P1.16 A STATISTICAL PROCEDURE TO FORECAST THE DAILY AMOUNT OF WARM SEASON LIGHTNING IN SOUTH FLORIDA

Phillip E. Shafer*
Henry E. Fuelberg
Department of Meteorology
Florida State University
Tallahassee, Florida

1. INTRODUCTION

Over the last 30 years, cloud-to-ground (CG) lightning has ranked above both tornadoes and hurricanes in weather related fatalities across the United States (Curran et al. 1997). Aside from the loss of life, lightning damages trees, buildings, and utility lines, and is one of the leading causes of power outages and disruptions to communications. For these reasons, accurate forecasts of the timing and location of thunderstorms and associated CG lightning are of great interest to all persons concerned with protecting life and property.

Florida leads the nation in lightning related casualties, a majority of which occur during the warm season months of May through September, the climatological peak for lightning activity in Florida. Many studies examining lightning patterns across the contiguous United States have found that Florida receives more CG lightning strikes annually than any other region (e.g., Orville and Silver 1997, Orville et al. 2002). Thus, Florida has been deservedly labeled the "lightning capital" of the United States.

Fig. 1 shows the spatial distribution of CG lightning for Florida on a 2.5 x 2.5 km grid for all warm season days (May through September) during the 14-year period from 1989-2002 (Stroupe 2003). Several areas of enhanced flash density are noted, specifically near Tampa Bay and Fort Myers on the west coast, as well as Cape Canaveral and a region stretching from West Palm Beach southward to Ft. Lauderdale and Miami on the east coast. These regions of enhanced flash density are due to many complex factors that have been studied in great detail. These include irregularly shaped and protruding coastlines, and thermal circulations such as the sea breeze and lake/river breezes, which interact with the prevailing synoptic flow.

* *Corresponding author address:* Phillip E. Shafer, Florida State University, Dept. of Meteorology, Tallahassee, FL 32306-4520; e-mail: Phil.Shafer@noaa.gov.

During the warm season, absent of synoptic or tropical disturbances, the Atlantic and Gulf of Mexico sea breeze circulations act as the primary triggering mechanism for afternoon convection and lightning in Florida. If adequate moisture and instability are present, the degree of afternoon convective activity that occurs and its location are primarily governed by the strength and inland penetration of the sea breeze boundary, which has been shown in previous studies to be highly dependent on the magnitude and direction of the prevailing low-level flow (López and Holle 1987, Camp et al. 1998, Lericos et al. 2002).

It is evident from Fig. 1 that many heavily populated areas along the east and west coasts of Florida are vulnerable to intense lightning activity. Consequently, the risks for casualties, damage, and disruptions to power and communications attributable to lightning are inherently much greater in these areas. The eastern halves of Miami-Dade and Broward counties in South Florida are especially vulnerable, since over 3.9 million people reside in the metropolitan areas of Miami and Fort Lauderdale (U.S. Census Bureau 2004). Here, power disruptions are not only problematic to customers but can pose major problems for the power companies responsible for repairing outages. For example, a company such as Florida Power & Light Corporation (FP&L) must determine well ahead of time whether lightning is likely during the late afternoon and evening hours anywhere within their service areas. If a high lightning threat is perceived, extra crews must be retained after normal business hours to deal with potential problems. If this threat is misjudged, the company either will not be able to respond to outages effectively, or, conversely, resources could be wasted on a threat that does not occur. Clearly, an accurate forecast of the timing, location and intensity of afternoon lightning activity in heavily populated areas of South Florida would be of great benefit to a power company, as well as to the customers they serve.

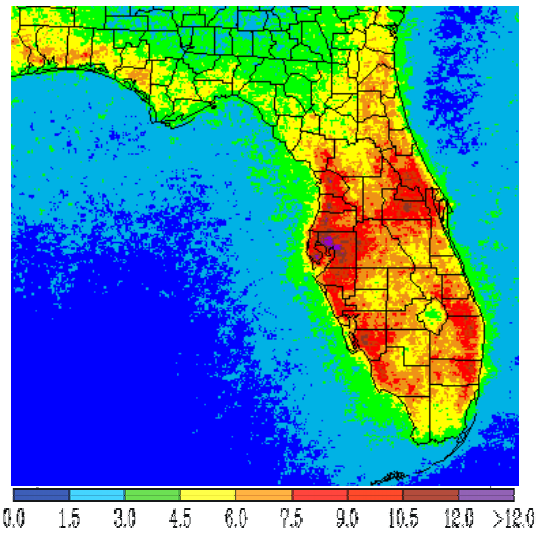


FIG. 1. Map of the spatial distribution of warm season CG lightning (flashes km^{-2} warm season^{-1}) for the state of Florida during a 14-year period from 1989-2002. Map obtained from http://bertha.met.fsu.edu/~jstroupe/flclimo.html.

The development of a lightning forecast procedure for these areas is a difficult problem. Despite the regular and predictable forcing produced by the sea breeze circulation, the nature of summertime convection and lightning over South Florida exhibits considerable spatial and temporal variability (López et al. 1984). This variability arises due to many complex localized and regional factors that govern the timing and preferred locations for convection and lightning on a given day. Even if one could pinpoint the exact locations that will experience convection on a particular day, it does not necessarily follow that those same areas will experience the most lightning, since this is governed by cloud microphysical processes that currently are unresolved by numerical models. For these reasons, one currently should not attempt to predict with any lead-time the exact number of lightning flashes that will occur within a small area during a specific time period. However, one can develop a prediction scheme that will provide useful guidance about the location and movement of the sea breeze and any associated convection, and, therefore, the degree of afternoon and evening lightning activity that can be expected, based on what has happened in the past under similar atmospheric conditions. In this regard, many studies have found statistical models to be useful for predicting thunderstorms and lightning. Some of the statistical methods that have been used include multiple linear regression, binary logistic regression, as well as Classification and Regression Trees (CART) (e.g., Livingston et al. 1996, Mazany et al. 2002, Burrows et al. 2004, Brenner 2004). These methods attempt to quantify the relationship between a set of predictors (i.e., sounding parameters or model data) and some outcome of interest such as thunderstorm probability at a particular location or spatial patterns of lightning frequencies (e.g., Neumann and Nicholson 1972; Reap 1994). Studies such as these have demonstrated the potential usefulness of statistical models for predicting thunderstorm and lightning activity during the warm season. However, most have focused on either a yes/no forecast of lightning or distinguishing between an active and an inactive day. Thus, no study has fully addressed the more complex issue of predicting the "amount" of lightning that will occur within a small domain such as the eastern halves of two counties.

