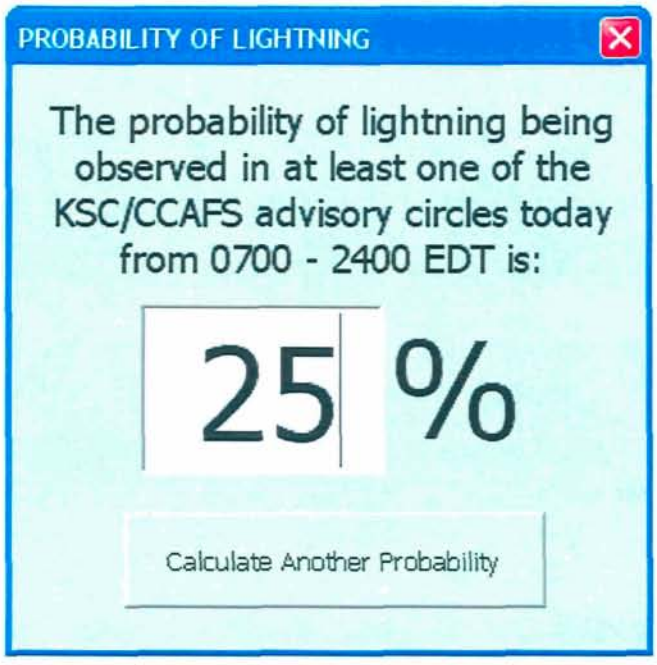

Figure 17. The equation output dialog box displaying the probability of lighting for the day based on the values input to the date and equation predictor dialog boxes.

6.2 MIDDS GUI

The MIDDS GUI was developed by Mr. Wahner of CSR using the Tool Command Language (Tcl)/Toolkit (Tk) capability in MIDDS. The design and function of this GUI is similar to the Excel GUI. It goes one step further by entering the sounding predictor values automatically into the dialog box. This removes the risk of having a forecaster enter an incorrect value while also reducing the time the forecaster would spend gathering and calculating the required parameter values. The GUI has two dialog boxes: the first displays the date and asks for equation predictor values, and the second displays the equation output.

6.2.1 Starting the GUI

This GUI is accessed through the MIDDS Toolbar by clicking on the 'FCST Tools' button and choosing 'Lightning Forecast Tool' from the drop-down list (Figure 18). This activates the GUI Tcl/Tk code to determine the date and gather the appropriate data for the equation from MIDDS. The code checks the time and date of the most recent sounding. If it does not match the current day and is within the time period 0900–1159 UTC, an error message stating that the data are missing is displayed (Figure 19). This ensures that data from the previous day and data from sounding times other than 1000 UTC are not used in the equations. The 0900–1159 UTC period allows for the fact that not all 1000 UTC soundings are released precisely at 1000 UTC.

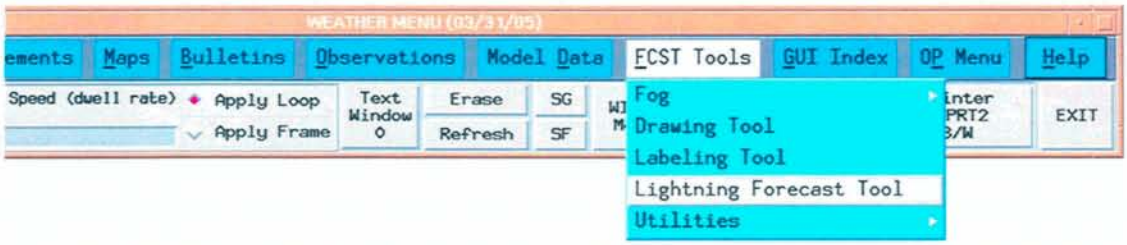

Figure 18. The MIDDS Toolbar showing the 'FCST Tools' button drop-down menu with 'Lightning Forecast Tool' highlighted.

Figure 19. The error dialog box displayed when a 1000 UTC XMR sounding for the current date is not available. Clicking the 'OK' button closes the box.

6.2.2 Equation Predictor Dialog Box

Whether or not the 1000 UTC XMR sounding for the current date is available, the equation predictor dialog box is displayed (Figure 20). This will allow the forecasters to use the GUI to create their seven-day forecasts even if data for the current day are not available. The dialog box has five tabs, one for each month. The tab of the current month is displayed initially. The current month, day and sounding time are printed along the top of the dialog box. If the current day's sounding is not available, 'No Current Sounding' will be displayed in place of the date and time in the upper right. The day value can be changed by the up/down arrows or by entering a value manually in the text box. This allows forecasters flexibility when making the seven-day Weekly Planning Forecast. The sounding date and time is formatted by year, day of year, and UTC time. The rest of the dialog box mirrors that of the Excel GUI (Figures 12 – 16).

Forecasters begin by choosing Yes or No for persistence, then a flow regime. They do not have to enter the sounding parameters as those values are already input by the GUI code and are displayed in their associated text boxes. The sounding parameters are retrieved by two MIDDS routines: PTLIST and SNDSKEWTJ. The SNDSKEWTJ routine retrieves most of the sounding stability parameters that are predictors in the equations using the local sounding data. The PTLIST routine retrieves the 850 and 500 mb temperatures for the VT calculation and all of the RH values in the 825–525 mb layer to calculate MRH. If there is not a current sounding, the text boxes will be populated with the values from the most recent sounding available. The 'No Current Sounding' message in the top right corner will inform the forecaster that this is the case. If the routines can not find a sounding file of any kind, the text boxes will be populated with the extreme low value in the range of available values for each sounding parameter.

