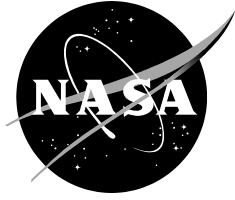


NASA/TM-2012-216314



Using Global Positioning System Integrated Precipitable Water Vapor to Forecast Lightning on KSC/CCAFS

*Lisa L. Huddleston
Applied Meteorology Unit
Kennedy Space Center, Florida*

December 2012

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 counter-part 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.

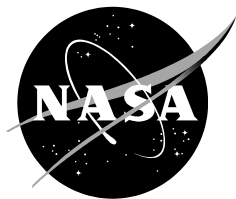
- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with 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 organizing and publishing research results, distributing specialized research announcements and feeds, providing help desk and personal search support, and enabling data exchange services.

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 to help@sti.nasa.gov
- Fax your question to the NASA STI Help Desk at 443-757-5803
- Phone the NASA STI Help Desk at 443-757-5802
- Write to:
STI Information Desk
NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320

NASA/TM-2012-216314



Using Global Positioning System Integrated Precipitable Water Vapor to Forecast Lightning on KSC/CCAFS

*Lisa L. Huddleston
Applied Meteorology Unit
Kennedy Space Center, Florida*

National Aeronautics and
Space Administration

*Kennedy Space Center
Kennedy Space Center, FL 32899-0001*

December 2012

Acknowledgements

The author would like to thank Dr. William Bauman and Ms. Winifred Crawford of the Applied Meteorology Unit, Dr. Francis Merceret of the Kennedy Space Center Weather Office, and Mr. Bill Roeder of the 45th Weather Squadron for lending their time, data, and statistical expertise to this task. The author would also like to thank Mr. Seth Gutman of the National Oceanic and Atmospheric Administration Earth System Research Laboratory for helpful information about the Ground-Based GPS Meteorology web site and data.

Available from:

NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320
443-757-5802

This report is also available in electronic form at

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

Executive Summary

Forecasters from the 45th Weather Squadron (45 WS) include a probability of lightning occurrence in their daily 24-hour and weekly planning forecasts. Major improvements have been made in forecasting the probability of lightning for the entire day including the Applied Meteorology Unit developed Objective Lightning Probability tool, which is used routinely during the warm season to forecast the probability of lightning occurrence. This tool outperformed the previous lightning probability techniques by 56% (Lambert 2007). However, the timing of the lightning remains a challenge. Timing is important in many operational decisions by customers for daily ground operation activities on Kennedy Space Center and Cape Canaveral Air Force Station. Examples include deciding to start or stop part of a launch operation like tower rollback, major ground processing operation, and routine ground processing. To help improve this forecast, the 45 WS and others have investigated techniques using Global Positioning System Integrated Precipitable Water (GPS-IPW) observations and changes over specified time periods to improve the skill in forecasting a first strike (Mazany et al. 2002; Inoue and Inoue 2007; Kehrer et al. 2008; Suparta et al. 2011a; and Suparta et al. 2011b). The previous work suggested that GPS-IPW could be useful to 45 WS operations. The purpose of this task was to determine if GPS-IPW would indeed be useful to 45 WS in predicting the probability of lightning at shorter time periods than the Objective Lightning Probability tool.

The data sources used for this task included 45th Space Wing Cloud-to-Ground component of the Four Dimensional Lightning Surveillance System (aka the Cloud-to-Ground Lightning Surveillance System), the GPS-IPW data from the GPS sensor near the Cape Canaveral, Fla. Lighthouse, and the lightning probabilities from the Objective Lightning Probability tool. Since data from the GPS-IPW site were not available before 2000, the period of record was 2000-2011 for the warm-season months of May-October. Because many of the variables considered in previous studies (Mazany et al. 2000 and Kehrer et al. 2008) such as K-Index, Total Totals, and upper level moisture variables, were already considered by the Objective Lightning Probability tool, only the objective lightning probability values along with the current GPS-IPW values and changes in GPS-IPW over half hour increments up to 24 hours were considered in this study. New models were built using binary (yes/no) logistic regression where the element to be forecast was the occurrence of lightning. Yes meant lightning occurred and no meant lightning did not occur within the specified time period and area of interest.

Although previous studies have shown GPS-IPW values to be promising in forecasting lightning, the results of this study did not find them to be very useful. This is likely because the level of noise in the Objective Lightning Probability (which dominates the regression equations) is greater than the increase in predictive capability offered by the inclusion of the GPS-IPW data.

Table of Contents

Executive Summary	2
List of Figures	4
List of Tables.....	7
1. Introduction.....	8
1.1 Previous Studies.....	8
1.2 Current Work	9
2. Data.....	10
2.1 CGLSS.....	12
2.2 GPS-IPW	13
2.3 Objective Lightning Probability Tool Output.....	15
3. Equation Development	16
3.1 Predictand.....	16
3.2 Candidate Predictors	16
3.3 Data availability.....	16
3.4 Development and Verification Data Sets	17
3.5 Logistic Regression	18
3.6 Predictor Selection Methodologies	18
4. Logistic Regression Equation Verification	21
4.1 Relative Operating Characteristics (ROC).....	21
4.2 Reliability Diagrams	22
4.3 Equation Performance	23
4.4 Predictor Collinearity and Scree Plots	35
5. Equation Performance.....	41
6. Summary and Conclusions	44
References.....	45
List of Acronyms	46

List of Figures

- Figure 1. Time series for June 2003 of GPS-IPW in centimeters (cm) (left vertical axis) from GPS observations made at the Cape Canaveral, Fla. Coast Guard GPS site (blue line) and IPW from rawinsonde observations from CCAFS (red pluses) compared with (right vertical axis) the Objective Lightning Probability tool daily probability and whether or not lightning occurred (1 = yes, 0 = no) within the 5 NM KSC/CCAFS lightning warning circles shown in Figure 4. 10
- Figure 2. Time series for June 2003 of GPS-IPW in centimeters (cm) (left vertical axis) from GPS observations made at the Cape Canaveral, Fla. Coast Guard GPS site (blue line) and IPW from rawinsonde observations from CCAFS (red pluses) compared with (right vertical axis) the CGLSS lightning flash count of lightning that occurred within the 5 NM KSC/CCAFS lightning warning circles shown in Figure 4. 11
- Figure 3. Scatter plot of Objective Lightning Probability tool daily probability vs. Δ IPW over a 24-hour period and a 2-hour lead time before a CGLSS flash occurred (blue diamond, lightning occurred) and when a CGLSS flash did not occur (red square, lightning did not occur) for the month of June during the entire POR. 12
- Figure 4. The 5-NM lightning warning circles on KSC (blue) and CCAFS (red). 13
- Figure 5. (a) Overview map of Cape Canaveral Air Force Station (CCAFS) and Kennedy Space Center (KSC) vicinity. (b) Inset map of CCAFS showing the location of the GPS receiver in relation to the skid strip and Cape Canaveral lighthouse. 14
- Figure 6. Flowchart depicting the forward selection method of choosing model predictors (Anderson et al. 2012). 19
- Figure 7. Flowchart depicting the backward elimination method of choosing model predictors (Anderson et al. 2012). 20
- Figure 8. A basic ROC diagram showing the false positive rate (or false alarm rate) on the x-axis and the true positive rate (or hit rate) on the y-axis. The 45° line is the chance line, representing a forecast system that has no skill, and the dashed curve bending towards the upper left corner represents a skillful forecast system. 22
- Figure 9. Example reliability diagram for an overconfident forecast. The dashed line represents a perfectly reliable forecast. Here the event occurs more frequently than indicated when the forecast indicates a decreased probability of the event occurring (to the left of the dotted line), but less frequently than indicated when the forecast indicates an increased probability of the event occurring (to the right of the dotted line). Although the forecasts correctly indicate increases and decreases in the probabilities of the events, the changes in probability are over-stated, and the forecasts are said to be over-confident. 23
- Figure 10. ROC curve for each variable in the 2-hour forecast regression equation using the forward selection method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability and the current GPS-IPW are the best predictors. The rest of the GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve. 25
- Figure 11. Reliability diagram for the 2-hour forecast equation using the forward selection method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 2-hour forecast equation. The

reliability is good with slight overforecasting in the small range available. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.....26

Figure 12. ROC curve for each variable in the 9-hour forecast regression equation using the forward selection method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability is the best predictor. GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.28

Figure 13. Reliability diagram for the 9-hour forecast equation using the forward selection method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 9-hour forecast equation. The reliability is good with slight overforecasting in the small range available. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.....29

Figure 14. ROC curve for each variable in the 2-hour forecast regression equation using the backward elimination method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability is the best predictor. GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.31

Figure 15. Reliability diagram for the 2-hour forecast equation using the backward elimination method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 2-hour forecast equation. The reliability is good in the 0 to 0.2 range, but increasing overforecasting in the 0.2 to 0.4 range. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.....32

Figure 16. ROC curve for each variable in the 9-hour forecast regression equation using the backward elimination method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability is the best predictor. GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.34

Figure 17. Reliability diagram for the 2-hour forecast equation using the backward elimination method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 2-hour forecast equation. The reliability is good with slight overforecasting in the small range available. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.....35

Figure 18. Scree plot for the predictors in the 2-hour forecast regression equation using the forward selection method of predictor selection. The first four variables have eigenvalues > 1 and are likely useful in a forecast equation.36

Figure 19. Scree plot for the predictors in the 9-hour forecast regression equation using the forward selection method of predictor selection. The first three variables have eigenvalues > 1 and are likely useful in a forecast equation.38

Figure 20. Scree plot for the predictors in the 2-hour forecast regression equation using the backward elimination method of predictor selection. The first two variables have eigenvalues > 1 and are likely useful in a forecast equation.39

Figure 21. Scree plot for the predictors in the 9-hour forecast regression equation using the backward elimination method of predictor selection. The first two variables have eigenvalues > 1 and are likely useful in a forecast equation.40

List of Tables

Table 1.	GPS-IPW records available initially for the 12-year POR and the final amount after removing missing data.	17
Table 2.	GPS-IPW records counts in the development data set and validation data set.	17
Table 3.	Area under the ROC curve and forecast skill index, S, for each predictor in the 2-hour forecast regression equation using the forward selection method of predictor selection, where b_0 is -7.157.....	24
Table 4.	Area under the ROC curve and forecast skill index, S, for each predictor in the 9-hour forecast regression equation using the forward selection method of predictor selection, where b_0 is -5.093.....	27
Table 5.	Area under the ROC curve and forecast skill index, S, for each variable in the 2-hour forecast regression equation using the backward elimination method of variable selection, where b_0 is -5.359.....	30
Table 6.	Area under the ROC curve and forecast skill index, S, for each variable in the 9-hour forecast regression equation using the backward elimination method of variable selection, where b_0 is -5.100.....	33
Table 7.	Comparison of accuracy measurements, skill scores, and bias for the 2-hour forecast equations, using the backward elimination method, forward selection method, and the truncated forecast equations using the backward elimination method or forward selection method. The scores are shown for a range of lightning index threshold values.....	42
Table 8.	Comparison of accuracy measurements, skill scores, and bias for the 9-hour forecast equations, using the backward elimination method, forward selection method, and the truncated forecast equations using the backward elimination method or forward selection method. The scores are shown for a range of lightning index threshold values.....	43

1. Introduction

Forecasters from the 45th Weather Squadron (45 WS) include a probability of lightning occurrence in their daily 24-hour and weekly planning forecasts. This value is used by personnel at Kennedy Space Center (KSC) and Cape Canaveral Air Force Station (CCAFS) to help plan daily ground operations. These probabilities are also the first step in the 45 WS lightning warning process and even influence the launch forecasts. A major improvement was realized in forecasting the probability of lightning for the day through the Applied Meteorology Unit (AMU)-developed Objective Lightning Probability tool, which is used routinely during the warm season to forecast the probability of lightning occurrence. This tool outperformed the 45 WS's previous objective lightning probability technique by 56% (Lambert 2007).

However, the timing of the lightning within the period covered by the Objective Lightning Probability tool remains a challenge. Timing is important in many operational decisions by customers for daily ground operation activities on KSC and CCAFS. Examples include deciding to start or stop a launch operation like tower rollback, a major ground processing operation, and routine ground processing. To help improve this forecast, the 45 WS and others have investigated techniques using Global Positioning System Integrated Precipitable Water (GPS-IPW) observations and changes over specified time periods to improve the skill in forecasting a first strike (Mazany et al. 2002; Inoue and Inoue 2007; Kehrer et al. 2008; Suparta et al. 2011a; and Suparta et al. 2011b). The purpose of this task was to determine the utility of using GPS-IPW and output from the Objective Lightning Probability tool to predict the probability of lightning at the temporal resolution of the 45 WS lightning warnings and major ground processing operations.

1.1 Previous Studies

The AMU used two previous studies as guides in conducting this task. Both studies used data from the KSC/CCAFS area with the goal of improving the lightning forecast using GPS-IPW and other observed data types.

1.1.1 Mazany et al. (2002)

Mazany et al. (2002) used data from the 1999 summer thunderstorm season and found that four predictors, maximum electric field mill values, GPS-IPW, 9-hour change in IPW, and K-Index (KI) proved important for forecasting lightning events at KSC. Their logistic regression model developed using these predictors, hereafter called the Mazany model, was shown to decrease false alarm rates by a minimum of 13.2% and improved lead time of forecasts by the Spaceflight Meteorology Group for KSC by 10%.

1.1.2 Kehrer et al. (2008)

Kehrer et al. (2008) verified the Mazany model using an expanded data set from the 2000-2003 summer thunderstorm seasons. They found that when using the expanded data set, the Mazany model underperformed expectations by examining the accuracy measures of probability of detection (POD), hit rate (HR), false alarm rate (FAR), and Kuipers skill score (KSS). They then developed and optimized two new models for operationally significant forecast intervals. Their new models were optimized based on the Operational Utility Index (OUI), which is a locally developed performance metric used to emphasize personnel safety (D'Arcangelo 2000). Their equation for OUI is

$$[(3 \times \text{POD}) + (2 \times \text{KSS}) - (1 \times \text{FAR})]/6.$$

They first developed a 2-hour forecast model that was optimized for the 45 WS lightning advisories. The 2-hour period was used to allow for the 30 minute desired lead time of these advisories, plus an additional 1.5 hours for sensor dwell, communication, calculation, analysis

and advisory decision by the forecaster. The second model was optimized for major ground processing operations. It was a 9-hour forecast model that allowed for a 7.5 hour lead time for the operation plus the same 1.5-hour discussed in the 2-hour forecast, for sensor dwell, communication, calculation, analysis, and decision making. Four predictors were significant for the 2-hour model:

- 0.5 hour change in GPS-IPW,
- 7.5 hour change in GPS-IPW,
- Current GPS-IPW value and
- KI.

Five predictors were significant for the 9-hour major ground processing forecast model:

- Current GPS-IPW value,
- 8.5 hour change in GPS-IPW,
- 2.5 hour change in GPS-IPW,
- 12 hour change in GPS-IPW, and
- KI.