The present study seeks to develop a system of statistical guidance equations describing the amount of warm season CG lightning activity that can be expected during the noon-midnight time period within two areas of South Florida serviced by FP&L, the eastern halves of Miami-Dade and Broward Counties in South Florida. The equations will give probabilities for different ranges of CG flash count, conditional on at least one flash occurring. The question of whether at least one flash will occur is a different problem that has been explored by Winarchick and Fuelberg (2005, *Conference on Meteorological Applications of Lightning Data*). The equations being derived are for the warm season months of May through September when the sea breeze is the dominant forcing mechanism for convection in Florida. Candidate predictors for the regression models include various wind, stability and moisture parameters calculated from 14 years of morning radiosonde data at Miami or West Palm Beach (1989-2002), as well as day number, morning lightning activity, and persistence. Lightning flash counts for each area are subdivided into quartile groups based on climatology, and separate logistic regression equations are derived for each domain giving the conditional probability of occurrence for different quartile ranges of flash count. Using probability thresholds for each of the equations, decision trees are constructed to determine the predicted lightning quartile for the day. Finally, the resulting models are evaluated and independently tested using k-fold cross-validation.

2. DATA

2.1 *Study Domain*

Statistical guidance equations were developed for two domains, the densely populated FP&L service areas of eastern Miami-Dade and Broward Counties in South Florida. A map of these areas is shown in Fig. 2, with the two regions outlined. In Broward County, the domain includes all land areas east of U.S. Route 27, north to the border with Palm Beach County and south to the border with Miami-Dade County. This domain includes the metropolitan areas of Hollywood and Fort Lauderdale. The Miami-Dade domain includes all land areas east of State Route 997 (a.k.a. Krome Avenue) from Homestead northward, and east of U.S. Route 1 from Homestead southward to the end of the peninsula. This region encompasses much of the Miami metropolitan area.

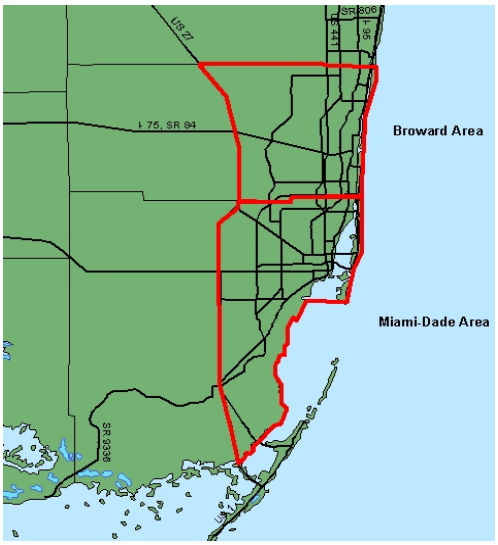


FIG. 2. Map of the two study regions. The Miami-Dade and Broward County domains are outlined.

2.2 *Lightning Data*

The study utilized CG lightning data from the National Lightning Detection Network (NLDN). This network, in operation since 1989, detects and records CG lightning flashes across the contiguous United States. The NLDN is owned and operated by Vaisala-Global Atmospherics Inc. (GAI), providing both real-time and historical data to electric utilities, the National Weather Service and other government, educational, and commercial users (Cummins et al. 1998). A complete description of sensors and methods of detection is given in Cummins et al. (1998).

The study period was the warm season months of May through September for the years 1989-2002. The location accuracy and detection efficiency of the NLDN has changed during this time due to system upgrades. Prior to 1994, detection efficiencies across the U.S. ranged between 65% and 85%, with location accuracies between 8 km and 16 km. A system upgrade in 1995 allowed a greater number of flashes to be detected, as well as improvements in location accuracy. Since the upgrade, the NLDN has a location accuracy of ~ 0.5 km over most of the U.S., and an estimated flash detection efficiency of 80-90% (Cummins et al. 1998). Detection efficiencies over Florida currently range from ~ 80% over most of the peninsula to only ~ 60% over the extreme southern part of the state. In this study, no corrections were applied to account for these spatial and temporal variations in detection efficiency or location accuracy. Thus, actual flash counts are underestimated.

Due to the improved detection efficiency of the NLDN, the same flash can be sensed multiple times, and non-CG discharges also can be detected (Cummins et al. 1998). Following the recommendation of Cummins et al. (1998), weak positive flashes with signal strengths less than 10 kA were removed from the data set. In addition, multiple flashes occurring during the same second and within 10 km of each other were assumed to be the same flash, and were combined into a single flash by retaining the first flash's time and location and adding the multiplicities.

The procedure to count the number of CG flashes in our areas of interest was straightforward. A rectangular array of 2.5 km by 2.5 km grid boxes was superimposed over the region, with the southwest corner at 25.0°N, 81.0°W, and the northeast corner at 26.5°N, 79.5°W. This created a 61 x 67 array of boxes that encompassed all of Broward and Miami-Dade Counties, including areas just west and offshore. Each flash was referenced by latitude/longitude coordinates which were converted to (i, j) coordinates to find the location of the flash within the array. The total flash count for each grid box was calculated by summing the number of hits in that box over the period of interest, noon-midnight local time (1600-0359 UTC). To obtain the total noon-midnight count within each domain, flash totals for the hours of 1600 UTC through 0359 UTC were accumulated for all grid boxes lying within the areas outlined in Fig. 2. A morning flash

count also was calculated for each domain as a potential predictor of afternoon lightning. This was obtained by summing all flashes between 0600 and 1159 local time (1000-1559 UTC).