The final step is to click on the 'Calculate Probability' button in the lower right corner of the dialog box. The 'Dismiss' button in the lower left closes the GUI. If the forecaster does not choose a persistence value or flow regime, one of two error messages is displayed informing the forecaster that a choice needs to be made. There is one error message for persistence and one for flow regime (Figure 21).

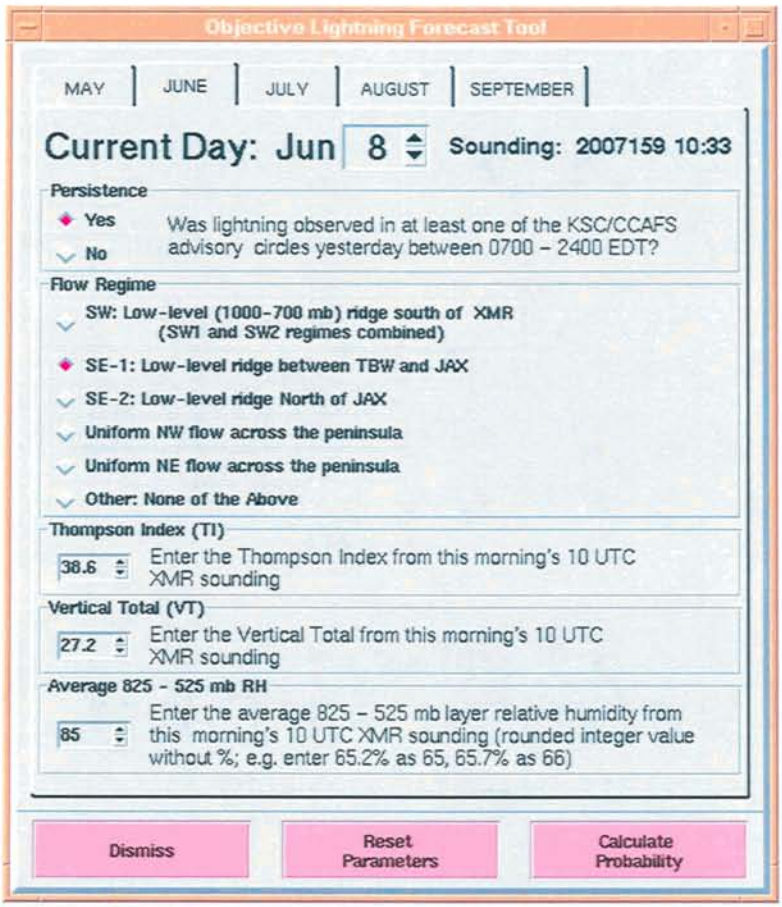

Figure 20. Equation predictor dialog box for June in MIDDS. A tab for each month is at the top, followed by the date and sounding time, then the predictor values. Clicking the 'Dismiss' button closes the GUI, the 'Reset Parameters' button resets the sounding stability parameters to original values, and the 'Calculate Probability' button displays the probability output dialog box (Figure 22).

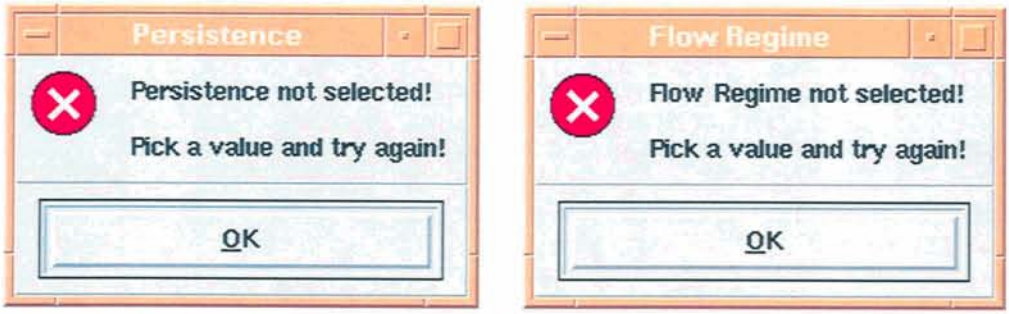

Figure 21. The error dialog box displayed when persistence is not chosen (left) or a flow regime is not chosen (right). Clicking the 'OK' button closes the box.

6.2.3 Output

When the user clicks the 'Calculate Probability' button in the equation predictor dialog box, the probability of lightning occurrence for the day is displayed in a dialog box similar to that of the Excel GUI (Figure 17). The MIDDS output dialog box is shown in Figure 22. The GUI code also outputs a file that contains all the parameter values input by the user to calculate the probability. This file is currently named LtgProb.txt, and resides in the MIDDS data directory.