1.2 Current Work

In this task, the AMU used the Objective Lightning Probability tool output, the current GPS-IPW value, and the changes in the GPS-IPW value over the last 0.5 to 24 hours in 30 minute increments to determine the time period for the GPS-IPW change that produces the best 2-hour and 9-hour probability forecasts, similar to Mazany et al. (2002) and Kehrer et al. (2008). The output from the Objective Lightning Probability/GPS-IPW models was compared to that of the Objective Lightning Probability tool alone to determine if adding the GPS-IPW data improved the current Objective Lightning Probability tool.

2. Data

The three data types used in this task were from the 45th Space Wing Cloud-to-Ground component of the Four Dimensional Lightning Surveillance System (CG-4DLSS), the GPS-IPW data from the sensor near the Lighthouse on CCAFS, and the lightning probabilities from the Objective Lightning Probability tool. CG-4DLSS is better known by the previous name for the system, the Cloud-to-Ground Lightning Surveillance System (CGLSS), therefore, that term will be used hereafter in this report. Since data from the GPS-IPW site were not available before 2000, the period of record (POR) is 2000-2011 for the warm-season months of May-October. These data were plotted by month at half hour intervals for each month and year in the POR in order to identify any data gaps and outliers. Example plots for June 2003 are shown in Figure 1 and Figure 2.

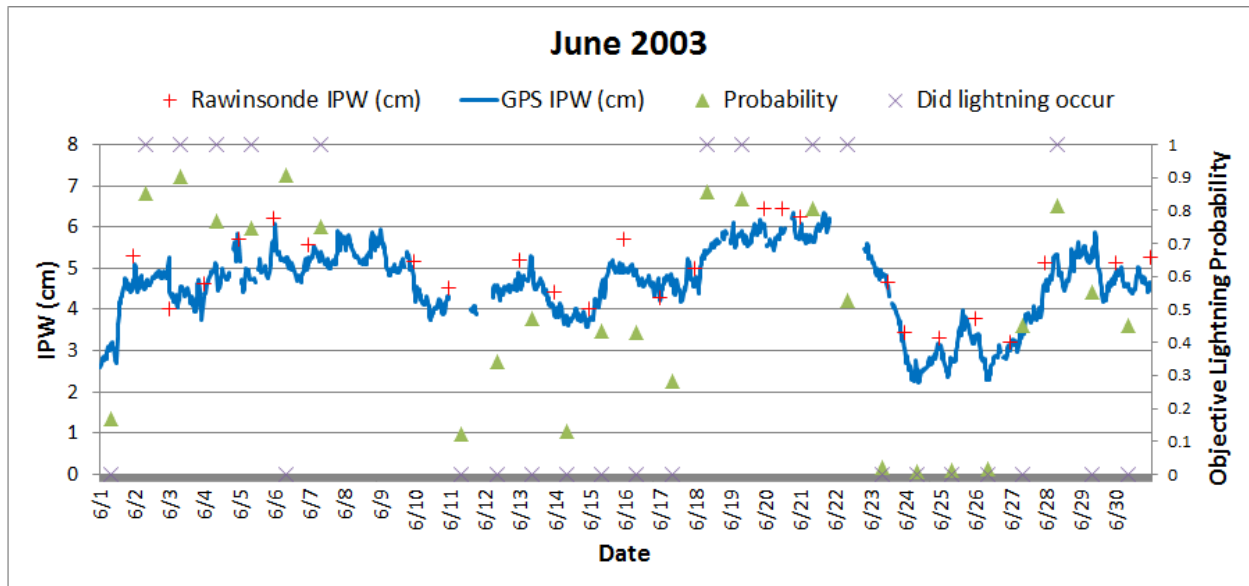


Figure 1. Time series for June 2003 of GPS-IPW in centimeters (cm) (left vertical axis) from GPS observations made at the Cape Canaveral, Fla. Coast Guard GPS site (blue line) and IPW from rawinsonde observations from CCAFS (red pluses) compared with (right vertical axis) the Objective Lightning Probability tool daily probability and whether or not lightning occurred (1 = yes, 0 = no) within the 5 NM KSC/CCAFS lightning warning circles shown in Figure 4.

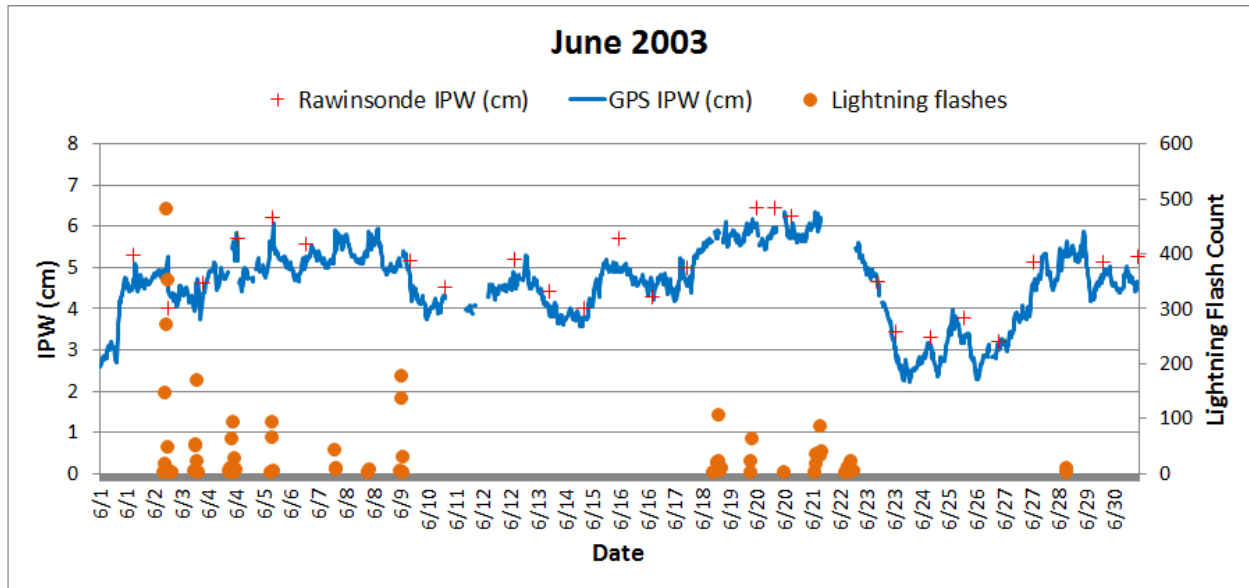


Figure 2. Time series for June 2003 of GPS-IPW in centimeters (cm) (left vertical axis) from GPS observations made at the Cape Canaveral, Fla. Coast Guard GPS site (blue line) and IPW from rawinsonde observations from CCAFS (red pluses) compared with (right vertical axis) the CGLSS lightning flash count of lightning that occurred within the 5 NM KSC/CCAFS lightning warning circles shown in Figure 4.

The 45 WS personnel requested equations with lead times of two and nine hours to support various operational requirements, as was done in Kehrner et al. (2008). The Objective Lightning Probability tool values were plotted against the 24-hour change in the GPS-IPW values for each month in all years in the POR. Figure 3 is an example of a plot of these data for June 2000-2011 and the 2-hour lead time. Plots for other months for both the 2-hour and 9-hour lead time showed similar results. The red squares indicate when lightning did not occur and the blue diamonds indicate when lightning did occur. While the density of points when lightning occurred is higher at higher values of AMU tool probabilities, lightning still occurred at lower values of AMU tool probabilities. Therefore, no IPW trends or associations with lightning occurrence were evident for any of the warm season months plotted for either the 2-hour or 9-hour lead times.

**June 2000 - 2011 GPS Δ IPW vs.
Objective Lightning Probability
2 hour lead time**

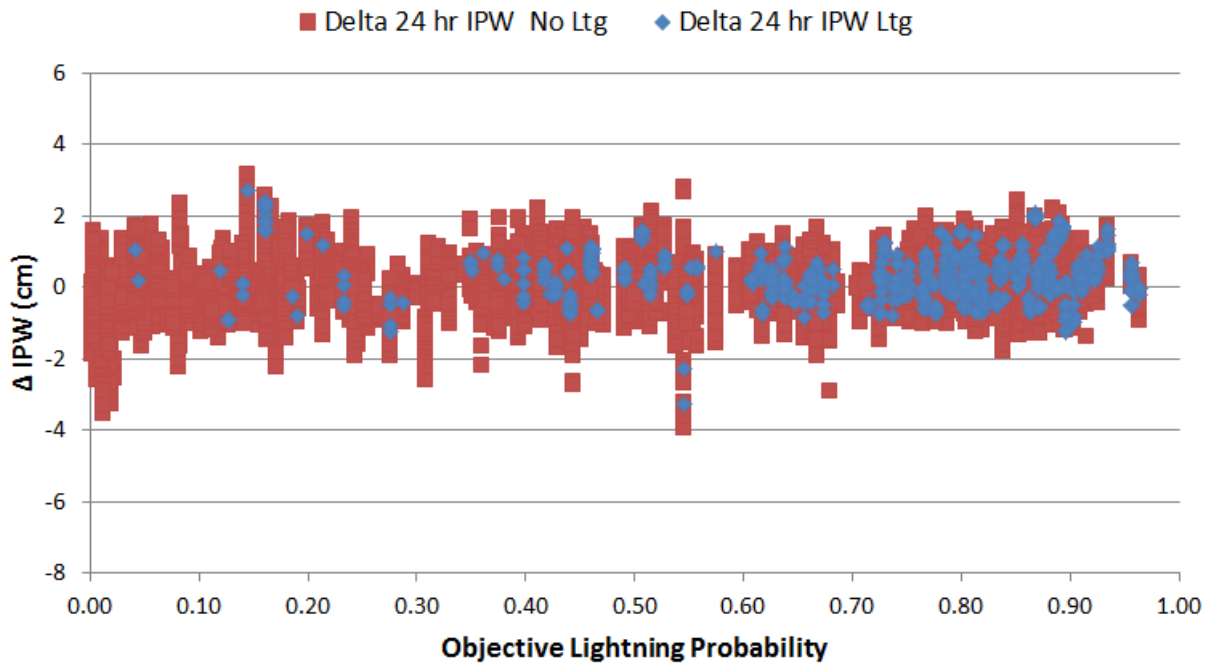


Figure 3. Scatter plot of Objective Lightning Probability tool daily probability vs. Δ IPW over a 24-hour period and a 2-hour lead time before a CGLSS flash occurred (blue diamond, lightning occurred) and when a CGLSS flash did not occur (red square, lightning did not occur) for the month of June during the entire POR.

2.1 CGLSS

CGLSS is a network of six sensors that detects cloud-to-ground strokes on KSC/CAFS and the surrounding area. It reports the date, time, latitude, longitude, peak current magnitude, and location error ellipse information of each detected stroke. This data set was used as the predictand in the equations, determining whether or not lightning occurred on a particular day and during a particular half-hour interval in the database. These data were provided by Ms. Crawford of the AMU. The CGLSS data were filtered to include only lightning strikes recorded during the warm season between 0700-2400 EDT and only within the 5 NM lightning warning circles shown in Figure 4 to match the forecast from the Objective Lightning Probability tool, which in turn was selected to match the 45 WS daily planning forecast. Development of the predictand was based on whether lightning was observed in the time period and the warning circles on each day. The calculations considered only if at least one stroke was detected, not how many lightning strokes were detected. Calculation of the predictand was straightforward: a '1' was assigned as the predictand if at least one cloud-to-ground lightning stroke was detected within the defined time frame and spatial area on a specific day, otherwise a '0' was assigned (Lambert 2007).

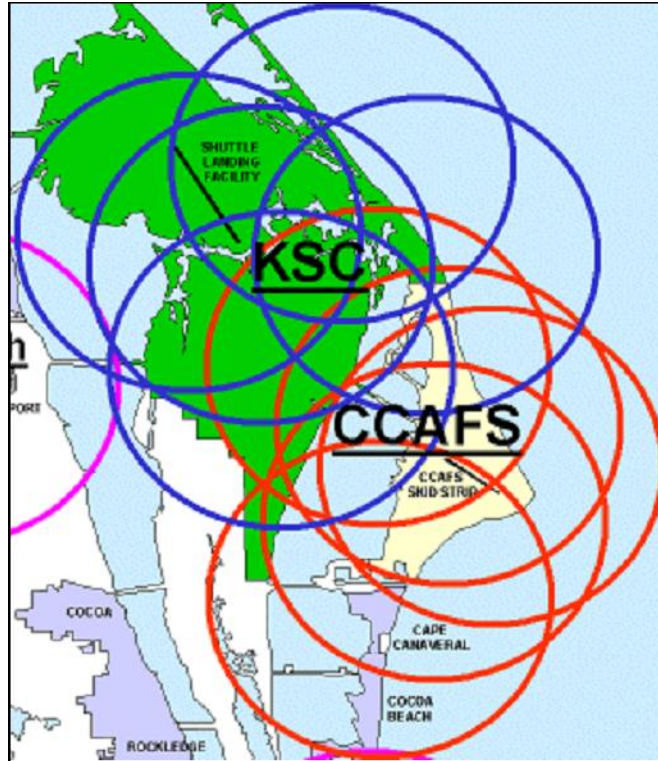


Figure 4. The 5-NM lightning warning circles on KSC (blue) and CCAFS (red).

2.2 GPS-IPW

The GPS-IPW data from May through October 2000-2011 were downloaded from the National Oceanic and Atmospheric Administration Earth System Research Laboratory (ESRL) Ground-Based GPS Meteorology web site <http://gpsmet.noaa.gov/cgi-bin/gnuplots/rti.cgi>. While the physical site near the Cape Canaveral Lighthouse (Figure 5) remained approximately the same over the 12-year POR, the 4-letter site ID changed when there were significant changes/upgrades in the equipment. Mr. Seth Gutman of the ESRL provided the information as to when these changes/upgrades occurred.