Figure 3a shows the hourly distribution of flash count for the Miami-Dade domain for all warm season days with lightning data available (2097) during the 14-year period. The hourly distribution for Broward County (not shown) is very similar. A diurnal peak in lightning activity occurs between 2 PM and 3 PM local time, with a rapid decrease after 7 PM. Similar diurnal variations have been observed in previous studies (e.g., Neumann and Nicholson 1972; Maier et al. 1984; Reap and MacGorman 1989; Reap 1994; Livingston et al. 1996; Lericos et al. 2002; Mazany et al. 2002). The noon-midnight period considered in this study captures most of the daily activity in each region, accounting for 89% of the daily total in the Miami-Dade domain and 92% of the daily total in the Broward region. Although early afternoon is the most active period for lightning, the evening hours were included in the forecast period so that FP&L officials can ensure that sufficient manpower will be available to repair outages after normal business hours.

The frequency distribution of flash count for all available warm season days during the 14-year period is shown in Fig. 3b for the Miami-Dade domain (the graph for Broward County is very similar). The distribution clearly is right-skewed, with the greatest number of days having either no activity or between 1 and 50 CG flashes during the noon-midnight period. Very few days have greater than 300 flashes. A similar distribution was observed by Burrows et al. (2004) in their study of flash densities over Canada and the northern United States.

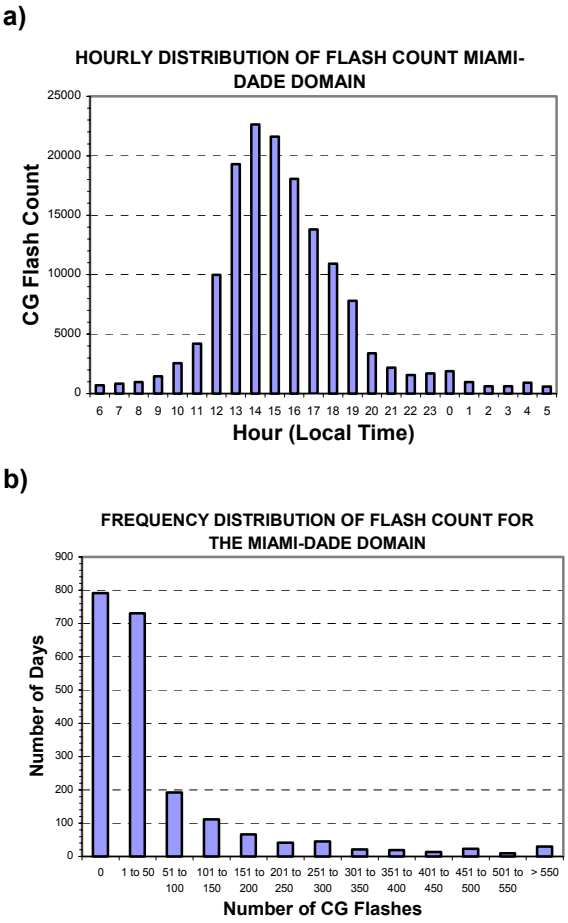


FIG. 3. a) Hourly distribution of CG flash count and b) frequency distribution of flash count during the noon-midnight period for the Miami-Dade domain for all warm season days with lightning data available during the 14 year period 1989-2002.

2.3 *Radiosonde Data*

Morning radiosonde data for Miami/West Palm Beach were used to calculate various wind, moisture, and stability parameters that serve as candidate predictors for the regression models. Data for the years 1989-1999 were obtained from the "Radiosonde Data of North America" CD-ROM prepared by the Forecast Systems Laboratory (FSL) and the National Climatic Data Center (NCDC) (FSL and NCDC 1999). Data for the remaining years 2000-2002 were obtained directly from FSL's "Radiosonde Database Access" Web site (http://raob.fsl.noaa.gov).

After 1977 and prior to 1995 the Miami (MFL) radiosonde site was located in West Palm Beach (PBI). López et al. (1984), Blanchard and López (1985), and Lericos et al. (2002) determined that basic features of the two soundings are very similar, with only minor differences in the boundary layer due to local phenomena. Thus, given the close proximity of both radiosonde sites (~ 100 km apart), it is assumed that the conditions at either site generally are representative of the conditions within the two study areas. Therefore, no adjustments were made to account for the difference in location of the two soundings.

Fifty-four parameters were calculated from the soundings, many of which have been found in previous studies to be useful predictors of thunderstorms and lightning during the warm season. These include variables that describe wind direction, wind speed, moisture in various layers, temperature, and stability. A complete list of the parameters is given in Table 1. For ease of

Table 1. Radiosonde-derived parameters and abbreviations.