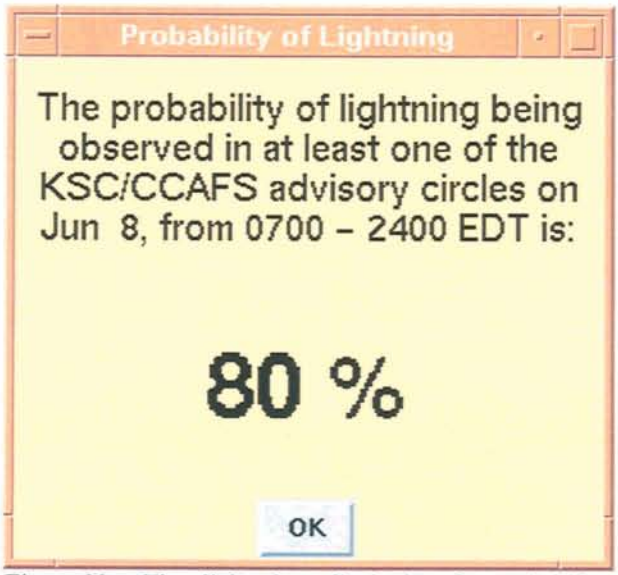

Figure 22. The dialog box displaying the probability of lightning occurrence for the day as calculated by the equation. Clicking the 'OK' button closes the box.

6.3 Predictor Responses

As done by Mr. Roeder of the 45 WS in Phase I, the AMU 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 five months.

6.3.1 May

The response charts for May 15 are shown in Figure 23. The probability response curves due to changes in the predictors VT and KI are given in Figure 23a. The flow regime and persistence values were held constant at SW and Yes, respectively. As VT was varied from 10 to 50, KI was held constant at its May mean value of 17. Conversely, as KI was varied from -30 to 70, VT was held at its May mean value of 25. 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 VT than in KI. The probability changed from 10 to 80% over a change in VT from 20 to 35. It took a much larger change in KI, from 0 to 55, to effect the same change in probability. The median value for KI, the most important predictor in the equation (Table 7), is below 20. This is the threshold value above which thunderstorm formation becomes more probable. Even when KI = 20 in Figure 23a, the probability is still only 28%. Although the mean value for VT is conducive for thunderstorm formation, this predictor contributed less to the reduction in residual deviance than KI and has only a moderate effect on the probability outcome. In Figure 23a, VT = 25 yielded a 25% probability of lightning occurrence.

The bar chart in Figure 23b shows the results of varying flow regime and persistence with VT and KI held constant at their May mean values. The SW flow regime produced the highest probability, and the probabilities were higher for every flow regime when persistence = Yes. The probability values were quite low for all flow regimes and both persistence categories, ranging from 3% (SE-2 and NE, No) to 25% (SW, Yes). This is likely an artifact of the lightning climatology for May. Lightning occurred on only 91 of the 518 available days in May, yielding an 18% monthly climatology of lightning occurrence (Lambert 2006).

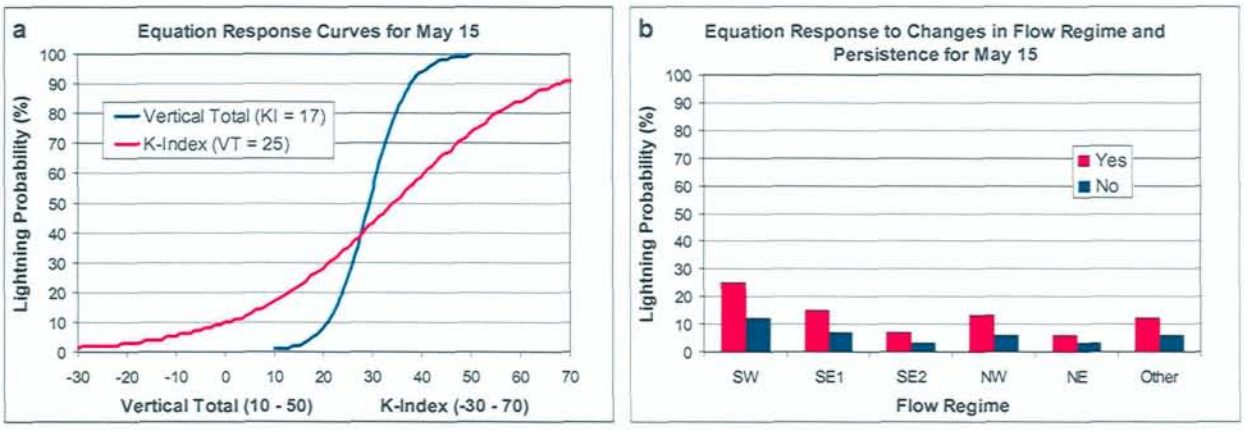

Figure 23. Equation response charts for May 15: (a) change in probability due to changes in VT and KI with flow regime = SW, persistence = Yes, KI = 17 when VT was varied from 10 to 50 (blue), and VT = 25 when KI was varied from -30 to 70 (red); (b) change in probability due to changes in flow regime and persistence with KI = 17 and VT = 25. The red bars represent persistence = Yes and the blue bars represent persistence = No.