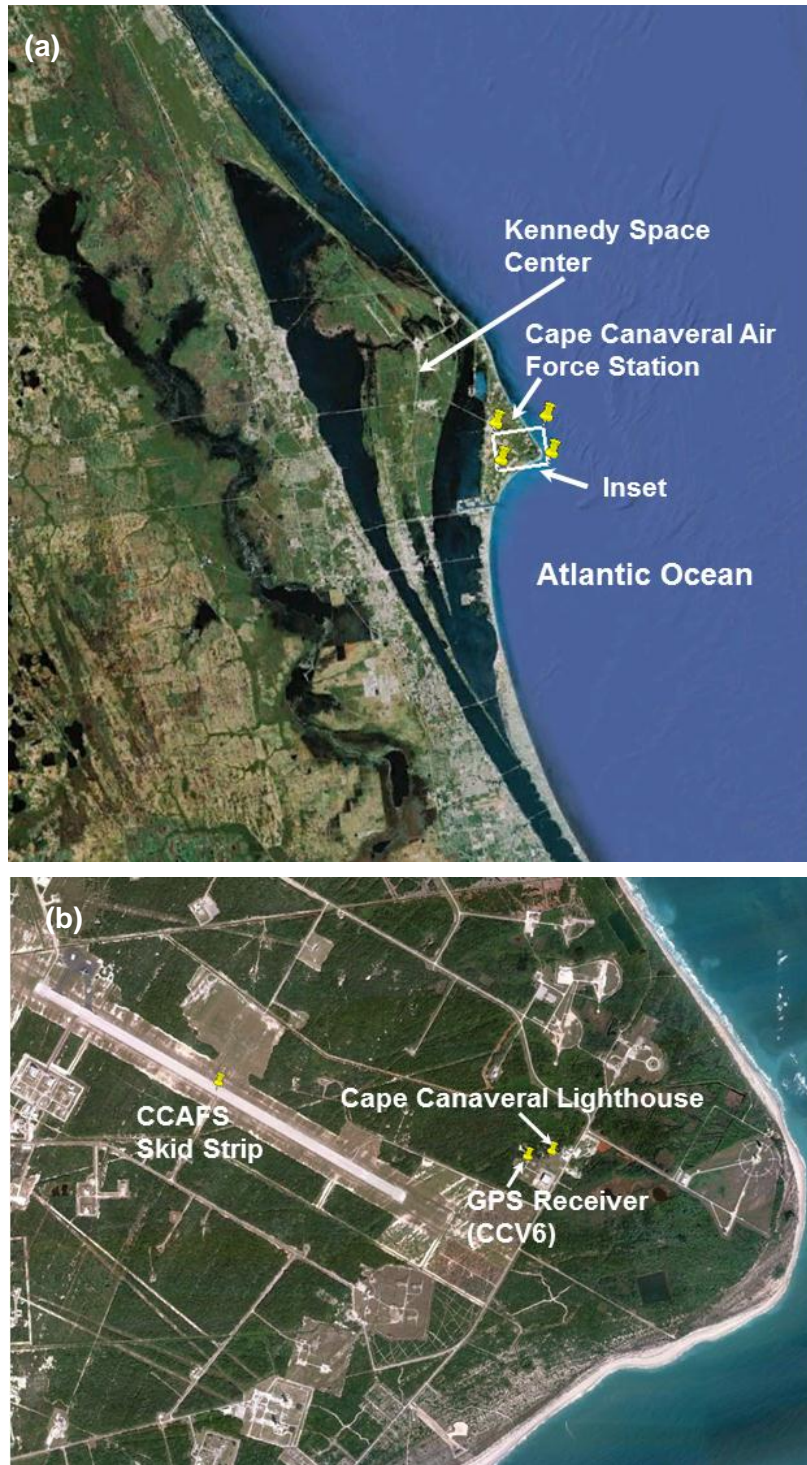


Figure 5. (a) Overview map of Cape Canaveral Air Force Station (CCAFS) and Kennedy Space Center (KSC) vicinity. (b) Inset map of CCAFS showing the location of the GPS receiver in relation to the skid strip and Cape Canaveral lighthouse.

2.3 Objective Lightning Probability Tool Output

The Objective Lightning Probability tool is a set of logistic regression equations that calculates the probability of lightning occurrence for the day in the 5 NM lightning warning circles on KSC and CCAFS (Lambert 2007).

Ms. Crawford provided the probability output from the Objective Lightning Probability tool for the POR. Along with the probability, the data included year, month and day, whether lightning occurred on each day (0 if no, 1 if yes), and the flow regime name for each day. The tool provides the probability of lightning for the entire day (valid between 0700-2400 EDT), so there was only one value per day.

3. Equation Development

The element to be forecast is lightning occurrence, which is a binary variable. That is, did lightning occur within the specified time period and area of interest: Yes, or 1, meant lightning occurred and no, or 0, meant lightning did not occur. Accordingly, a multiple logistic regression model is appropriate for this study. As in Kehrer et al. (2008), 2-hour and 9-hour forecast models were developed using data from the entire warm season POR from 2000 to 2011. These models were limited to the hours between 0700-2400 EDT to be consistent with the Objective Lightning Probability tool. The AMU divided the data into development and verification data sets and logistic regression equations were computed from the development data set. The output of the logistic regression equation is a lightning index that gives the probability that lightning will occur. The performance of the logistic regression equations is tested using the validation data set.

3.1 Predictand

The predictand is a lightning index between 0 (lightning did not occur) and 1 (lightning did occur). The predictand is then compared with various thresholds in an attempt to predict when lightning will occur within any of the 5 NM warning circles on KSC/CCAFS. As in Kehrer et al. (2008), this is used for a 2-hour forecast model as well as 9-hour forecast model. The 2-hour and 9-hour forecast models are limited to the hours between 0700-2400 EDT to be consistent with the Objective Lightning Probability tool.

3.2 Candidate Predictors

Because many of the variables considered in previous studies (Mazany et al. 2000 and Kehrer et al. 2008) such as KI, Total Totals (TT), and upper level moisture variables, were already considered by the AMU Objective Lightning Probability tool, only the objective lightning probability values along with the current GPS-IPW values and changes in GPS-IPW values were considered in this task. With the objective lightning probability value, the current GPS-IPW and the change in GPS-IPW over 0.5 to 24 hours in half-hour increments, there were 50 predictors available for equation development.

3.3 Data availability

For the entire warm season POR, 105,888 records were possible in the dataset, recording GPS-IPW every 0.5 hour between May 1 and October 31 for 12 years. However, there were numerous GPS-IPW sensor outages during this period. These outages affected numerous records because to find the GPS-IPW differences over time periods varying from 0.5 to 24.0 hours, the outages would affect various GPS-IPW time period differences. The only records selected for the final total data set were ones in which no data was missing for any of the GPS-IPW time period differences and also no data missing from the Objective Lightning Probability tool. The final record count, removing missing data, was 33,630 records. The record counts by month are shown in Table 1.

Table 1. GPS-IPW records available initially for the 12-year POR and the final amount after removing missing data.				
Month	Initial # Records		Final # Records	
	# Records	Percent	# Records	Percent
May	17,856	17%	6,161	18%
June	17,280	16%	6,498	19%
July	17,856	17%	6,299	19%
August	17,856	17%	5,656	17%
September	17,280	16%	4,395	13%
October	17,760	17%	4,621	14%
Total	105,888	100%	33,630	100%

3.4 Development and Verification Data Sets

The regression equations needed to be tested on a sample of data not involved in its development so the model can be independently tested (Wilks 2006). The AMU developed the verification data set using a random sample of 25% of the GPS-IPW data available every 0.5 hour between May 1 and October 31 for 12 years (Table 1). The counts of records in the development and validation data sets are shown in Table 2. This allowed 75% of the data to be used to develop the equations.

Table 2. GPS-IPW records counts in the development data set and validation data set.				
Month	Development		Validation	
	# Records	Percent	# Records	Percent
May	4,599	18%	1,562	19%
June	4,843	19%	1,655	20%
July	4,757	19%	1,542	18%
August	4,227	17%	1,429	17%
September	3,273	13%	1,122	13%
October	3,516	14%	1,105	13%
Total	25,215	100%	8,415	100%

3.5 Logistic Regression

A multiple, logistic regression model was the best choice model for this study. Logistic regression is the appropriate model to use when the predictand, or element to be forecast is binary and probabilities of the predictand are desired. In this case, the element to be forecast was lightning occurrence: yes, or 1, meant lightning occurred and no, or 0, meant the lightning did not occur within the specified time period and area of interest. More details about the background and theory of multiple logistic regression can be found in Wilks (2006).

3.6 Predictor Selection Methodologies

Because there were 50 candidate predictors for these models, the goal was to determine a subset of predictors that affected the predictand to create a model that fits the data well without the negative effects of overfitting the model. Models that are overfit tend to be too closely molded to the development data set and the fitted relationship degrades when used with independent, verification data, such as for operational forecasts (Wilks 2006).

The AMU used two methods of predictor selection for this task: forward selection and backward elimination. These selection methods are described in Sections 3.6.1 and 3.6.2. The two selection methods chose different regressors for the final model. Predictors were added or removed based on a p-value¹ of a stringent 0.01 significance level in order to match the level selected by Mazany et al. (2002).

The equation output is a lightning index that gives the probability that lightning will occur. Performance metrics for each model included the HR, FAR, POD, KSS, OUI, and bias at various lightning index thresholds. The lightning index threshold is the point at which lightning is predicted when the model probability falls above the threshold and not predicted when model probability falls below the threshold (Kehrer et al. 2008).

Multicollinearity occurs when two or more independent variables in the model are approximately determined by a linear combination of other independent variables in the model. Of the 50 candidate predictor variables, 48 are a linear combination of the current GPS-IPW value and one of the half hour interval values between 0.5 and 24.0 hours. Therefore, there could be some multicollinearity of the independent variables. In Section 4.4, a factor analysis is performed to remove some of the independent variables. In Section 5, performance metrics are recalculated to shown any degradation in model performance due to the removal of some of the independent variables due to collinearity.

3.6.1 Forward Selection Method

Forward selection (Figure 6) starts with no predictors (independent variables) in the model. Each predictor is evaluated individually and if any predictors possess a p-value below the pre-specified value, 0.01 in this case, the one with the lowest p-value is added to the model and remains in the model. This process continues until no predictor has a p-value better than the cut-off specified (Anderson et al. 2012).

¹ The p-value gives the probability that a regression coefficient, significantly different from zero, could have occurred by chance (merely as a result of sampling error), assuming the hypothesis that the regression coefficient equals zero is true. An independent variable with a regression coefficient not significantly different from 0 ($p > 0.01$) can be removed from the regression model. If $p < 0.01$ then the variable contributes significantly to the prediction of the outcome variable. (Kohler 1985)

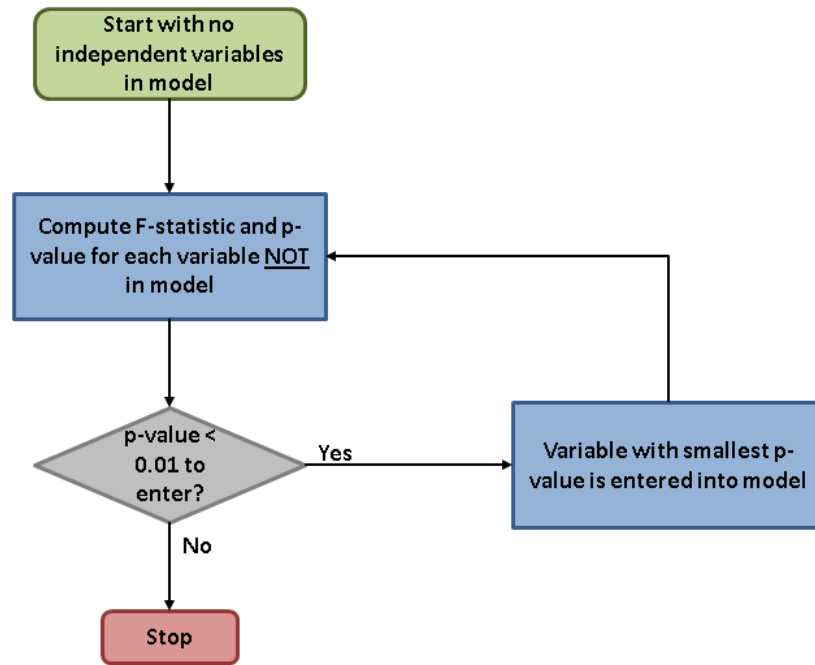


Figure 6. Flowchart depicting the forward selection method of choosing model predictors (Anderson et al. 2012).

3.6.2 Backward Elimination Method

Backward elimination (Figure 7) begins with a model consisting of all potential predictors. The model's results are obtained and if any predictors are above the pre-specified p-value of 0.01, the predictor with the highest p-value is removed from the model and not allowed to return. Once every predictor p-value is smaller than the p-value cut-off specified, the backward elimination method stops (Anderson et al. 2012).

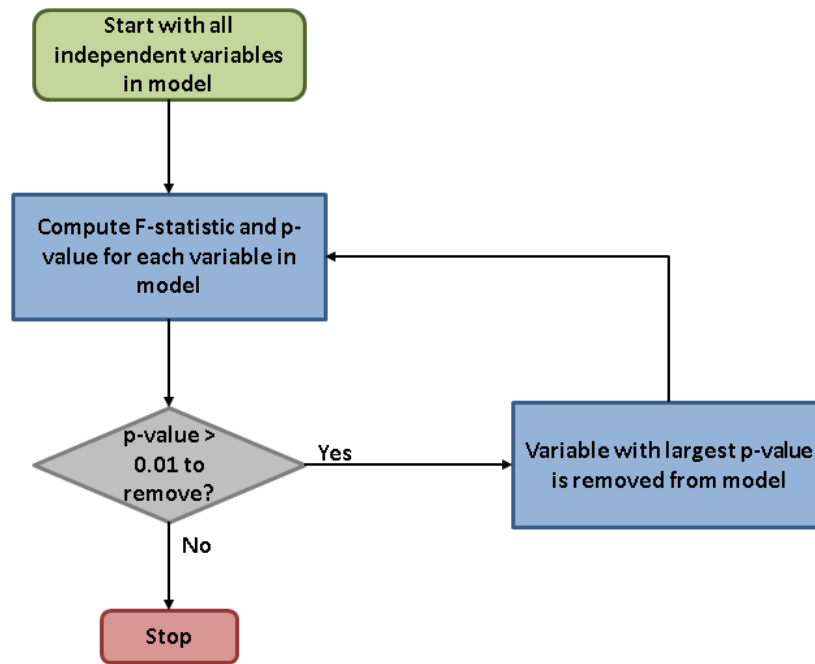


Figure 7. Flowchart depicting the backward elimination method of choosing model predictors (Anderson et al. 2012).

4. Logistic Regression Equation Verification

The logistic regression equation predictand is a binary lightning index where 0 means no lightning occurred and 1 means lightning did occur. The predictand is then compared with various thresholds in an attempt to predict when lightning will occur within any of the 5 NM warning circles on KSC/CCAFS. The final logistic regression equations were in the form of

$$\hat{y} = \frac{1}{1 + \exp(b_0 + b_1x_1 + \dots + b_kx_k)},$$

where b_0 is the intercept, b_k is the coefficient for the predictor, and x_k is the predictor. The number of predictors selected for the final equations ranged from 8 to 19. The predictors selected and their coefficients for the four predictor selection methods are shown in Section 4.3. Validation of the model is accomplished through an evaluation of the Relative Operating Characteristics (ROC) curves and the associated reliability diagrams. In addition, the AMU examined collinearity of the predictors.

4.1 Relative Operating Characteristics (ROC)

The AMU developed ROC curves for each model as a graphical way of showing the model's ability to correctly anticipate the occurrence or non-occurrence of lightning. The ROC curve (Figure 8) is useful in helping to identify an optimum warning criterion by indicating the trade-off between misses and false alarms (Mason and Graham 1999). Generally, for a skillful forecast system, the ROC curve bends towards the top left, where hit rates are larger than false alarm rates and the area under the curve (AUC) is greater than 0.5. An example is shown in Figure 8. The point nearest the upper left corner is the If the curve is close to the 45° diagonal, the AUC is near 0.5 and the forecast system does not provide any useful information. If the curve lies below the 45° diagonal, the AUC is less than 0.5 and the forecast system provides negative skill (Mason and Graham 1999). Mason and Graham (1999) provides a simple transformation, S, such that $S = 2 \times (AUC - 0.5)$. The range of S is from 1.0 (for a perfect forecast) to -1.0 (for the worst forecast), with 0.0 indicating no skill. For the equations shown in Section 4.3, $S < 0.5$ for most of the predictors except the objective lightning probability, which was always > 0.5 . S was also > 0.5 for the logistic regression equation in each case, but not significantly greater than S for the objective lightning probability.