Parameter Description	Abbreviation	Parameter Description	Abbreviation
Mean 1000-700 hPa wind direction	(MNDIR)	Surface wet bulb temperature	(SFCTWB)
Mean 1000-700 hPa wind speed	(MNSPD)	Precipitable water	(PW)
Mean 1000-700 hPa cross-shore component	(UPERP)	Mean surface-825 hPa mixing ratio	(WSFC-825)
Mean 1000-700 hPa along-shore component	(VPARLL)	K-index	(KI)
sin (mean 1000-700 hPa wind direction)	sin(MNDIR)	Vertical totals	(VT)
sin (wind direction at 950 hPa)	sin(DIR950)	Cross totals	(CT)
sin (wind direction at 700 hPa)	sin(DIR700)	Total totals index	(TT)
Mean sfc-850 hPa u component	(USFC-850)	Severe WEAther Threat index	(SWEAT)
Mean sfc-850 hPa v component	(VSFC-850)	Convective Available Potential Energy	(CAPE)
Mean sfc-850 hPa wind speed	(SPDSFC-850)	Modified CAPE	(MCAPE)
Mean 850-700 hPa u component	(U850-700)	Lifted Index	(LI)
Mean 850-700 hPa v component	(V850-700)	Modified Lifted Index	(MLI)
Mean 850-700 hPa wind speed	(SPD850-700)	Showalter Stability Index	(SSI)
Mean 700-500 hPa u component	(U700-500)	Temperature at 900 hPa	(T900)
Mean 700-500 hPa v component	(V700-500)	Sfc-1000hPa temperature difference	(DTSFC-1000)
Mean 700-500 hPa wind speed	(SPD700-500)	Sfc-850 hPa temperature difference	(DTSFC-850)
Wind speed at 900 hPa	(SPD900)	850-700 hPa temperature difference	(DT850-700)
Surface dew point	(SFCDWP)	850-500 hPa temperature difference	(DT850-500)
Modified surface dew point	(MSFCDWP)	500-300 hPa temperature difference	(DT500-300)
Mean sfc-900 hPa relative humidity	(RHSFC-900)	1000 hPa height	(Z1000)
Mean 800-600 hPa relative humidity	(RH800-600)	850 hPa height	(Z850)
Mean 700-500 hPa relative humidity	(RH700-500)	Equilibrium level pressure	(EL)
Mean 600-400 hPa relative humidity	(RH600-400)	Freezing level height	(FRZLVL)
Mean sfc-500 hPa relative humidity	(RHSFC-500)	Wet bulb zero height	(WBZLVL)
Mean 500-300 hPa relative humidity	(RH500-300)	1000-500 hPa thickness	(THICK)
Mean 800-600 hPa dew point depression	(DD800-600)	Convective temperature	(TCON)
Mean sfc-500 hPa dew point depression	(DDSFC-500)	Temperature at the Equilibrium level	(T@EL)

calculation, the raw sounding data were interpolated to 25 hPa levels, with the first level being the surface and then decreasing by 25 hPa increments from 1000 hPa to 100 hPa.

As discussed previously, many lightning studies in Florida have found that the magnitude and direction of the prevailing flow are important factors determining the degree of lightning activity in a particular location, since this flow exerts a major influence on the strength and inland penetration of the sea breeze (e.g., Gentry and Moore 1954; López and Holle 1987; Reap 1994; Lericos et al. 2002). For our study areas, the Atlantic Coast sea breeze has a significant influence on the amount of afternoon lightning that will occur, and its location and strength depend greatly on whether the low-level flow is offshore or onshore. To include this effect in the prediction scheme, various wind direction and wind speed predictors were calculated from the morning soundings. These include mean wind direction (MNDIR) and speed (MNSPD) in the 1000-700 hPa layer, the mean 1000-700 hPa cross-shore (UPERP) and along-shore (VPARLL) wind components, as well as the mean speed and *u* and *v* wind components in various other layers (see Table 1). The 1000-700 hPa layer was chosen because previous studies have determined it to represent best the combined motion of the sea breeze front and thunderstorms over the Florida peninsula during the warm season (López and Holle 1987; Camp et al. 1998).

The wind parameters listed above are vector-averaged quantities, calculated by computing the *u* and *v* components at each 25 hPa level, finding the mean of each component (U and V) through the layer of interest, and then computing the inverse tangent of (U/ V) to obtain a mean wind

direction (radians) in that layer. A mean layer wind speed was obtained by taking $(U^2 + V^2)^{1/2}$. The wind components perpendicular and parallel to the coastline, UPERP and VPARLL, were calculated by assuming an average coastline orientation of 15° clockwise from the north-south direction.

The equations being derived are for situations when the sea breeze is the dominant forcing mechanism for convection, and are not meant for days when large-scale forcing leads to thunderstorms. Therefore, some effort must be made to remove these synoptically influenced days from the analysis before the model building process begins. This was done by removing any day whose 1000-700 hPa layer mean wind speed was greater than three standard deviations from the climatological mean value (any day $\geq$ 25.53 knots). A total of 28 days was removed, leaving 2019 days available. Eleven of the 28 excluded days contained tropical storms or hurricanes in the vicinity of South Florida, and 13 days had some form of large-scale synoptic disturbance in the area. This was not an exhaustive effort, and it does not guarantee that every synoptically disturbed day was removed.

Atmospheric stability and moisture content also influence thunderstorm activity. As found by Fuelberg and Biggar (1994) for the Florida panhandle, adequate moisture and instability are prerequisites for thunderstorm formation during the warm season. Since this study only concerns days having at least one lightning flash in the two domains, adequate moisture and instability are assumed to be present. Nonetheless, there still may be relationships between the moisture/ stability parameters and the degree of convection and lightning that occurs (e.g., López et al. 1984; Reap 1994; Brenner 2004). To investigate, several parameters describing moisture and stability were calculated from the soundings (Table 1). The moisture parameters include surface dew point, dew point depression in various layers, mean relative humidity in various layers, and precipitable water. The mean relative humidity parameters are pressure-weighted averages based on the 25 hPa data within the layer of interest. A 12-h change in PW also was calculated (e.g., Mazany et al. 2002) by computing the difference in PW values between the 1200 UTC and previous 0000 UTC soundings.

A variety of stability indices were calculated (Table 1). K index, vertical totals, cross totals, total totals, SWEAT, and SSI were obtained from standard formulas as given in the *Glossary of Meteorology* (2000). CAPE and lifted index were calculated by lifting a standard surface parcel to its lifting condensation level (LCL) and then ascending the moist adiabat to determine the level of free convection (LFC) and equilibrium level (EL). Modified stability parameters also were calculated to better reflect afternoon conditions (Table 1), including a modified CAPE (MCAPE) and a modified lifted index (MLI). These were calculated in a manner similar to CAPE and lifted index, except they were based on a modified surface parcel heated to the convective temperature (TCON). TCON was obtained by descending dry adiabatically from the convective condensation level (CCL) to the surface, with the CCL based on a mean mixing ratio in the lowest 100 hPa.