6.3.2 June

The response charts for June 15 are shown in Figure 24. The probability response curves due to changes in the predictors TI, VT, and MRH are given in Figure 24a. The flow regime and persistence values were held constant at SW and Yes, respectively. As TI was varied from -20 to 60, VT and MRH were held constant at their June mean values of 25 and 59%, respectively. As VT was varied from 0 to 45, TI and MRH were held constant at their June mean values of 30 and 59%, respectively. Finally, as MRH was varied from 0 to 100, TI and MRH were held constant at their June mean values of 30 and 59%, respectively. The TI and VT curves are non-linear and are similar in shape to the logistic regression curve in Figure 6. The MRH curve is not shaped like that in Figure 6, and is almost linear. The probabilities appear most sensitive to changes in VT with a change in probability from 10–90% over a VT range of 10 to 30. The same change in probability occurred with a change in TI of -5 to 45. The probabilities only changed by 45%, from 40 – 85%, over the entire range of MRH values.

The bar chart in Figure 24b shows the case of varying flow regime and persistence with TI, VT, and MRH held constant at their June mean values. The flow regime ranked second in the equation, which indicates a strong effect on the calculated probability. The range of probability values, from 22 to 71% for persistence = Yes and 15 to 60% for persistence = No verifies this strong effect. The SW flow regime produced the highest probabilities and the NE flow regime had the lowest probabilities for both persistence categories. The probabilities were higher for every flow regime when persistence = Yes. Overall, the probabilities were much higher than the corresponding values for May. Unlike the low occurrence of lightning in May, the monthly climatology for June was 46% (Lambert 2006).

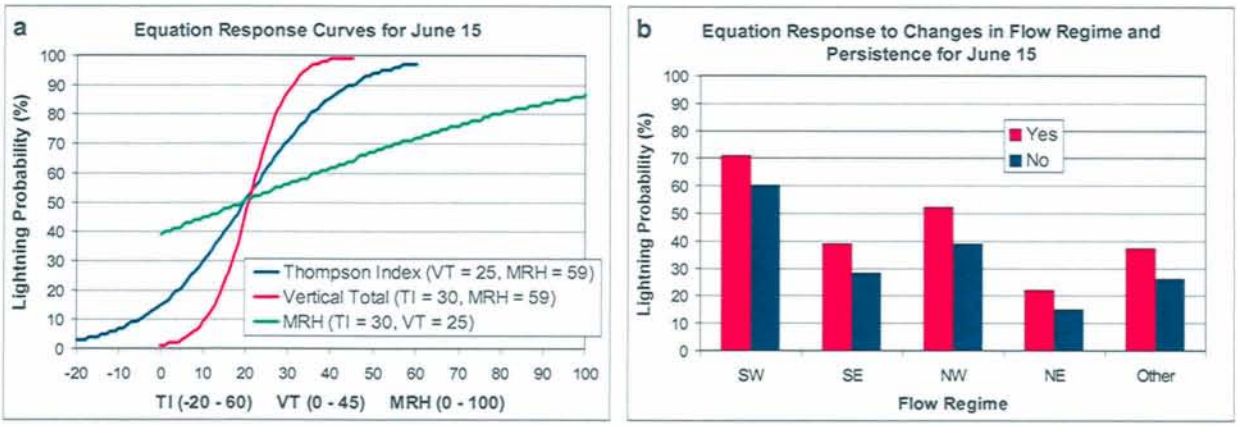

Figure 24. Equation response charts for June 15: (a) change in probability due to changes in TI, VT, and MRH with flow regime = SW, persistence = Yes, VT = 25 and MRH = 59% when TI was varied from -20 to 60 (blue), TI = 30 and MRH = 59% when VT was varied from 0 to 45 (red), and TI = 30 and VT = 25 when MRH was varied from 0 to 100% (green); (b) changes in probability due to changes in flow regime and persistence with TI = 30, VT = 25, and MRH = 59%. The red bars represent persistence = Yes and the blue bars represent persistence = No.

6.3.3 July

The response charts for July 15 are shown in Figure 25. The probability response curves due to changes in the predictors TI and TT are given in Figure 25a. The flow regime and persistence values were held constant at SW and Yes, respectively. As TI was varied from -20 to 70, TT was held constant at its July mean value of 44. Conversely, as TT was varied from 0 to 75, TI was held at its mean value of 31. The probabilities were more sensitive to changes in TT than TI. The probability changed from 20 to 80% over the TT range of 30 to 50, while a larger TI range of -10 to 45 was required to effect the same change. It ranked third among the predictors for July whereas TT ranked first in its reduction of the residual deviance. However, TI ranked first in its reduction of residual deviance in the equation, and TT ranked third. It is, therefore, likely that TI will have more influence on the probability than TT.

The bar chart in Figure 25b shows the case of varying flow regime and persistence with TI and TT held constant at their mean July values. The SW flow regime produced the highest probabilities, and the probabilities were higher for every flow regime when persistence = Yes. The probability values covered a large range for both persistence categories. The Yes values ranged from 22 to 67% and the No values ranged from 13 to 51%. The highest values in both categories were associated with the SW flow regime and the lowest values with the NE flow regime. Flow regime ranked second in the equation and persistence ranked fourth. Therefore, the flow regime will likely have a larger effect on the probability than persistence.