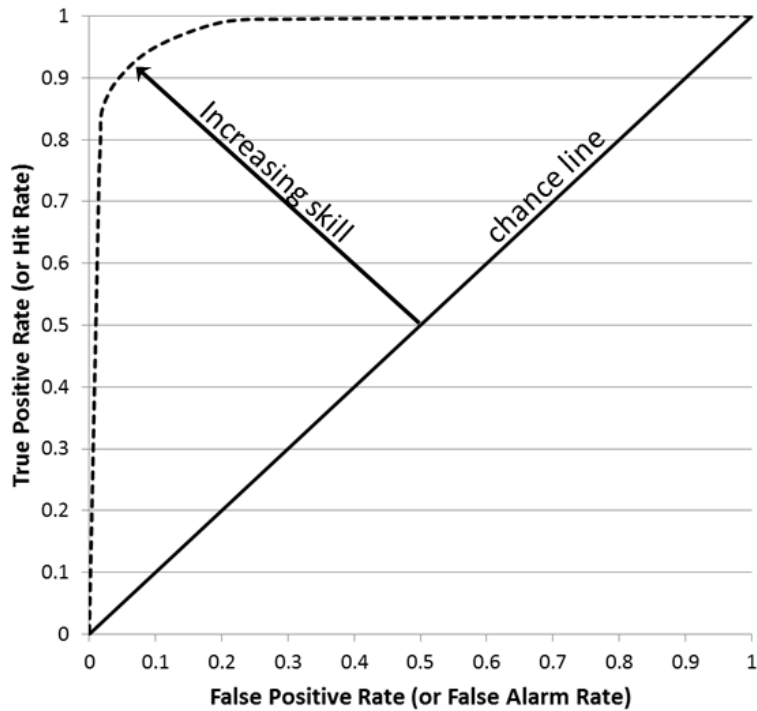


Figure 8. A basic ROC diagram showing the false positive rate (or false alarm rate) on the x-axis and the true positive rate (or hit rate) on the y-axis. The 45° line is the chance line, representing a forecast system that has no skill, and the dashed curve bending towards the upper left corner represents a skillful forecast system.

4.2 Reliability Diagrams

In a reliability diagram, a dashed diagonal line represents perfect reliability and another curve represents the reliability of the forecast equation. An example of an overconfident reliability diagram is shown in Figure 9. A histogram is usually shown with it and gives the frequency of the number of observations in each probability range. When the forecast equation curve is below the dashed diagonal line, the equation over-forecasted lightning occurrence. When the curve is above the dashed line, the equations under-forecasted lightning occurrence. The four logistic regression equations performed well at lightning probabilities of 0.2 and below. Above 0.2, however, the reliability curve of the equations is below the dashed line indicating that the equations over-forecasted lightning occurrence. Regardless of the selection method used, the logistic regression equations only produced lightning indexes greater than 0.4 only 0.1% of the time. Therefore these values were treated as extreme and are not shown in the reliability diagrams. It is important to note that reliability is just one measure of performance. One could have a perfectly reliable forecast and still have a forecast that adds no value. For example, one could forecast only climatology over a long period and the forecast would be very reliable, but the forecast would be little value added. Reliability must be combined with sharpness to measure overall value. Conversely, one could issue perfectly sharp forecasts, only forecasts of 0% and 100%, but would be of little value if they weren't reliable over a long period. Again, a good forecast needs to be both reliable and sharp. Sharpness is often shown in sharpness diagrams that show the frequency that probability bins were forecast.

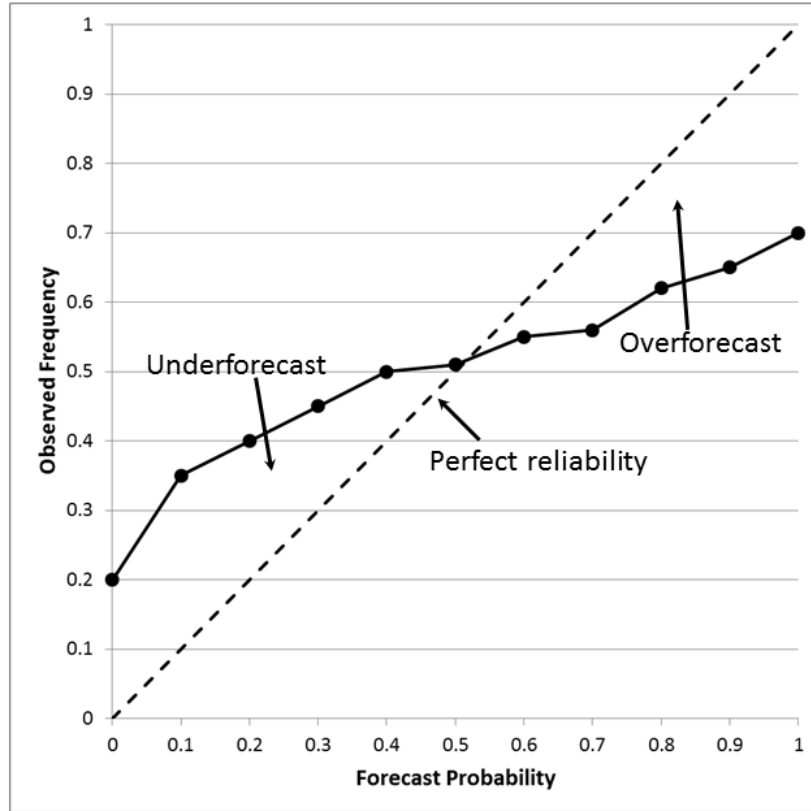


Figure 9. Example reliability diagram for an overconfident forecast. The dashed line represents a perfectly reliable forecast. Here the event occurs more frequently than indicated when the forecast indicates a decreased probability of the event occurring (to the left of the dotted line), but less frequently than indicated when the forecast indicates an increased probability of the event occurring (to the right of the dotted line). Although the forecasts correctly indicate increases and decreases in the probabilities of the events, the changes in probability are over-stated, and the forecasts are said to be over-confident.

4.3 Equation Performance

The following sections contain tables and charts that show the equation parameters and their performance statistics using ROC curves and reliability diagrams for each of the four equations developed by the AMU. The S values for the ROC curves, each predictor, and the AUC are shown in the tables in each section. The ROC curves and reliability diagrams for the logistic regression variables are shown after the tables in each section. Each table and chart shows clearly that the objective lightning probability predictor accounts for most of the variability in the logistic regression equations in all four selection methods. The current GPS-IPW predictor was important in the 2-hour forecast regression equation using the forward selection method of variable selection. However, the remaining predictors offered not much more predictability than would be expected by random chance.

4.3.1 Forward Selection, 2-Hour Forecast Equation

The 14 predictors and their coefficients for this equation are shown in the first two columns of Table 3. The AUC and S values are in the last two columns. Figure 10 contains the ROC curves for each of the variables and for the equation. Figure 11 shows the reliability curve up to a probability of 0.4. A sharpness diagram is inset into the corner of Figure 11. Higher probabilities were not shown because the logistic regression equations only produced lightning indexes greater than 0.4 only 0.1% of the time. These probabilities were treated as extreme and are not shown in the reliability diagrams.

Table 3. Area under the ROC curve and forecast skill index, S, for each predictor in the 2-hour forecast regression equation using the forward selection method of predictor selection, where b_0 is -7.157			
Coefficients (b_1 to $k=14$)	Predictor (x_1 to $k=14$)	Area Under Curve (AUC)	S
3.301	Objective Lightning Tool	0.781	0.56
0.454	Current GPS-IPW	0.758	0.52
0.351	-0.5 hr Δ GPS-IPW	0.582	0.16
-0.198	-1.0 hr Δ GPS-IPW	0.599	0.20
-0.252	-1.5 hr Δ GPS-IPW	0.613	0.23
0.719	-2.0 hr Δ GPS-IPW	0.655	0.31
-0.178	-2.5 hr Δ GPS-IPW	0.660	0.32
0.438	-3.0 hr Δ GPS-IPW	0.653	0.31
0.054	-3.5 hr Δ GPS-IPW	0.664	0.33
-0.173	-4.0 hr Δ GPS-IPW	0.673	0.35
0.335	-4.5 hr Δ GPS-IPW	0.669	0.34
0.837	-5.5 hr Δ GPS-IPW	0.668	0.34
-0.349	-15.0 hr Δ GPS-IPW	0.573	0.15
0.226	-23.5 hr Δ GPS-IPW	0.623	0.25
	Logistic Regression	0.834	0.67

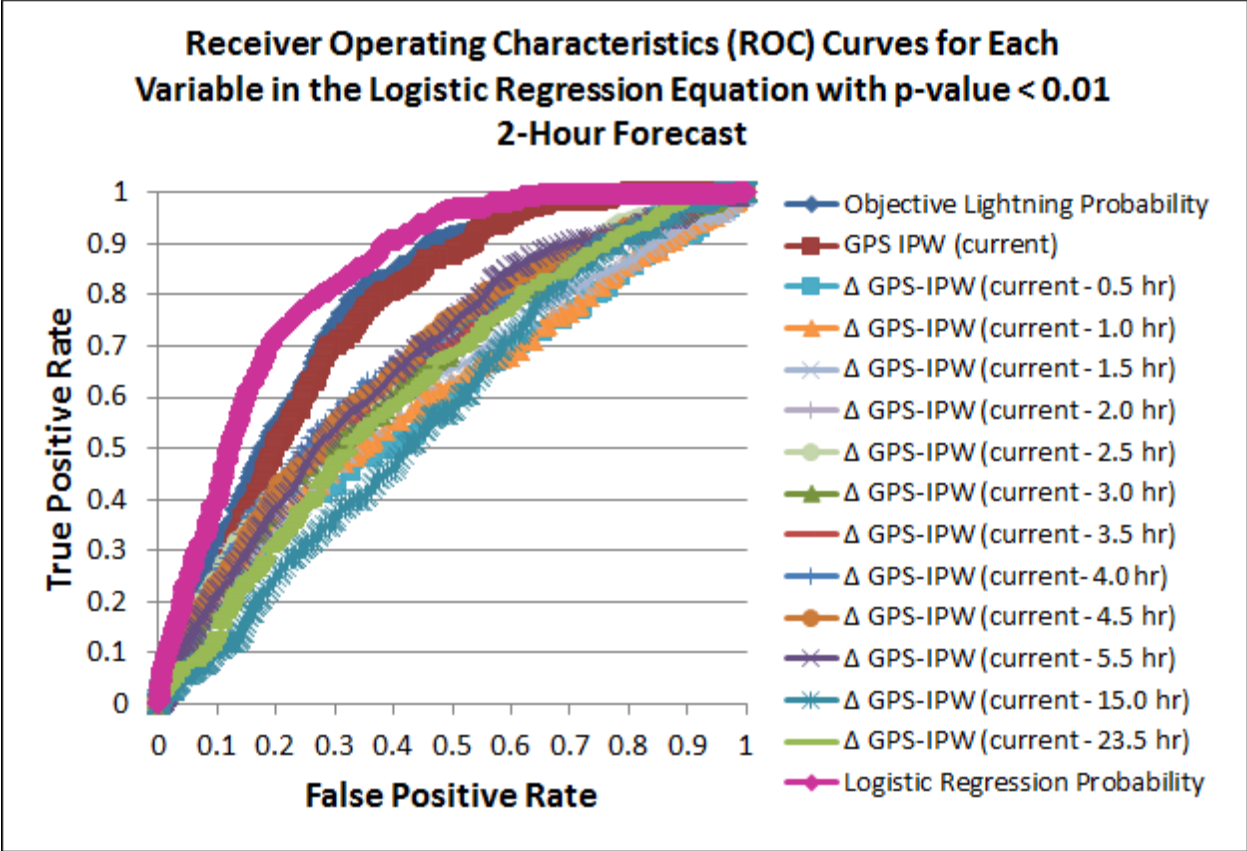


Figure 10. ROC curve for each variable in the 2-hour forecast regression equation using the forward selection method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability and the current GPS-IPW are the best predictors. The rest of the GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.

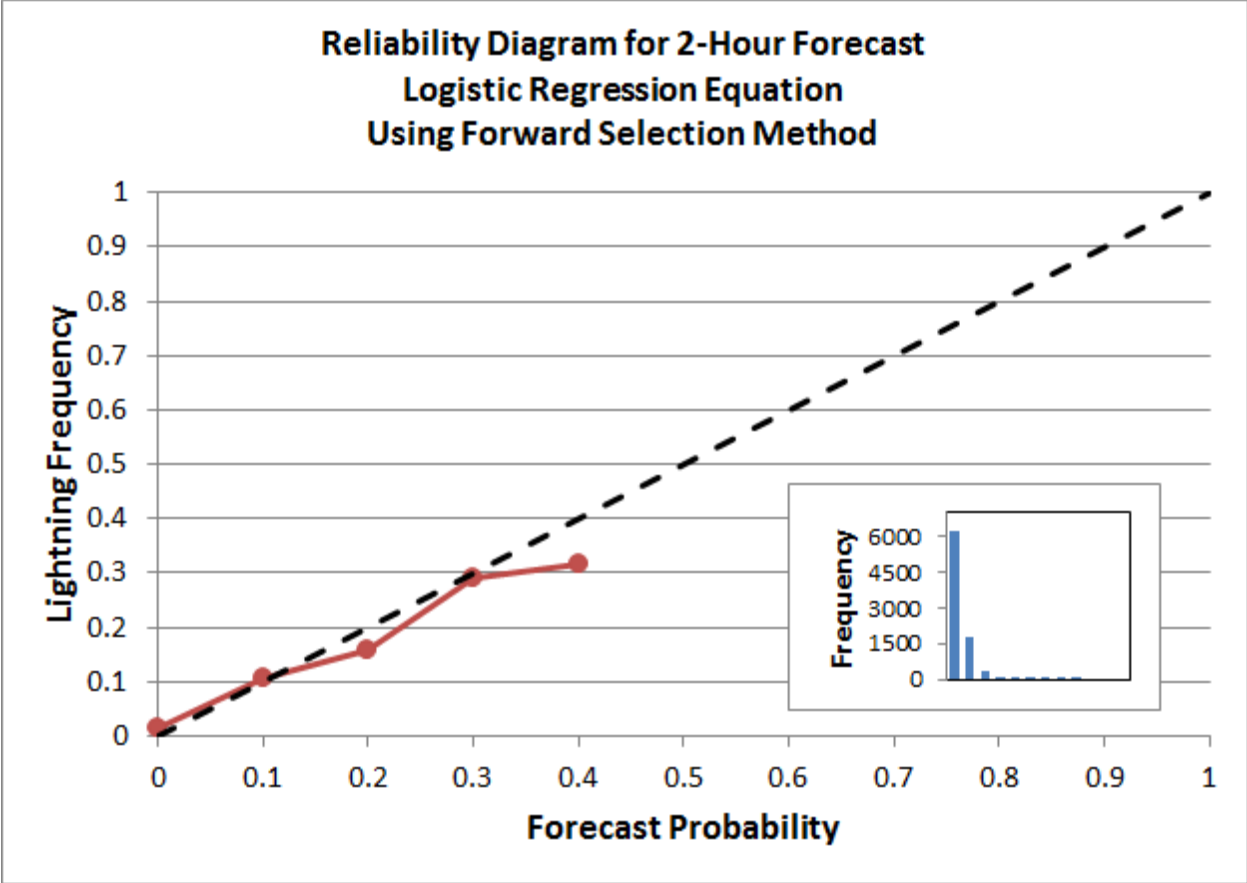


Figure 11. Reliability diagram for the 2-hour forecast equation using the forward selection method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 2-hour forecast equation. The reliability is good with slight overforecasting in the small range available. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.