Several additional parameters were calculated (Table 1) which have been found by previous studies to be useful predictors of afternoon convection in Florida during the warm season. These include temperature differences in various layers, 1000-500 hPa thickness, height of the freezing level and wet bulb zero, as well as 1000 hPa and 850 hPa heights.

Since this study is concerned with forecasting the amount of lightning that can be expected conditional on at least one flash occurring, only lightning days were retained in the data set. A day qualified as a lightning day for either domain if at least one flash occurred somewhere within those areas during the noon-midnight time period, regardless of what occurred elsewhere (e.g., on the other side of Krome Avenue/ U.S. Route 27). Of the 2019 days for which both lightning and sounding data were available, 1223 days had at least one flash in eastern Miami-Dade County and 1189 had at least one flash in eastern Broward County. These days constitute the final data sets used in the statistical analysis described in the following sections.

2.4 *Statistics Software*

Two statistical software packages were used to perform the regression analysis. Most of the exploratory work was done using S-PLUS, version 6.1 for Windows, distributed by Insightful Corporation. The final model development and testing were performed using the Statistical Package for the Social Sciences (SPSS), version 11.5 for Windows, distributed by SPSS, Inc. Both are powerful, state-of-the-art software packages with a wide range of capabilities.

3. MODEL DEVELOPMENT AND TESTING

3.1 *Modeling Options*

A major decision was to determine the form of the predictand, i.e., whether to estimate the actual flash count for the noon-midnight period or to transform the counts into discrete categories and predict a range of flash count. The final choice depended on which approach yielded the best predictions.

Several statistical techniques were attempted initially but were found to yield undesirable results. The first efforts were aimed at predicting an actual flash count for the noon-midnight period using multiple linear regression and Poisson log-linear regression (Wilks 1995 gives a complete description of these methods). Predictors were chosen for inclusion in the model through forward stepwise selection of variables. This procedure began by first selecting the independent variable that resulted in the largest reduction of the residual sum of squares (RSS). It then selected the next variable, which, together with the first, further reduced the RSS by the greatest amount (i.e., had the highest partial correlation with flash count). After each step, the algorithm performed a backward check to determine if removing any previously selected variable did not significantly increase the RSS. Any variables meeting this criterion were removed from the model, and all others were retained. This process continued until the RSS could not be changed by a significant amount, or until no other variables remained.

Several problems were encountered with the resulting regression equations. In short, several collinear (redundant) predictors were selected for inclusion in the equations, and there also was a high degree of unexplained variance in flash count, especially for larger counts. Wilks (1995) warns that stepwise selection procedures can sometimes result in highly correlated variables being included, which can have undesirable effects on the estimates of the regression coefficients and on model performance. In an attempt to correct these issues, several different transformations of the response variable were attempted in order to stabilize the variance, and non-linear and interaction effects were included as additional candidate predictors. However, the resulting equations still displayed a high degree of unexplained variance and collinearity among the selected predictor variables. Classification and Regression Trees (CART) also was attempted as an alternative to linear methods; however, the trees had a tendency to over-fit the training data and did not perform well under cross-validation. At this point, it was clear that attempting to estimate an actual flash count with any degree of confidence would be impossible given the large amount of variance that was present in the data set. In addition, some other method of variable selection was needed such that only the most important non-redundant variables would be included in the equations.

3.2 *Development and Testing of Final Model*

a. Binary Logistic Regression

Given the many problems encountered while attempting to predict an actual flash count, it became apparent that predicting a range of counts was the more feasible option. Therefore, the flash counts were grouped into four quartiles of activity based on climatology (Fig. 4), and the four quartiles were used as the predictand. Rather than developing one model to forecast the quartile, it was found that better results would be achieved if there were separate equations to distinguish the lowest quartile of activity (Q1) from all other days, the highest quartile (Q4) from all other days, and an equation to differentiate the upper two quartiles (Q3, Q4) from the lower two (Q1, Q2). The three equations would give conditional probabilities for each outcome, and thresholds could be determined for each equation to forecast the most likely quartile.

For situations when the outcome is binary or dichotomous (i.e., 1 for "yes" or 0 for "no") the most often used technique is "binary logistic regression" (BLR) (Hosmer and Lemeshow 1989). Let π denote the probability of a success for some outcome of interest (e.g., the probability of a Q4 event versus all other days). BLR relates this probability to a linear combination of predictor variables, X_K by the following equation:

$$\ln [\pi/(1- \pi)] = g(X_K) = b_0 + b_1X_1 + \ldots + b_KX_K \qquad (1)$$

where ln is the natural logarithm. The term on the left side of (1) is the "logit link function," which may be continuous and can range from $-\infty$ to $+\infty$ depending on the range of X_K (Hosmer and Lemeshow 1989). The probability of a success is then given by:

$$\pi = \exp(g(X_K)) / [1 + \exp(g(X_K))], \qquad (2)$$

and the probability of a failure (i.e., the probability of not observing a Q4) is just $1-\pi$.