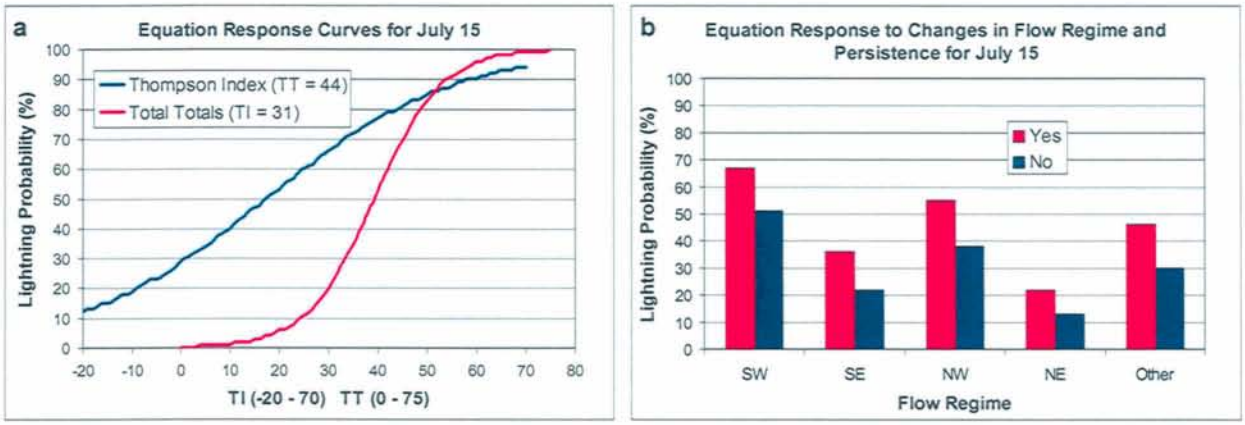

Figure 25. Equation response charts for July 15: (a) change in probability due to changes in the values of TT and TI with flow regime = SW, persistence = Yes, TT = 44 when TI was varied from -20 to 70 (blue), and TI = 31 when TT was varied from 0 to 75 (red); (b) changes in probability due to changes in flow regime and persistence with TT = 44 and TI = 31. The red bars represent persistence = Yes and the blue bars represent persistence = No.

6.3.4 August

The response charts for August 15 are shown in Figure 26. The probability response curves due to changes in the predictors TI, MRH, and VT are given in Figure 26a. The flow regime was held constant at SW. Persistence was not a predictor in the August equation. As TI was varied from -20 to 70, MRH and VT were held constant at their August mean values of 59% and 24, respectively. As MRH was varied from 0 to 100%, TI and VT were held at their August mean values of 31 and 24, respectively. Finally, as VT was varied from 0 to 50, TI and MRH were held at their August mean values of 31 and 59%, respectively. The MRH curve exhibits the same degree of linearity as that for June, and the slope indicates that changes in MRH would effect a small change on the resulting probability. The TI and VT curves exhibit the classic logistic regression shape shown in Figure 6. The probability values change more quickly in response to changes in VT than TI. The probability changed from 10 to 90% over a TI range of -10 to 55 and a VT range of 10 to 30. However, TI explained most of the residual deviance in the equation development while VT explained the least. It is likely that the resulting probability will be more influenced by the value for TI than VT.

The bar chart in Figure 26b shows the case of varying flow regime with TI, MRH, and VT held constant at their mean August values. The SW flow regime produced the highest probability and the NE regime produced the lowest. The August flow regime lightning probability is 12% for NE flow and 68% for SW flow. Since the flow regime ranked second in the equation, it had a large effect on the resulting probability values. This influence on the probability out come is shown in Figure 26b where the probability values exhibit a large range from 18% for NE flow to 64% for SW flow.

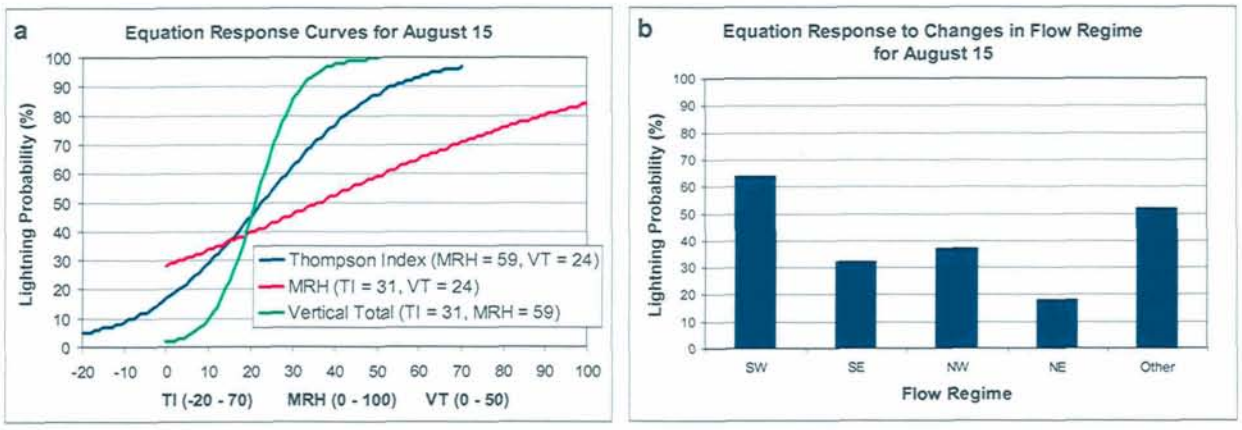

Figure 26. Equation response charts for August 15: (a) change in probability due to changes in the values of TI, MRH, and VT with flow regime = SW, persistence = Yes, MRH = 59% and VT = 34 when TI was varied from -20 to 70 (blue), TI = 31 and VT = 24 when MRH was varied from 0 to 100% (red), and TI = 31 and MRH = 59% when VT was varied from 0 to 50 (green); (b) changes in probability due to changes in flow regime with TI = 31, MRH = 59%, and VT = 24.