4.3.2 Forward Selection, 9-Hour Forecast Equation

The 19 predictors and their coefficients for this equation are shown in the first two columns of Table 4. The AUC and S values are in the last two columns. Figure 12 contains the ROC curves for each of the variables and for the equation. Figure 13 shows the reliability curve up to a probability of 0.4. A sharpness diagram is inset into the corner of Figure 13. Higher probabilities were not shown because the equations either did not calculate values higher than 0.4 or calculated very few values in the higher range.

Table 4. Area under the ROC curve and forecast skill index, S, for each predictor in the 9-hour forecast regression equation using the forward selection method of predictor selection, where b_0 is -5.093.

Coefficients (b_1 to $k=19$)	Predictor (x_1 to $k=19$)	Area Under Curve (AUC)	S
3.706	Objective Lightning Tool	0.806	0.61
0.650	-5.0 hr Δ GPS-IPW	0.441	-0.12
0.192	-5.5 hr Δ GPS-IPW	0.436	-0.13
-0.495	-6.0 hr Δ GPS-IPW	0.438	-0.12
0.046	-6.5 hr Δ GPS-IPW	0.427	-0.15
-0.993	-7.0 hr Δ GPS-IPW	0.424	-0.15
-0.374	-13.5 hr Δ GPS-IPW	0.460	-0.08
0.250	-14.5 hr Δ GPS-IPW	0.479	-0.04
-0.235	-16.5 hr Δ GPS-IPW	0.504	0.01
-0.239	-17.5 hr Δ GPS-IPW	0.529	0.06
0.112	-18.0 hr Δ GPS-IPW	0.535	0.07
-0.380	-18.5 hr Δ GPS-IPW	0.541	0.08
0.424	-19.0 hr Δ GPS-IPW	0.549	0.10
0.167	-19.5 hr Δ GPS-IPW	0.549	0.10
-0.294	-20.0 hr Δ GPS-IPW	0.555	0.11
0.313	-20.5 hr Δ GPS-IPW	0.561	0.12
0.223	-21.0 hr Δ GPS-IPW	0.564	0.13
0.125	-22.5 hr Δ GPS-IPW	0.571	0.14
0.379	-24.0 hr Δ GPS-IPW	0.571	0.14
	Logistic Regression	0.822	0.64

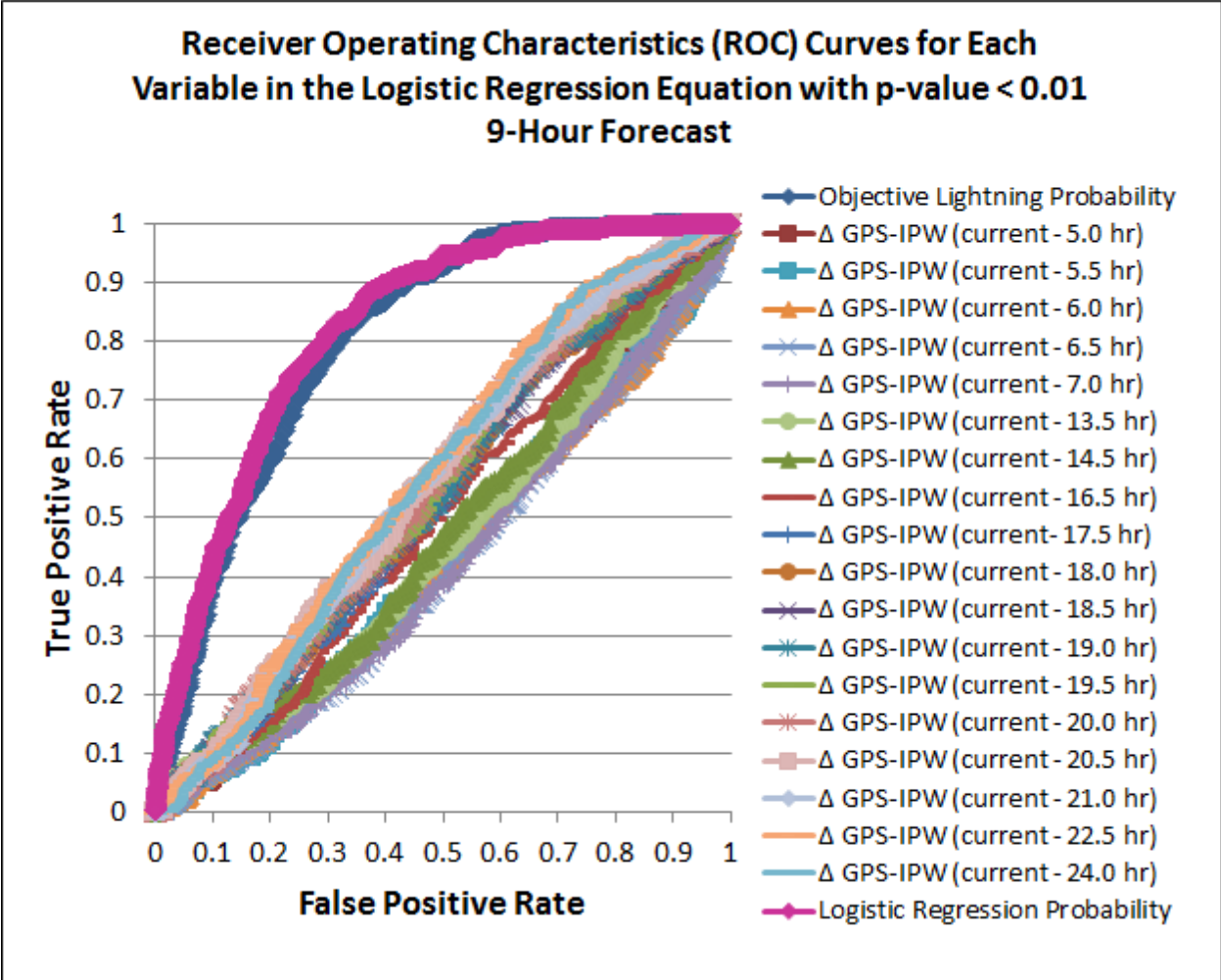


Figure 12. ROC curve for each variable in the 9-hour forecast regression equation using the forward selection method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability is the best predictor. GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.

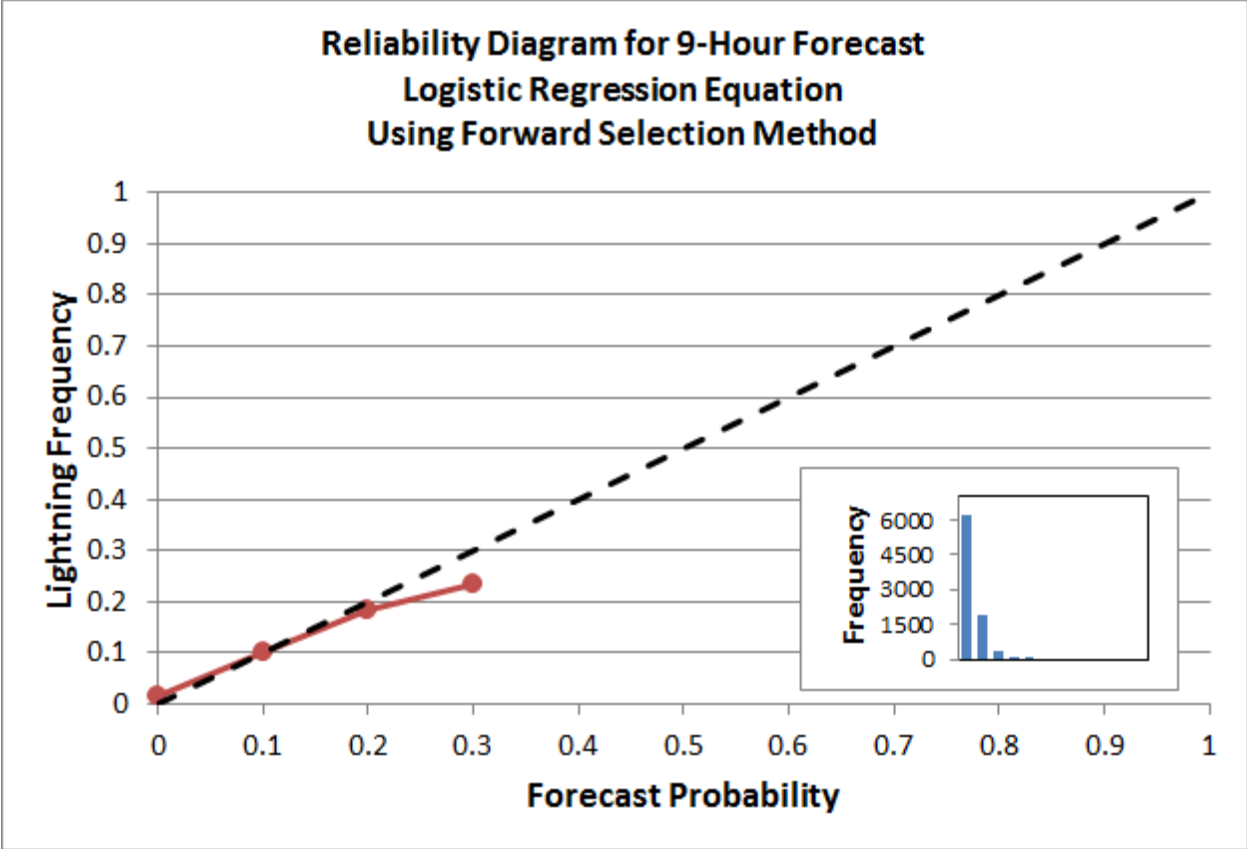


Figure 13. Reliability diagram for the 9-hour forecast equation using the forward selection method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 9-hour forecast equation. The reliability is good with slight overforecasting in the small range available. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.

4.3.3 Backward Elimination, 2-Hour Forecast Equation

The eight predictors and their coefficients for this equation are shown in the first two columns of Table 5. The AUC and S values are in the last two columns. Figure 14 contains the ROC curves for each of the variables and for the equation. Figure 15 shows the reliability curve up to a probability of 0.4. A sharpness diagram is inset into the corner of Figure 15. Higher probabilities were not shown because the equations either did not calculate values higher than 0.4 or calculated very few values in the higher range.

Table 5. Area under the ROC curve and forecast skill index, S, for each variable in the 2-hour forecast regression equation using the backward elimination method of variable selection, where b_0 is -5.359.

Coefficients (b_1 to $k=8$)	Predictor (x_1 to $k=8$)	Area Under Curve (AUC)	S
3.901	Objective Lightning Tool	0.781	0.56
0.729	-3.0 hr Δ GPS-IPW	0.654	0.31
1.018	-5.5 hr Δ GPS-IPW	0.668	0.34
0.616	-18.0 hr Δ GPS-IPW	0.565	0.13
0.742	-18.5 hr Δ GPS-IPW	0.565	0.13
-0.866	-20.0 hr Δ GPS-IPW	0.571	0.14
-0.591	-21.0 hr Δ GPS-IPW	0.589	0.18
0.472	-24.0 hr Δ GPS-IPW	0.627	0.25
	Logistic Regression	0.828	0.66

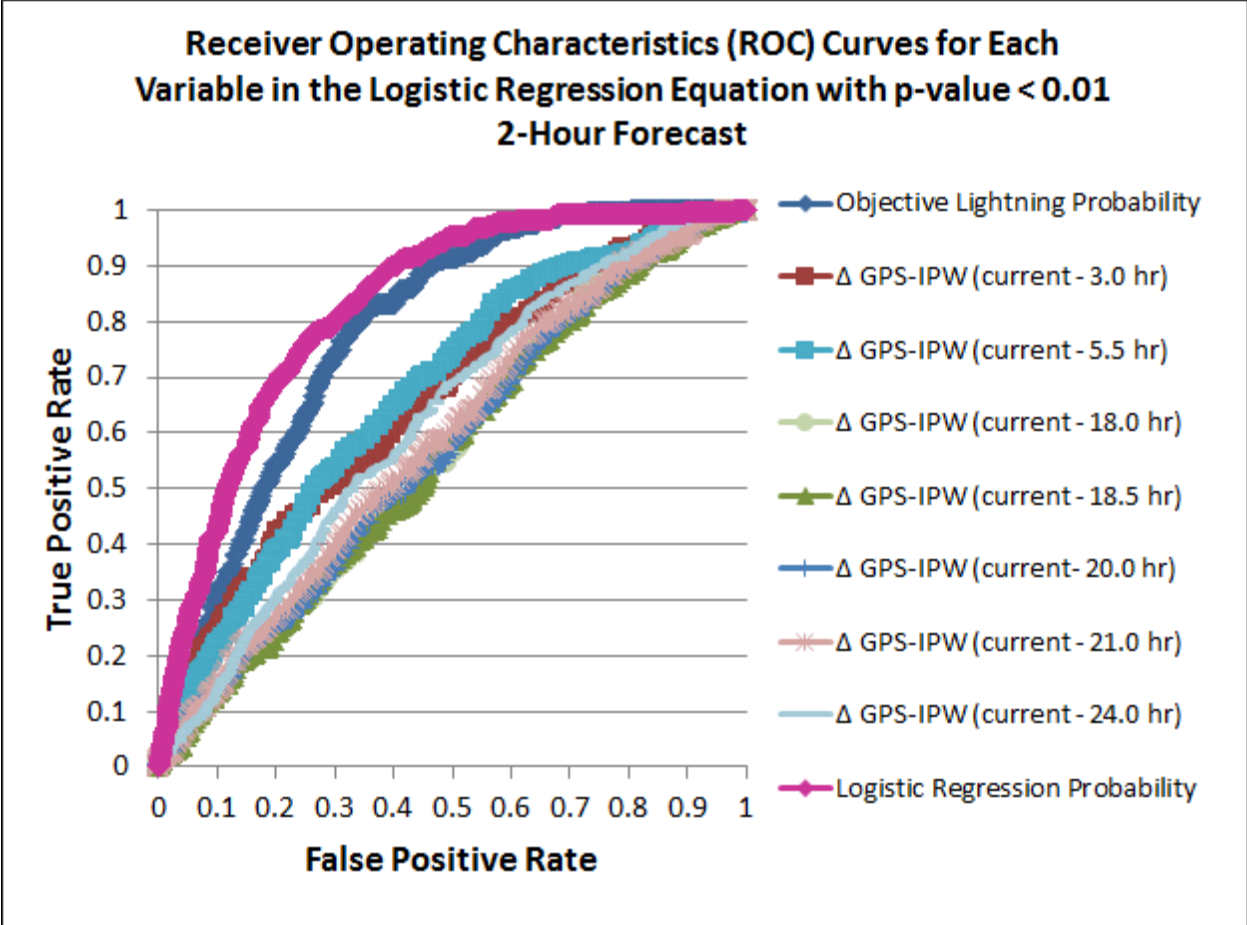


Figure 14. ROC curve for each variable in the 2-hour forecast regression equation using the backward elimination method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability is the best predictor. GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.