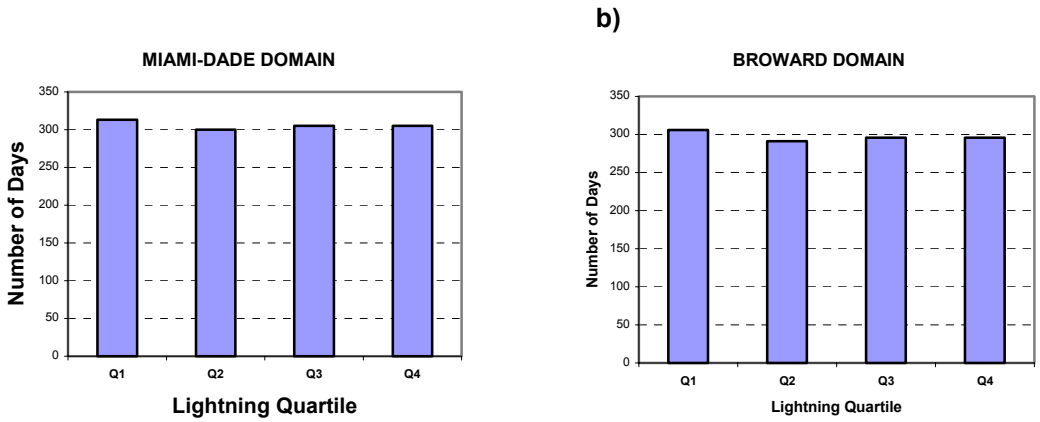


FIG. 4. The four quartiles of CG flash count for the (a) Miami-Dade and (b) Broward County domains. For the Miami-Dade domain, Q1: 1-7 flashes, Q2: 8-39, Q3: 40-125, Q4: > 125 flashes, and for the Broward County domain, Q1: 1-9 flashes, Q2: 10-47, Q3: 48-143, Q4: > 143 flashes.

BLR generally has less stringent assumptions than linear regression. Unlike multiple linear regression, logistic regression does not assume a linear relationship between the independent variables and the dependent (binary) outcome. Rather, the logit function on the left side of (1) is assumed to be linear in its parameters, although explicit interaction and power terms can be added as additional variables on the right side. In addition, the form of (2) guarantees that BLR will always produce probability estimates that are bounded between zero and one (Hosmer and Lemeshow 1989).

To create the binary response variables for the three models, three separate binomial indicators were assigned to each day in the Miami-Dade and Broward County data sets according to that day's lightning quartile value. For example, a "1" was assigned to Q1 days and a "0" otherwise, a "1" for Q4 days and a "0" otherwise, and a "1" for days in the upper two quartiles (Q3 or Q4) and a "0" otherwise. This resulted in three predictands for both domains, each having a climatological frequency of occurrence that is approximately 25%, 25%, and 50%, respectively.

b. Principal Component Analysis

To address the collinearity problem discussed previously, a principal component analysis (PCA) (Wilks 1995) was performed on all of the potential sounding predictors (Table 1). PCA is a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components (PCs). Wilks (1995) gives a complete description of this procedure. In this study, the PCs were used as a classification method to cluster all of the highly correlated predictors into groups for purposes of physical interpretation. This was done by grouping together all sounding parameters that had the highest weights (or "loadings") on each component. A total of eleven PCs were extracted during this process using the SPSS software. The first three PCs consisted of parameters describing moisture, wind direction and wind speed, respectively. The next three PCs were consolidated into one group representing unmodified and modified stability-related parameters. The last five components were combined into a set of miscellaneous parameters that were not directly assigned to any other component, although many of them are interrelated.

Table 2 lists two-tailed Pearson correlation coefficients between each predictor and the three binomial indicators of lightning activity for both domains. Asterisks indicate whether the correlations are significantly different from zero at the 0.01 (**) and 0.05 (*) significance levels. The correlations generally are low (all are below 0.3), suggesting that no single parameter from the morning sounding is a good indicator of the amount of afternoon lightning that will occur. Since many of the correlations in each group differ by only the hundredths place, the correlations themselves were not useful for determining which parameters out of each group to retain for the regression analysis. Instead, one parameter was chosen from each of the five groups that was

Table 2. Two-tailed Pearson correlation coefficients between each predictor and the three binomial indicators of lightning activity for both domains. Asterisks indicate whether the correlations are significantly different from zero at the 0.01 (**) and 0.05 (*) significance levels.

	MIAMI-DADE BINOMIAL INDICATORS			BROWARD BINOMIAL INDICATORS		
	Q1 yes/no (< 8 flashes)	Q3 or Q4 yes/no (≥ 40 flashes)	Q4 yes/no (> 125 flashes)	Q1 yes/no (< 10 flashes)	Q3 or Q4 yes/no (≥ 48 flashes)	Q4 yes/no (> 143 flashes)
GROUP 1: MOISTURE (COMP 1)						
DD800-600	0.120**	0.087**	-0.055	0.052	-0.044	-0.046
DDSFC-500	0.116**	-0.086**	-0.050	0.046	-0.036	-0.035
KI	-0.134**	0.120**	0.069*	-0.053	0.071*	0.071*
PW	-0.105**	0.065*	0.018	-0.034	0.011	0.007
RH500-300	-0.058*	0.021	0.033	-0.010	0.011	0.041
RH600-400	-0.083**	0.048	0.044	-0.033	0.014	0.031
RH700-500	-0.094**	0.056	0.039	-0.038	0.022	0.020
RH800-600	-0.086**	0.050	0.023	-0.015	0.012	0.015
RHSFC-500	-0.083**	0.047	0.017	-0.012	0.003	0.001
WBZLVL	-0.075**	0.058*	0.011	-0.042	0.006	-0.015
GROUP 2: WIND DIRECTION (COMP 2)						
MNDIR	-0.163**	0.222**	0.213**	-0.149**	0.217**	0.194**
sin(DIR700)	0.214**	-0.256**	-0.227**	0.176**	-0.234**	-0.205**
sin(DIR950)	0.208**	-0.251**	-0.217**	0.169**	-0.229**	-0.202**
sin(MNDIR)	0.209**	-0.282**	-0.254**	0.203**	-0.273**	-0.232**
U700-500	-0.167**	0.196**	0.182**	-0.166**	0.190**	0.171**
U850-700	-0.194**	0.239**	0.206**	-0.183**	0.228**	0.190**
UPERP	-0.207**	0.275**	0.236**	-0.182**	0.230**	0.206**
USFC-850	-0.205**	0.272**	0.219**	-0.185**	0.230**	0.198**
GROUP 3: WIND SPEED (COMP 3)						
MNSPD	0.101**	-0.178**	-0.151**	0.114**	-0.123**	-0.147**
SPD700-500	0.049	-0.084**	-0.040	0.039	-0.061*	-0.030
SPD850-700	0.083**	-0.145**	-0.106**	0.088**	-0.097**	-0.108**
SPD900	0.094**	-0.173**	-0.171**	0.127**	-0.127**	-0.166**
SPDSFC-850	0.102**	-0.177**	-0.161**	0.111**	-0.126**	-0.160**

Table 2 (continued).