6.3.5 September

The response charts for September 15 are shown in Figure 27. The response curves due to changes in the predictors MRH and VT are given in Figure 27a. The flow regime and persistence values were held constant at SW and Yes, respectively. As MRH was varied from 0 to 100%, VT was held constant at its September mean value of 24. As VT was varied from 0 to 45, MRH was held at its September mean value of 57%. Again, the MRH curve exhibits the same degree of linearity as that for June and August, and the slope indicates that changes in MRH would effect a small change on the resulting probability. The VT curve is similar to that in Figure 6, and the probability values change more quickly in response to changes in VT than MRH in Figure 27a. The probability range for the range of VT values is 0 to 100%, but the probability range for all the MRH values was 15 to 85%. However, MRH explained most of the residual deviance in the equation development while VT explained only a small amount. It is likely that the resulting probability will be more influenced by the value for MRH than VT.

The bar chart in Figure 27b shows the case of varying flow regime and persistence with MRH and VT held constant at their September mean values. Flow regime ranked second behind MRH in its reduction of residual deviance and had a large effect on the calculated probability, evident in Figure 27b. Persistence ranked third behind flow regime, and its influence was also evident in the difference in probability between Yes and No for each flow regime. The SW regime had highest values for both Yes and No persistence at 56 and 37%, respectively. The NW and NE regimes have similar values at 25 and 13% for Yes and No persistence, respectively.

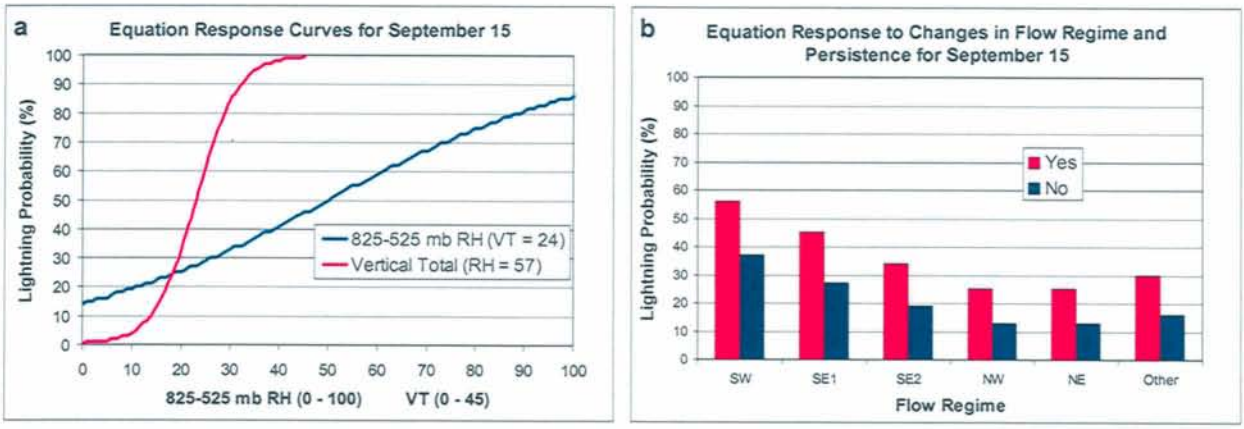

Figure 27. Equation response charts for September 15: (a) change in probability due to changes in the values of MRH and VT with flow regime = SW, persistence = Yes, VT = 24 when MRH was varied from 0 to 100% (blue), and MRH = 57% when VT was varied from 0 to 45 (red); (b) changes in probability due to changes in flow regime and persistence with MRH = 57% and VT = 24. The red bars represent persistence = Yes and the blue bars represent persistence = No.

7. Summary and Conclusions

The AMU created five logistic regression equations that predict the probability of cloud-to-ground lightning occurrence for the day in the KSC/CCAFS 5 n mi warning circles for each month in the warm season. These equations are based on equations developed in Phase 1 (Lambert and Wheeler 2005), but with five modifications:

- Increase the POR from 15 to 17 years,
- Modify the valid area to eliminate areas that are not within the 5 n mi warning circles,
- Include the XMR 1000 UTC sounding in determining the flow regime of the day,
- Use a different smoothing function for the daily lightning climatology, and
- Determine the optimal layer for the average RH calculation.

The P-2 equations described in this report outperformed the P-1 equations by an overall 8%, and showed better performance than the P-1 equations in four other tests. As a result, the new P-2 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.

Results from the P-2 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.

7.1 Equation Performance Review

The P-2 equations were tested using five methods described in Sections 5.3.1–5.3.5:

- The Brier SS,
- Probability distributions for lighting and non-lightning days,
- Reliability and bias,
- MSE decomposition, and
- Contingency table statistics.