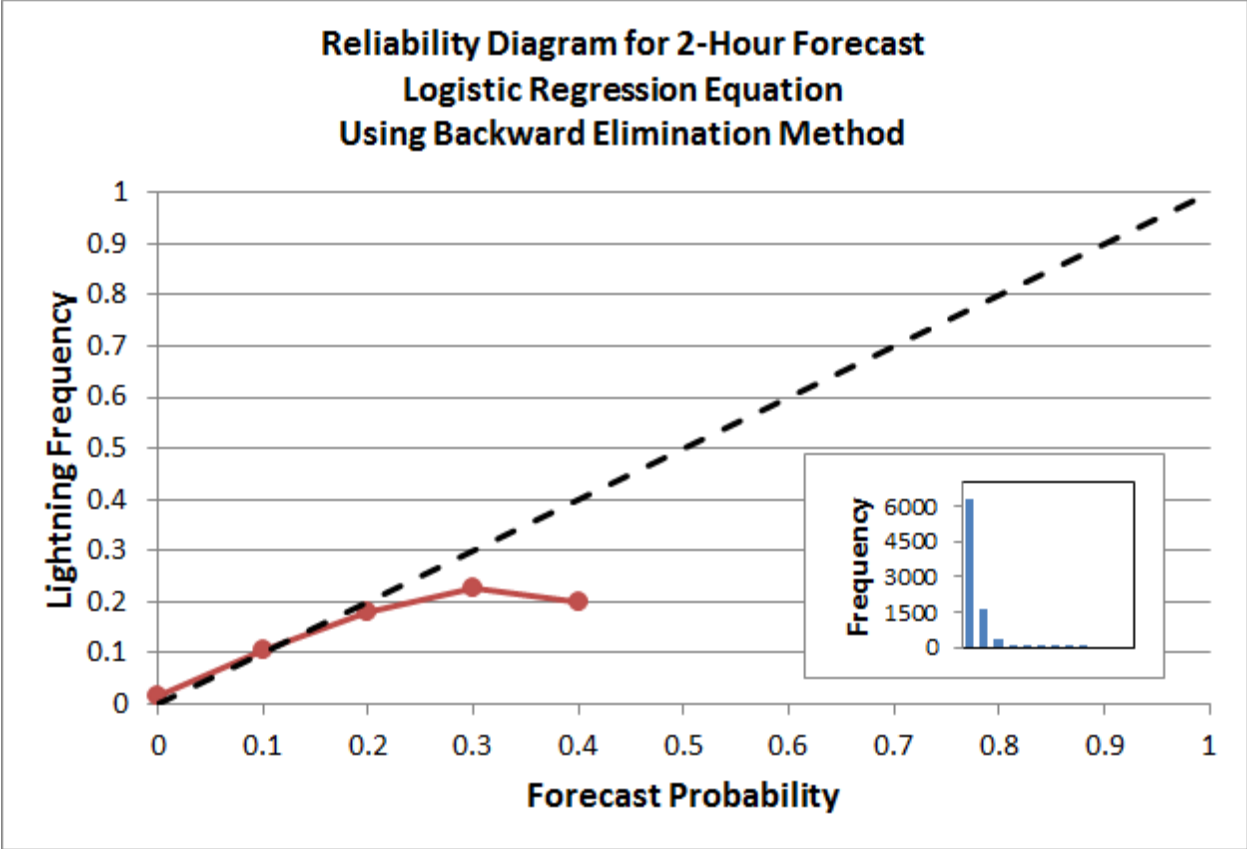


Figure 15. Reliability diagram for the 2-hour forecast equation using the backward elimination method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 2-hour forecast equation. The reliability is good in the 0 to 0.2 range, but increasing overforecasting in the 0.2 to 0.4 range. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.

4.3.4 Backward Elimination, 9-Hour Forecast Equation

The eight predictors and their coefficients for this equation are shown in the first two columns of Table 6. The AUC and S values are in the last two columns. Figure 16 contains the ROC curves for each of the variables and for the equation. Figure 17 shows the reliability curve up to a probability of 0.4. A sharpness diagram is inset into the corner of Figure 17. Higher probabilities were not shown because the equations either did not calculate values higher than 0.4 or calculated very few values in the higher range.

Table 6. Area under the ROC curve and forecast skill index, S, for each variable in the 9-hour forecast regression equation using the backward elimination method of variable selection, where b_0 is -5.100.

Coefficients (b_1 to $k=8$)	Predictor (x_1 to $k=8$)	Area Under Curve (AUC)	S
3.705	Objective Lightning Tool	0.806	0.61
0.604	-4.0 hr Δ GPS-IPW	0.468	-0.06
-1.107	-7.0 hr Δ GPS-IPW	0.436	-0.13
0.536	-10.5 hr Δ GPS-IPW	0.535	0.07
-0.748	-11.5 hr Δ GPS-IPW	0.541	0.08
-0.328	-17.0 hr Δ GPS-IPW	0.555	0.11
0.566	-21.0 hr Δ GPS-IPW	0.564	0.13
0.432	-24.0 hr Δ GPS-IPW	0.571	0.14
	Logistic Regression	0.816	0.63

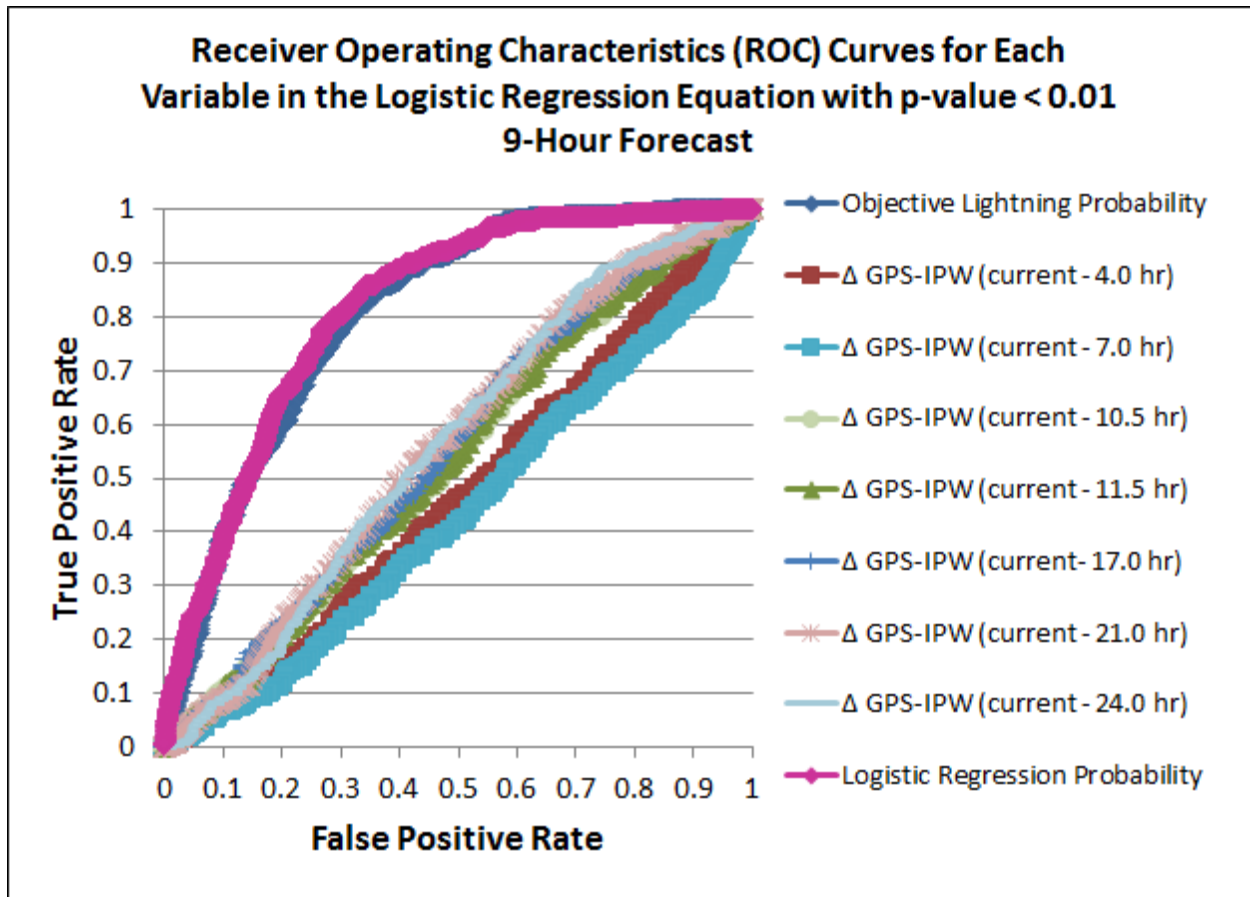


Figure 16. ROC curve for each variable in the 9-hour forecast regression equation using the backward elimination method of variable selection. Contributing variable plots are combined to create the overall logistic regression probability curve. Those individual plots most closely aligned with the logistic regression probability curve are the most accurate predictors. In this case, the objective lightning probability is the best predictor. GPS-IPW variables do not contribute much to the skill of the logistic regression probability curve.

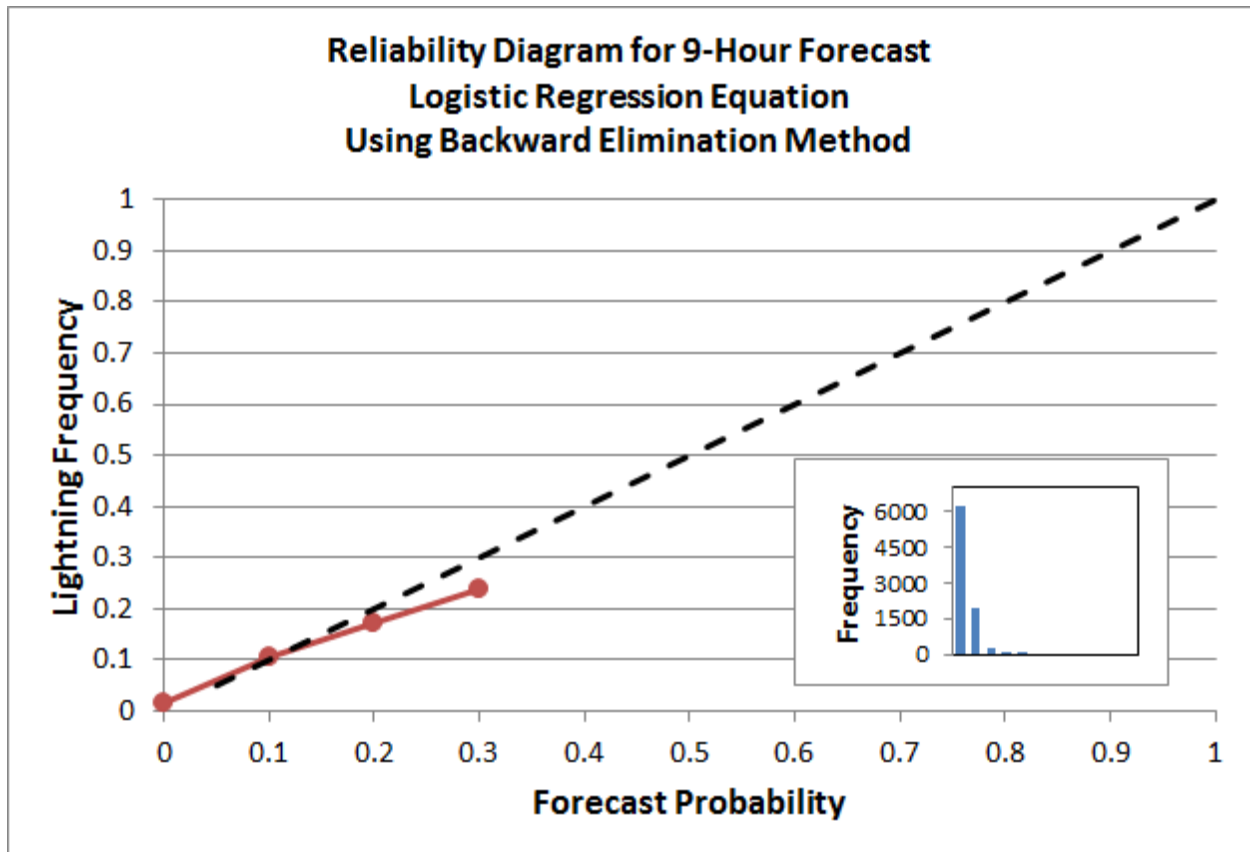


Figure 17. Reliability diagram for the 2-hour forecast equation using the backward elimination method of variable selection. The dashed diagonal line represents perfect reliability and the red curve represents the reliability of the 2-hour forecast equation. The reliability is good with slight overforecasting in the small range available. The histogram in the lower right shows the frequency of the number of observations in each probability range. The x-axis of the histogram is the same as the forecast probability axis on the reliability diagram.

4.4 Predictor Collinearity and Scree Plots

Collinearity can make the predictors redundant in some cases. It is common to evaluate a scree plot to determine how many predictors should be included in a model. The scree plot is a two dimensional graph with predictors on the x-axis and eigenvalues on the y-axis. The scree plot is a graphical representation of the incremental variance accounted for by the number of predictors in the model. Generally, the number of predictors that should be in the model are limited to those with an eigenvalue $> 1^2$, but selecting the number of factors involves a certain amount of subjective judgment. A predictor with an eigenvalue < 1 means the variable is not contributing an average amount to explaining the variance in the model (Walker and Maddan 2009). The selection of variables can be automated, but subjectivity is still required in picking thresholds to accept or reject a variable. Based on this analysis and the results of the above

² In factor analysis, a component's eigenvalue is the amount of variance the component explains. The major reason for this is the eigenvalue's definition as a weighted sum of squared correlations. However, the actual variance of the component scores also equals the eigenvalue. Thus in factor analysis the "factor variance" and "amount of variance the factor explains" are equal. Therefore the two phrases are often used interchangeably, even though conceptually they stand for very different quantities (Darlington 2012).

ROC diagrams, new equations were developed to see if reducing the number of factors in the model substantially impacted the performance metrics (shown in Table 7 and Table 8.)