	MIAMI-DADE BINOMIAL INDICATORS			BROWARD BINOMIAL INDICATORS		
	Q1 yes/no (< 8 flashes)	Q3 or Q4 yes/no (≥ 40 flashes)	Q4 yes/no (> 125 flashes)	Q1 yes/no (< 10 flashes)	Q3 or Q4 yes/no (≥ 48 flashes)	Q4 yes/no (> 143 flashes)
GROUP 4: STABILITY (COMPONENTS 4-6)						
CAPE	0.002	-0.008	-0.037	0.006	-0.017	-0.015
CT	-0.090**	0.103**	0.035	-0.052	0.051	0.054
DT850-700	0.062*	-0.080**	-0.124**	0.048	-0.098**	-0.098**
DTSFC-1000	-0.051	0.108**	0.116**	-0.126**	0.134**	0.117**
DTSFC-850	-0.096**	0.122**	0.121**	-0.124**	0.128**	0.109**
EL	0.032	-0.019	0.003	-0.005	-0.011	-0.013
FRZLVL	0.019	-0.025	-0.051	0.037	-0.040	-0.065*
LI	0.017	-0.014	-0.001	0.027	-0.024	-0.021
MCAPE	-0.060*	0.103**	0.089**	-0.122**	0.099**	0.086**
MLI	0.074**	-0.109**	-0.121**	0.125**	-0.124**	-0.124**
SFCDWP	-0.036	0.023	-0.025	-0.041	0.007	-0.011
SFCTWB	-0.024	0.001	-0.036	-0.018	-0.016	-0.019
SSI	0.102**	-0.129**	-0.056*	0.063*	-0.058*	-0.053
SWEAT	-0.048	0.020	-0.052	0.018	-0.045	-0.040
T@EL	0.012	0.024	0.061*	-0.019	0.026	0.031
TT	-0.108**	0.129**	0.082**	-0.089**	0.097**	0.106**
VT	-0.055	0.075**	0.107**	-0.087**	0.106**	0.118**
GROUP 5: MISCELLANEOUS (COMPONENTS 7-11)						
DT500-300	-0.071*	0.060*	0.091**	-0.057	0.061*	0.083**
DT850-500	0.011	-0.016	-0.017	0.057*	-0.038	-0.050
MSFCDWP	-0.038	0.047	0.023	-0.063*	0.023	-0.009
RHSFC-900	0.027	-0.042	-0.059*	0.040	-0.053	-0.088**
T900	-0.139**	0.161**	0.124**	-0.139**	0.106**	0.119**
TCON	-0.090**	0.088**	0.082**	-0.105**	0.085**	0.129**
THICK	-0.061*	0.032	-0.016	-0.047	-0.010	-0.012
V700-500	0.050	-0.093**	-0.119**	0.016	-0.018	-0.058*
V850-700	0.032	-0.082**	-0.088**	-0.023	0.019	-0.028
VPARLL	-0.047	0.021	-0.004	-0.092**	0.104**	0.043
VSFC-850	0.009	-0.052	-0.065*	-0.037	0.040	-0.016
WSFC-825	-0.044	0.049	0.027	-0.065*	0.025	-0.008
Z1000	0.088**	-0.062*	-0.038	0.021	-0.038	0.007
Z850	0.059*	-0.030	-0.016	-0.009	-0.017	0.028

thought to have the most physical relevance to afternoon convection in South Florida, while still possessing as high a correlation as possible with the three binary predictands. From the wind direction and wind speed groups, UPERP and MNSPD were chosen because they have the greatest influence on the strength and inland penetration of the sea breeze front (e.g., López and Holle 1987; Camp et al. 1998). From the group of moisture-related parameters, K index was chosen since it is one of the most widely used indices for assessing thunderstorm potential during the warm season (e.g., Reap 1994; Livingston et al. 1996; Mazany et al. 2002). MLI was selected from the group of stability-related parameters because it is representative of afternoon conditions. Finally, the temperature at 900 hPa (T900) was chosen from the group of miscellaneous parameters because it has a higher correlation with lightning activity than all others in this group (and all three are significant at the 0.01 level). Since 900 hPa is near the top of the nocturnal inversion, it is thought that T900 may be indicative of the degree of afternoon heating that will occur. Neumann and Nicholson (1972) found T900 to be a good predictor of afternoon thunderstorms at the Kennedy Space Center.

It should be noted that the correlations in Table 2 indicate the degree of "linear" association between the predictors and lightning activity. Thus, these correlations do not recognize the strength of any non-linear (or "curvature") relationships that may exist (Wilks 1995). To incorporate possible curvature effects between each of the five physical variables and lightning activity, power terms up to the fourth degree were submitted as additional candidate predictors (e.g., Neumann and Nicholson 1972; Reap 1994). To avoid collinearity problems among the power terms, the five physical variables first were converted to z-scores (i.e., "standardized anomalies" formed by subtracting the sample mean and dividing by the sample standard deviation) before raising them to a power.

c. Climatology and Persistence

Climatology was incorporated into the prediction equations by computing climatological frequencies for each of the three binary response variables. These frequencies were included as additional candidate predictors, which has the effect of smoothing any discontinuities from one month to the next and avoiding the need for separate prediction equations for each month. Six persistence variables also were created as candidate predictors for the equations. These include the previous day's actual flash count (PDCNT), a yes/no indicator of at least 1 flash the previous day (PDYN), the previous day's quartile of activity (PDQRT), and the previous day's yes/no indicator for a Q1 (PDQ1YN), the upper two quartiles (PDQ3Q4), and Q4 (PDQ4YN).