The results from each of these tests showed marginal to superior performance of the P-2 equations over the P-1 equations with 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 and the P-1 equations. Given that most of the test results showed that the P-2 equations exhibited superior performance over the P-1 equations, they replaced the operational P-1 equations at the beginning of the 2007 lightning season.

7.2 GUI Issues

The Excel and MIDDS GUIs described in Section 6 interface with the equations and facilitate user-friendly input and fast output of the lightning probability for the day. The MIDDS GUI accesses and calculates all of the sounding parameter values needed for each equation and displays them in their text boxes when a 1000 UTC sounding is available. If a sounding is not available for the current day, the GUI will display values from the most recent 1000 UTC sounding and 'No Current Sounding' in the upper right corner. The values must be entered manually into the Excel GUI.

7.2.1 Calculated Values for Excel GUI

Most of the values for manual input to the Excel GUI are available to the user through the MIDDS Skew-T program, but VT, TI, and MRH must be calculated. The equation for VT is

$$VT = T_{850} - T_{500},$$

where T_{850} is the temperature at 850 mb and T_{500} is the temperature at 500 mb; and the equation for TI is

$$TI = KI - LI,$$

where KI and LI are output by the MIDDS Skew-T program. The MRH value should be calculated using Equations 5 and 6 with all sounding RH observations between 825 and 525 mb, inclusive.

7.2.2 Flow Regime Determination

Forecasters must determine and choose the flow regime manually for both GUIs. The very first step forecasters should take before determining the flow regime is to refer to Lambert (2006) and Section 3.3 of this report to understand how a flow regime was determined in this work. The flow regime for each day was determined by first using the 1200 UTC soundings at MFL, TBW, and JAX. The 1000 UTC XMR sounding was used only when one or more of the 1200 UTC soundings was missing or the flow regime determined by the 1200 UTC soundings was Other. Since 1200 UTC soundings were the main source 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 estimate the flow regime for the day:

- 1000 UTC XMR sounding;
- Pressure and wind field output from the most recent initializations of the
  - Rapid Update Cycle (RUC),
  - North American Mesoscale (NAM,), 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 flow in the XMR 1000–700 mb layer should be combined with another source, such as one of the model initializations. 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 a discussion of other factors influencing the formation of convection for the day.

7.2.3 Local vs NOAAPort XMR Sounding Data

The AMU began testing the new lightning probability forecast equations on 1 May 2007 in an effort to archive the output and verify equation performance for the warm season of 2007. For input to the equations, the AMU used the 1000 UTC XMR sounding parameters found in MetWise Net (http://extremeforecasting.com/net/overview.htm). On 24 May, the AMU calculated a probability of 2% and the 45 WS calculated a probability of 4%. While this difference was small, it was cause for concern that either the 45 WS version of the GUI had a problem, or the data entered into the GUI by the AMU or 45 WS were incorrect. The initial investigation revealed a discrepancy in the KI value between MIDDS and MetWise Net. The AMU checked the KI value on the Advanced Weather Interactive Processing System (AWIPS) in the AMU and found that it was the same as that in Metwise Net. A check of the sounding available in the archive on the GSD website confirmed that the values in this sounding created a KI equal to that in Metwise Net and AWIPS. MetWise Net, AWIPS, and GSD receive their sounding data through the NOAAPort Receive System (NRS).

Mr. Wahner of CSR checked the mandatory and significant level data in MIDDS that came directly from Weather Station A at CCAFS and those that were re-transmitted to MIDDS through the NRS. He noted large discrepancies in the dewpoint temperatures at 700 and 500 mb. After some investigation by Mr. Wahner, Mr. Herring, and CSR programmers at Weather Station A, they found a line of code in the software that prepares the data for transmission through the NRS. The logic of the code sets the dewpoint depression (DD) equal to 30° C if the RH is 20% or less. This was a requirement found in an older version of the Federal Meteorological Handbook (FMH) #3, but is not in the current FMH #3. Mr. Herring of CSR initiated a requirements statement to remove this algorithm from the code. Details of this investigation can be found in Bauman (2007).

The routines in the MIDDS GUI use the local data directly from Weather Station A. Forecasters must use caution to make sure they are using these data and not data from a sounding on AWIPS or one that was transmitted to MIDDS through the NRS.

7.3 Future Work

At the most recent AMU Tasking Meeting in April 2007, a third phase to this task was approved in which two changes will be made in an effort to further improve equation performance. The first will be to include October data to the current POR. In looking at Figure 5a, it appears that the end of the lightning season is beyond September 30. The daily climatology values at the end of September are approximately 10% higher than at the beginning of May. The second modification will be to stratify the warm seasons by the progression of the daily climatology instead of by month. Stratifying the data by the progression of the daily climatology through the warm season is more natural than by date. The AMU will develop a method to determine four to five sub-seasons in the overall warm season:

- A pre-lightning season in early May,
- A spin-up transition season from mid-May to early or mid June,
- The core lightning season from June to mid or late August,
- A spin-down transition season through September, and
- A possible post-lightning season in October.

The number of sub-seasons can not be determined until the October data are added and a new daily climatology chart is created. One equation will be developed for each sub-season and their performance compared to the P-2 equations.