Four predictors had eigenvalues greater than 1 in the 2-hour forecast regression equation using the forward selection method for predictor selection. The four predictors in the truncated regression equation were selected by examining the Figure 10 ROC curve and selecting the four predictors that had the highest AUC. For this case, the variables selected were the objective lightning probability, the current value of the GPS-IPW, 4 hour change in GPS-IPW and the 4.5 hour change in GPS-IPW. The scree plot for the 2-hour forecast equation using the forward selection method of predictor selection is shown in **Error! Reference source not found.** The new equation is shown below.

$$\hat{y} = \frac{1}{1 + \exp(-7.509 + 2.864x_1 + 0.599x_2 + 1.310x_3 + 0.100x_4)}$$

where

- x_1 = Objective lightning probability,
- x_2 = Current GPS-IPW,
- x_3 = 4.0 hr Δ GPS-IPW and
- x_4 = 4.5 hr Δ GPS-IPW.

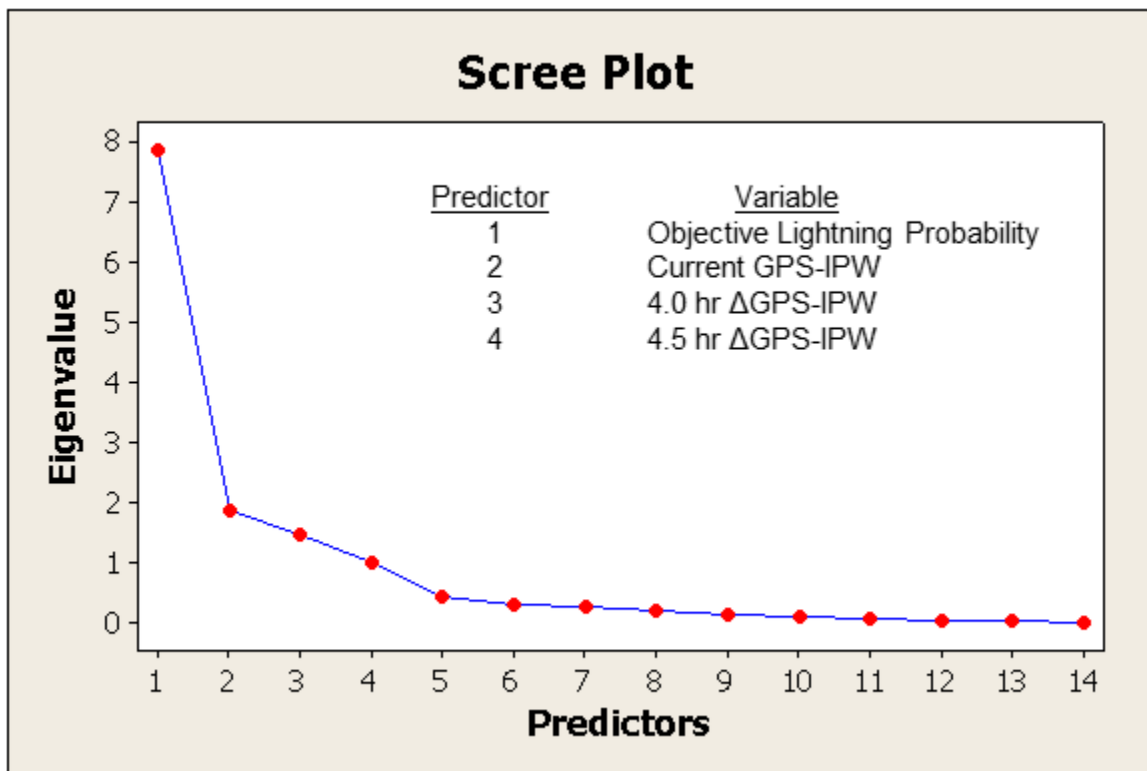


Figure 18. Scree plot for the predictors in the 2-hour forecast regression equation using the forward selection method of predictor selection. The first four variables have eigenvalues > 1 and are likely useful in a forecast equation.

The final equation from Kehrer et al (2008) using the forward selection method for the 2-hour forecast is shown below.

$$\hat{y} = \frac{1}{1 + \exp(-2.366 + 2.053x_1 + 0.538x_2 + 0.0310x_3 + 0.322x_4)}$$

where

$x_1 = 0.5$ hr Δ GPS-IPW,

$x_2 = 7.5$ hr Δ GPS-IPW,

$x_3 =$ K-index and

$x_4 =$ current GPS-IPW.

Three predictors had eigenvalues greater than one in the 9-hour forecast regression equation using the forward selection method for predictor selection. The three predictors in the truncated regression equation were selected by examining the Figure 12 ROC curve and selecting the three predictors that had the highest AUC. For this case, the variables selected were the objective lightning probability, the 22.5 hour change in GPS-IPW and the 24 hour change in GPS-IPW. The scree plot for the 9-hour forecast equation using the forward selection method of predictor selection is shown in Figure 19. The new equation is shown below.

$$\hat{y} = \frac{1}{1 + \exp(-5.137 + 4.038x_1 + 0.396x_2 - 0.127x_3)}$$

where

$x_1 =$ Objective lightning probability,

$x_2 = 22.5$ hr Δ GPS-IPW and

$x_3 = 24.0$ hr Δ GPS-IPW.

The final equation from Kehrer et al (2008) using the forward selection method for the 9-hour forecast is shown below.

$$\hat{y} = \frac{1}{1 + \exp(-4.885 + 0.541x_1 + 0.346x_2 - 0.446x_3 + 0.235x_4 + 0.071x_5)}$$

where

$x_1 =$ current GPS-IPW,

$x_2 = 3.5$ hr Δ GPS-IPW,

$x_3 = 8.5$ hr Δ GPS-IPW,

$x_4 = 12.0$ hr Δ GPS-IPW and

$x_5 =$ K-index.

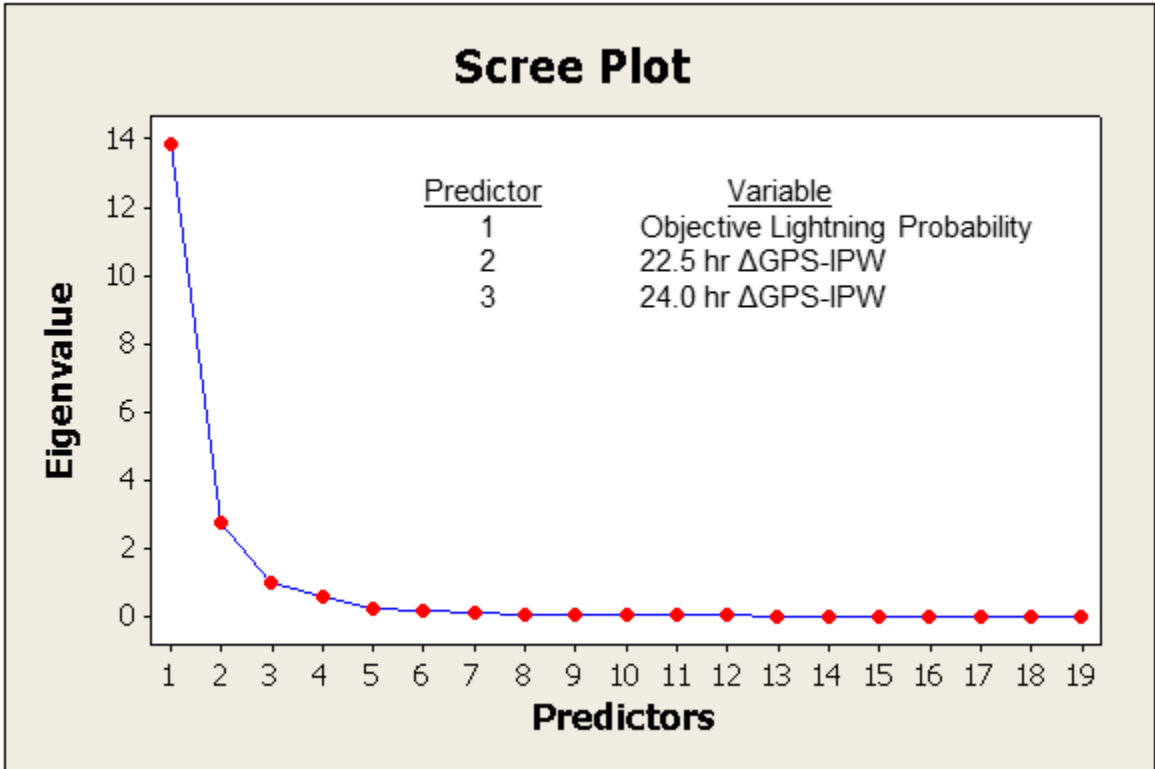


Figure 19. Scree plot for the predictors in the 9-hour forecast regression equation using the forward selection method of predictor selection. The first three variables have eigenvalues > 1 and are likely useful in a forecast equation.

Two predictors had eigenvalues greater than one in the 2-hour forecast regression equation using the backward elimination method for predictor selection. The two predictors in the truncated regression equation were selected by examining the Figure 14 ROC curve and selecting the two predictors that had the highest AUC. For this case, the variables selected were the objective lightning probability and the 5.5 hour change in GPS-IPW. The scree plot for the 2-hour forecast equation using the backward elimination method of predictor selection is shown in Figure 20. The new equation is shown below.

$$\hat{y} = \frac{1}{1 + \exp(-5.044 + 3.633x_1 + 1.357x_2)}$$

where

x_1 = Objective Lightning Probability

x_2 = Δ GPS-IPW (current – 5.5 hr)

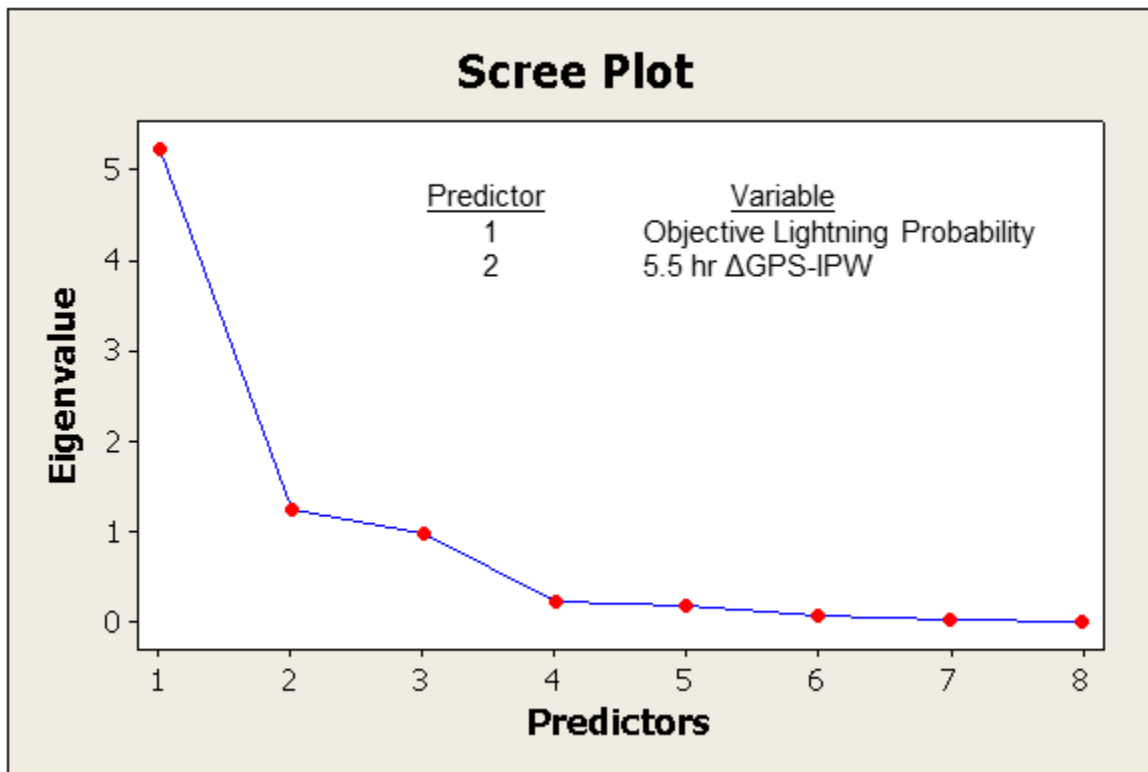


Figure 20. Scree plot for the predictors in the 2-hour forecast regression equation using the backward elimination method of predictor selection. The first two variables have eigenvalues > 1 and are likely useful in a forecast equation.

Two factors had eigenvalues greater than one in the 9-hour forecast regression equation using the backward elimination method for predictor selection. The two variables in the truncated regression equation were selected by examining the Figure 16 ROC curve and selecting the two predictors that had the highest AUC. For this case, the variables selected were the objective lightning probability and the 24 hour change in GPS-IPW. The scree plot for the 9-hour forecast equation using the backward elimination method of predictor selection is shown in Figure 21. The new equation is shown below.

$$\hat{y} = \frac{1}{1 + \exp(-5.133 + 4.027x_1 + 0.235x_2)}$$

where

x_1 = Objective Lightning Probability

x_2 = Δ GPS-IPW (current – 24.0 hr)

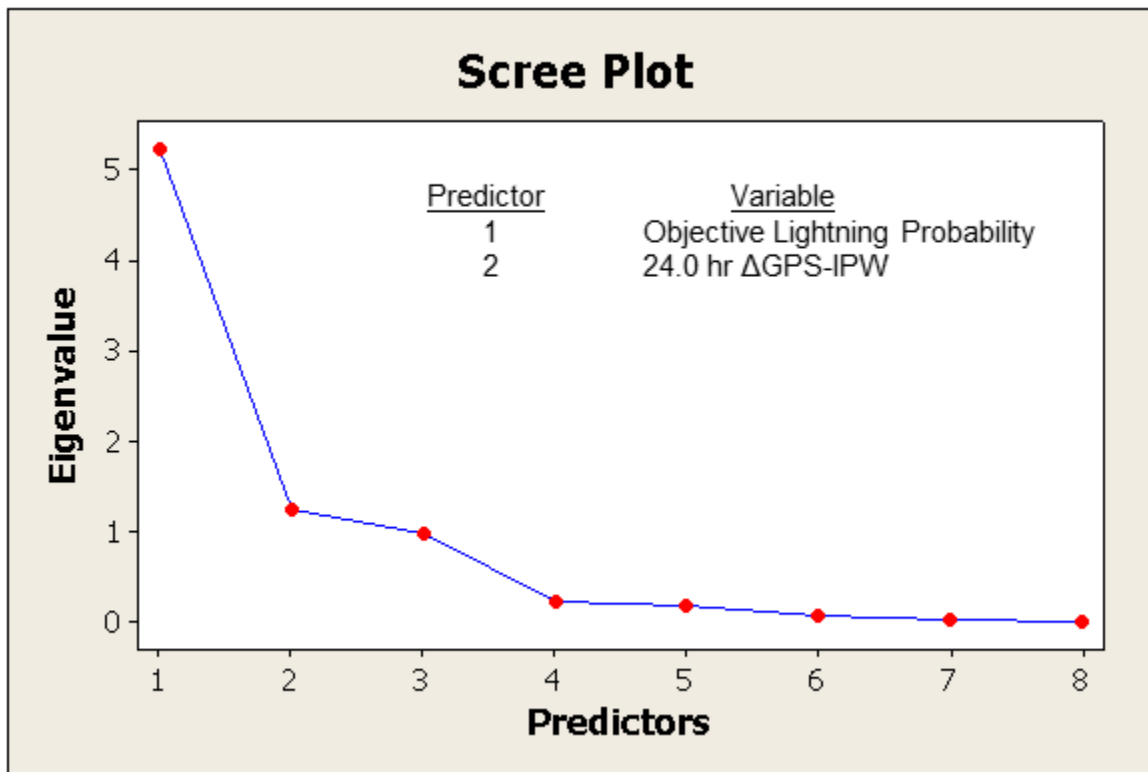


Figure 21. Scree plot for the predictors in the 9-hour forecast regression equation using the backward elimination method of predictor selection. The first two variables have eigenvalues > 1 and are likely useful in a forecast equation.

5. Equation Performance

The performance metrics for the equations perform differently depending on the lightning threshold index value chosen. To maximize the OUI, the lightning threshold index had to be reduced to 0.05, and even then the OUI was not as good as the model by Kehrler et al (2008) for either the 2-hour or 9-hour forecast model. Performance metrics for the 2-hour forecast using the forward selection and backward elimination methods of predictor selection and lightning index values are shown in Table 7. Performance metrics for the 9-hour forecast using the forward selection and backward elimination methods of predictor selection and lightning index values are shown in Table 8. For both the 2-hour and 9-hour forecasts, regardless of predictor selection method used, HR was acceptable for any index value above 0. However, POD dropped precipitously above an index value of 0.05. The FAR was unacceptably high regardless of index value. KSS and OUI were maximized at an index value of 0.05 for the 2-hour and 9-hour forecasts, but both scored below 53% and 42%, respectively. Bias was closest to 1 at an index value of 0.2. Below index values of 0.2, lightning was overforecast and lightning was underforecast above an index value of 0.2.

Table 7. Comparison of accuracy measurements, skill scores, and bias for the 2-hour forecast equations, using the backward elimination method, forward selection method, and the truncated forecast equations using the backward elimination method or forward selection method. The scores are shown for a range of lightning index threshold values.

Index	Selection Method	HR (%)	POD (%)	FAR (%)	KSS (%)	OUI (%)	Bias
0.00	Backward	4.2	100.0	95.8	0.0	34.0	32.91
	Forward	4.2	100.0	95.8	0.0	34.0	23.91
	2 predictor back	4.2	100.0	95.8	0.0	34.0	23.91
	4 predictor fwd	4.2	100.0	95.8	0.0	34.0	23.91
0.05	Backward	76.9	72.4	87.9	49.5	38.1	5.98
	Forward	76.2	76.7	87.7	52.8	41.3	6.24
	2 predictor back	74.9	74.7	88.5	49.7	39.2	6.49
	4 predictor fwd	74.0	78.4	88.4	52.3	41.9	6.77
0.1	Backward	87.8	42.9	84.5	32.7	18.3	2.77
	Forward	87.2	44.0	85.1	33.1	18.9	2.95
	2 predictor back	87.1	43.2	85.4	32.2	18.1	2.95
	4 predictor fwd	86.8	45.2	85.3	33.8	19.6	3.06
0.2	Backward	94.2	12.8	79.9	10.6	-3.4	0.64
	Forward	94.5	11.1	79.4	9.2	-4.6	0.54
	2 predictor back	94.6	7.7	83.0	6.0	-8.0	0.45
	4 predictor fwd	94.4	10.5	80.5	8.6	-5.3	0.54
0.3	Backward	95.4	3.1	81.7	2.5	-11.2	0.17
	Forward	95.5	3.4	74.5	3.0	-9.7	0.13
	2 predictor back	95.6	1.1	85.7	0.8	-13.4	0.08
	4 predictor fwd	95.6	2.6	77.5	2.2	-10.9	0.11
0.32	Backward	95.5	1.7	86.7	1.2	-13.2	0.13
	Forward	95.6	2.6	76.9	2.2	-10.8	0.11
	2 predictor back	95.6	0.9	85.7	0.6	-13.7	0.06
	4 predictor fwd	95.6	2.3	75.8	2.0	-10.8	0.09
0.4	Backward	95.6	0.3	94.1	0.1	-15.5	0.05
	Forward	95.7	1.1	81.0	0.9	-12.6	0.06
	2 predictor back	95.7	0.6	84.6	0.4	-13.7	0.04
	4 predictor fwd	95.7	0.0	100.0	-0.1	-16.7	0.03
0.5	Backward	95.7	0.0	N/A	-0.1	N/A	0.03
	Forward	95.7	0.0	N/A	-0.1	N/A	0.02
	2 predictor back	95.8	0.0	100.0	-0.1	-16.7	0.01
	4 predictor fwd	95.8	0.0	100.0	-0.1	-16.7	0.01

Table 8. Comparison of accuracy measurements, skill scores, and bias for the 9-hour forecast equations, using the backward elimination method, forward selection method, and the truncated forecast equations using the backward elimination method or forward selection method. The scores are shown for a range of lightning index threshold values.

Index	Selection Method	HR (%)	POD (%)	FAR (%)	KSS (%)	OUI (%)	Bias
0.00	Backward	4.2	100.0	95.8	0.0	34.0	23.90
	Forward	4.2	100.0	95.8	0.0	34.0	23.84
	2 predictor back	4.2	100.0	95.8	0.0	34.0	23.91
	3 predictor fwd	4.2	100.0	95.8	0.0	34.0	23.84
0.05	Backward	76.9	71.9	88.6	47.6	37.0	6.29
	Forward	76.2	74.8	88.3	50.2	39.4	6.36
	2 predictor back	74.9	74.7	88.5	49.7	39.2	6.49
	3 predictor fwd	74.0	74.2	89.2	47.3	38.0	6.89
0.1	Backward	87.8	35.8	85.8	26.3	12.4	2.53
	Forward	87.2	40.8	84.6	31.0	16.6	2.64
	2 predictor back	87.1	43.2	85.4	32.2	18.1	2.95
	3 predictor fwd	86.8	48.2	86.1	35.1	21.5	3.46
0.2	Backward	94.2	6.5	73.6	5.7	-7.0	0.25
	Forward	94.5	7.7	76.5	6.6	-6.7	0.33
	2 predictor back	94.6	7.7	83.0	6.0	-8.0	0.45
	3 predictor fwd	94.4	4.5	82.4	3.6	-10.3	0.26
0.3	Backward	95.4	0.6	60.0	0.5	-9.5	0.01
	Forward	95.5	0.6	86.7	0.4	-14.0	0.04
	2 predictor back	95.6	1.1	85.7	0.8	-13.4	0.08
	3 predictor fwd	95.6	0.0	N/A	0.0	N/A	0.00
0.32	Backward	95.5	0.6	60.0	0.5	-9.5	0.01
	Forward	95.6	0.6	81.8	0.5	-13.2	0.03
	2 predictor back	95.6	0.9	85.7	0.6	-13.7	0.06
	3 predictor fwd	95.6	0.0	N/A	0.0	N/A	0.00
0.4	Backward	95.6	0.0	N/A	0.0	N/A	0.00
	Forward	95.7	0.0	N/A	0.0	N/A	0.00
	2 predictor back	95.7	0.6	84.6	0.4	-13.7	0.04
	3 predictor fwd	95.7	0.0	N/A	0.0	N/A	0.00
0.5	Backward	95.7	0.0	N/A	0.0	N/A	0.00
	Forward	95.7	0.0	N/A	0.0	N/A	0.00
	2 predictor back	95.8	0.0	100.0	-0.1	-16.7	0.01
	3 predictor fwd	95.8	0.0	N/A	0.0	N/A	0.00

6. Summary and Conclusions

The AMU investigated the utility of using GPS-IPW as compared to the previous Objective Lightning Probability tool to predict the probability of lightning occurrence for two important operation products from 45 WS: 1) lightning advisories, and 2) major ground processing operations. The data sources used for this task included CGLSS, the GPS-IPW data from the GPS sensor at CCAFS, and the lightning probabilities from the Objective Lightning Probability tool. Since data from the GPS-IPW site were not available before 2000, the period of record (POR) was 2000-2011 for the warm-season months of May-October. Because many of the variables considered in previous studies (Mazany et al. 2000 and Kehrler et al. 2008) such as KI, TT, and upper level moisture variables, were already considered by the Objective Lightning Probability tool, only the objective lightning probability values along with the current GPS-IPW values and changes in GPS-IPW over half hour increments up to 24 hours were considered in this study. New models were built using binary (yes/no) logistic regression where the element to be forecast was the occurrence of lightning. Yes meant lightning occurred and no meant lightning did not occur within the specified time period and area of interest.

Using the proven methodology of multiple logistic regression to evaluate the binary predictand, the AMU evaluated a total of 50 candidate predictors to determine a subset of predictors that affected the predictand. The AMU used the forward selection method and the backwards elimination method to select the predictor. Of the 50 candidate predictor variables, 48 were a linear combination of the current GPS-IPW value and one of the half hour interval values between 0.5 and 24.0 hours. Therefore, there was some multicollinearity of the independent variables. The AMU performed a factor analysis to remove some of the independent variables and then they recalculated performance measures to investigate if reducing the number of factors in the model substantially impacted the performance metrics.

Although previous studies showed the GPS-IPW values to be promising in forecasting lightning, the results of this task did not find them to be value added over the current Objective Lightning Probability tool. This is likely because the level of noise in the objective lightning probability (which dominates the regression equations) is greater than the increase in predictive capability offered by the inclusion of the GPS-IPW data. The Objective Lightning Probability tool was designed to predict the probability of lightning for the day (from 0700 to 2400 EDT). The equations used in the Objective Lightning Probability tool were not designed for the temporal resolution of the 45 WS lightning advisories and major ground processing operations. As a result, this work demonstrated that inclusion of the GPS-IPW data into the objective lightning probability as a predictor in the equations did not improve model performance.

References

- Anderson, D., D. Sweeney, and T. Williams, 2012: *Statistics for Business and Economics, Revised, 11th Edition*. South-Western, Cengage Learning, 1088 pp.
- D'Arcangelo, D., 2000: Forecasting the Onset of Cloud-Ground Lightning Using Layered Vertically Integrated Liquid Water, M.S. Thesis, Pennsylvania State University, August 2000, 60 pp.
- Darlington, Richard H., cited 2012: Factor Analysis. [Available online at <http://www.psych.cornell.edu/darlington/factor.htm>.]
- Inoue, H. and T. Inoue, 2007: Characteristics of the Water-Vapor Field over the Kanto District Associated with Summer Thunderstorm Activities, *SOLA*, **3**, 101-104.
- Kehrer, K., B. Graf, and W. Roeder, 2008: Global Positioning System (GPS) Precipitable Water in Forecasting Lightning at Spaceport Canaveral, *Wea. Forecasting*, **23**, 219-232.
- Kohler, H. 1985: *Statistics for Business and Economics*: Scott, Foresman and Company, 770 pp.
- Lambert, W., 2007: Objective Lightning Probability Forecasting for Kennedy Space Center and Cape Canaveral Air Force Station, Phase II. NASA Contractor Report CR-2007-214732, Kennedy Space Center, FL, 57 pp. [Available from ENSCO, Inc., 1980 N. Atlantic Ave., Suite 830, Cocoa Beach, FL, 32931 and online at <http://science.ksc.nasa.gov/amu/final.html>.]
- Mason, S. and N. Graham, 1999: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Wea. Forecasting*, **14**, 713-725.
- Mazany, R., S. Businger, S. Gutman, and W. Roeder, 2002: Operational Multiple-Doppler Wind Retrieval Inferred from Long-Range Radial Velocity Measurements, *Wea. Forecasting*, **17**, 1034-1047.
- Suparta, W., J. Adnan, and M. Ali, 2011a: Detection of Lightning Activity Using GPS PWV Measurements. *Proceeding of the 2011 IEEE International Conference on Space Science and Communication (IconSpace)*, 12-13 July, 2011, Penang, Malaysia, IEEE, 115-120.
- Suparta, W., J. Adnan, M. Ali, 2011b: Monitoring the Association between GPS PWV and Lightning Activity during the 2009 Winter Monsoon over Bangi Malaysia. *Proceeding of the 2011 International Conference on Environment Science and Engineering (ICESE 2011)*, 28-30 Sept. 2011, Singapore, Singapore, IEEE, 101-106.
- Walker, J. and S. Maddan, 2009: *Statistics in Criminology and Criminal Justice*. Jones and Bartlett Publishers, 498 pp.
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 627 pp.

List of Acronyms

45 WS	45th Weather Squadron	HR	Hit Rate
45 SW	45th Space Wing	KI	K-Index
AMU	Applied Meteorology Unit	KSC	Kennedy Space Center
AUC	Area Under Curve	KSS	Kuipers Skill Score
CCAFS	Cape Canaveral Air Force Station	NM	Nautical Mile
CGLSS	Cloud-to-Ground Lightning Surveillance System	OUI	Operational Utility Index
EDT	Eastern Daylight Time	POD	Probability of Detection
ESRL	Earth System Research Laboratory	POR	Period of Record
FAR	False Alarm Ratio	ROC	Relative Operating Characteristics
GPS-IPW	Global Positioning System Integrated Precipitable Water	TT	Total Totals

NOTICE

Mention of a copyrighted, trademarked or proprietary product, service, or document does not constitute endorsement thereof by the author, 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) 14-12-2012		2. REPORT TYPE NASA/TM		3. DATES COVERED (From - To) March 2012 to December 2012	
4. TITLE AND SUBTITLE Using Global Positioning System Integrated Precipitable Water Vapor to Forecast Lightning on KSC/CCAFS				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Lisa L. Huddleston				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Applied Meteorology Unit Mail Code: GP-B Kennedy Space Center, FL 32899				8. PERFORMING ORGANIZATION REPORT NUMBER NASA/TM-2012-216314	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSORING/MONITOR'S ACRONYM(S)	
				11. SPONSORING/MONITORING REPORT NUMBER	
12. DISTRIBUTION/AVAILABILITY STATEMENT Unclassified, Unlimited					
13. SUPPLEMENTARY NOTES Electronic versions of all Applied Meteorology Unit final reports can be found at http://science.ksc.nasa.gov/amu/final.html					
14. ABSTRACT The AMU investigated the utility of using Global Positioning System Integrated Precipitable Water (GPS-IPW) as compared to the previous Objective Lightning Probability tool to predict the probability of lightning occurrence for lightning advisories, and major ground processing operations. New models were built using binary logistic regression where the element to be forecast was the occurrence of lightning. Yes meant lightning occurred and no meant lightning did not occur within the specified time period and area of interest. This work demonstrated that inclusion of the GPS-IPW data into the objective lightning probability as a predictor in the equations did not improve model performance. This is likely because the level of noise in the objective lightning probability is greater than the increase in predictive capability offered by the inclusion of the GPS-IPW data.					
15. SUBJECT TERMS Objective Lightning Probability, Global Positioning System Integrated Precipitable Water, Binary Logistic Regression, Lightning Advisories, Ground Processing Operations					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Lisa L. Huddleston
U	UU	UU	UU	47	19b. TELEPHONE NUMBER (include area code) (321) 861-4952