Data from more years will be added in future phases. This will help develop more robust statistical relationships in the equations and provide more data for verification. Also, new techniques may be available over the next few years that could also help improve equation performance. Any such techniques should be considered and tested in future phases. 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

Bauman, W., 2007: Dewpoint Temperature Discrepancy in MIDDS vs. NOAAPort XMR Soundings. NASA Applied Meteorology Unit Memorandum, 2 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL 32931, or Dr. Bauman at bauman.bill@ensco.com].

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].

Insightful Corporation, 2005a: *S-PLUS 7 for Windows User's Guide*, Insightful Corp., Seattle, WA, 664 pp.

Insightful Corporation, 2005b: *S-PLUS 7 for Windows Guide to Statistics, Volume 1*, Insightful Corp., Seattle, WA, 730 pp.

Lambert, W., 2006: New Lightning Probabilities Based on Flow Regime. NASA Applied Meteorology Unit Memorandum, 7 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL 32931, or Ms. Lambert at lambert.winnie@ensco.com].

Lambert, W. and M. Wheeler, 2005: Objective lightning probability forecasting for Kennedy Space Center and Cape Canaveral Air Force Station. NASA Contractor Report CR-2005-212564, Kennedy Space Center, FL, 54 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 230, Cocoa Beach, FL, 32931, and at http://science.ksc.nasa.gov/amu/final.html]

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

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

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

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2d ed. Academic Press, Inc., San Diego, CA, 467 pp.

List of Acronyms

45 WS	45th Weather Squadron
AMU	Applied Meteorology Unit
AWIPS	Advanced Weather Interactive Processing System
CAPE	Convective Available Potential Energy
CCAFS	Cape Canaveral Air Force Station
CG	Cloud-to-Ground
CGLSS	CG Lightning Surveillance System
CIN	Convective INhibition
CSI	Critical Success Index
CSR	Computer Sciences Raytheon
CT	Cross Totals
FAR	False Alarm Ratio
GFS	Global Forecast System model
GSD	Global Systems Division
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	Kuyper Skill Score
LFC	Level of Free Convection
LI	Lifted Index
McIDAS	Man-computer Interactive Data Access System
MFL	Miami, FL 3-letter identifier
MIDDS	Meteorological Interactive Data Display System
MRH	Mean RH in the 825-525 mb layer
MSE	Mean Square Error
NAM	North American Mesoscale model
NE	Northeast flow regime
NLDN	National Lightning Detection Network
NPTI	Neumann-Pfeffer Thunderstorm Index
NW	Northwest flow regime
P-1	Phase I equations
P-2	Phase II equations
POD	Probability of Detection
POR	Period of Record
RH	Relative Humidity
RUC	Rapid Update Cycle model
SE	Southeast flow regime
SMG	Spaceflight Meteorology Group
SS	Skill Score
SW	Southwest flow regime
TBW	Tampa, FL 3-letter identifier
Tcl/Tk	Tool Command Language/Toolkit
TI	Thompson Index
TT	Total Totals
VT	Vertical Totals
WMO	World Meteorological Organization
XMR	CCAFS rawinsonde 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 for the purpose of fully informing the reader of the resources used to conduct the work reported herein.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
31-07-2007	Final	June 2006 - July 2007

4. TITLE AND SUBTITLE

Objective Lughtning Probability Forecasting for Kennedy Space Center and Cape Canaveral Air Force Station, Phase II

5a. CONTRACT NUMBER

NNK06MA70C

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Winifred Lambert

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

ENSCO, Inc.
1980 N. Atlantic Ave. Suite 230
Cocoa Beach, FL 32931

8. PERFORMING ORGANIZATION REPORT NUMBER

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

NASA
John F. Kennedy Space Center
Code KT-C-H
Kennedy Space Center, FL 32899

10. SPONSORING/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING REPORT NUMBER

NASA/CR-2007-214732

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified, Unlimited

13. SUPPLEMENTARY NOTES

An electronic version can be found at http://science.ksc.nasa.gov/amu/final.html

14. ABSTRACT

This report describes the work done by the Applied Meteorology Unit (AMU) to update the lightning probability forecast equations developed in Phase I. In the time since the Phase I equations were developed, new ideas regarding certain predictors were formulated and a desire to make the tool more automated was expressed by 45 WS forecasters. Five modifications were made to the data: 1) increased the period of record from 15 to 17 years, 2) modified the valid area to match the lighting warning areas, 3) added the 1000 UTC CCAFS sounding to the other soundings in determining the flow regime, 4) used a different smoothing function for the daily climatology, and 5) determined the optimal relative humidity (RH) layer to use as a predictor. The new equations outperformed the Phase I equations in several tests, and improved the skill of the forecast over the Phase I equations by 8%. A graphical user interface (GUI) was created in the Meteorological Interactive Data Display System (MIDDS) that gathers the predictor values for the equations automatically. The GUI was transitioned to operations in May 2007 for the 2007 warm season.

15. SUBJECT TERMS

weather, thunderstorm, lightning, convection, statistical forecasting

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Francis J. Merceret
U	U	U	UU	59	19b. TELEPHONE NUMBER *(Include area code)* (321) 867-0818