Table 3 shows a 4 x 4 contingency table for the number of observed days in each lightning quartile versus the previous day's quartile of activity for all cases when there was at least 1 flash the previous day for the Miami-Dade domain. It is clear that persistence is a powerful predictor of the degree of lightning activity during the warm season in South Florida. Thus, it should be included as a candidate predictor in the equations. Overall, persistence forecasts the correct quartile ~ 34% of the time and is correct to within one quartile of the observed ~ 73% of the time. To put these statistics into perspective, the most naive forecast would be the climatological average for each of the quartiles, in which case one would forecast the correct quartile by chance ~ 25% of the time and would be correct to within one quartile of the observed by chance ~ 62.5% of the time (if the random choice is constrained to being unbiased). Since persistence typically produces a more accurate forecast than does climatology, persistence is used as the primary standard of reference for assessing the overall skill of the prediction equations derived in this study.

Table 4 lists the final set of candidate predictors that were used to derive the prediction equations for each domain. There were twenty physical variables (including powers up to the fourth degree), three climatic predictors, two indicators of morning activity, and the six persistence variables, for a total of 31 potential predictors. All days with a missing value for any predictor were removed from the data set, leaving 1209 days available for the Miami-Dade domain and 1177 days for the Broward domain.

d. Model development and testing

Three separate logistic regression equations were derived for both the Miami-Dade and Broward County domains (total of six equations), one giving the conditional probability of a Q1 lightning event, one for the probability of an event in the upper two quartiles (Q3/Q4), and the last giving the probability of a Q4 event. The logistic regression algorithm in SPSS was used to build the equations and screen the variables (Table 4) for selection into each model.

Table 3. 4 x 4 contingency table for the number of observed days in each lightning quartile versus the previous day's quartile for all cases with at least one flash the previous day.

Miami-Dade Persistence

	PREVIOUS DAY QUARTILE						
	Q1	Q2	Q3	Q4	Total	% Correct	% Within 1Q
OBSERVED							
Q1	**63**	62	44	37	206	31	61
Q2	54	**66**	56	45	221	30	80
Q3	51	54	**72**	67	244	30	79
Q4	36	40	77	**117**	270	43	72
Total	204	223	248	266	941	**34**	**73**

Table 4. List of final candidate predictors for the regression models. The asterisk denotes all first-order terms that are standardized quantities (z-scores) based on the sample mean and standard deviation of each variable for each domain. The higher order terms are products of z-scores.

Physical variables:

UPERP*, UPERP ^{2}, UPERP ^{3}, UPERP ^{4}
MNSPD*, MNSPD ^{2}, MNSPD ^{3}, MNSPD ^{4}
KI*, KI ^{2}, KI ^{3}, KI ^{4}
MLI*, MLI ^{2}, MLI ^{3}, MLI ^{4}
T900*, T900 ^{2}, T900 ^{3}, T900 ^{4}

Climatology (as a function of day number):

Frequency of a Q1	(CLIQ1)
Frequency of upper two quartiles	(CLIQ3Q4)
Frequency of a Q4	(CLIQ4)

Persistence/morning activity:

Previous day noon-midnight flash count	(PDCNT)
Previous day yes/no indicator of at least 1 flash	(PDYN)
Previous day lightning quartile	(PDQRT, 0 if no activity or 1-4)
Yes/no indicator for Q1 the previous day	(PDQ1YN)
Yes/no indicator for Q3 or Q4 the previous day	(PDQ3Q4)
Yes/no indicator for Q4 the previous day	(PDQ4YN)
Morning Flash count 0600-1159 local time	(MORNCNT)
Morning yes/no indicator	(MORNYN)

To derive each equation, a procedure combining forward stepwise screening and cross-validation was used to select the best combination of variables that is most likely to generalize to independent data. In short, the process began by randomly dividing the working data set into two parts. One set, containing 75% of the cases, was used as a "learning" sample for screening the predictors for selection into each of the models. The remaining 25% of the cases were reserved as an "evaluation" sample to test the model each time a predictor was added or removed during the stepwise selection process. The screening procedure in SPSS uses "forward conditional" stepwise selection, with a test for backward elimination. This procedure is similar to that described previously for the multiple linear regression, except now the selection criterion was the significance (p-value) of the variable's log-likelihood (LL) chi-square statistic. This procedure is described in greater detail in Wilks (1995). The stepwise selection procedure generated a sequence of candidate models, one for each time a predictor was either added to the model or removed at a later step. At each step in the sequence, SPSS produced a 2 x 2 contingency table showing the number of observed "1s" and "0s" versus the number predicted, one table for the 75% learning sample and another showing test results for the 25% evaluation sample. The predictors comprising the model at the step with the highest percentage of correctly classified days (i.e., the highest "hit rate") for the 25% evaluation sample were noted.

The above steps were repeated an additional four times, each time using a different random number seed to divide the training data into 75% (learning) and 25% (evaluation) samples. Each time, the predictors that were in the model at the step with the highest hit rate for the independent (evaluation) cases were noted. The rationale for this procedure was to identify only those predictors most likely to generalize to independent data and not "over-fit" the learning sample. The reason for repeating the random sampling multiple times was to guard against the possibility that the evaluation results depended heavily on how the division was made, or in other words, which cases ended up in the learning set and which were in the evaluation set. The list of "best" predictors identified during this process were re-entered for stepwise screening on the entire data set to obtain final logistic regression equations for each of the three outcome responses (prob(Q1), prob(Q3/Q4), and prob(Q4)). Out of this sequence of candidate models, the step with the highest overall percentage of correctly classified days (highest hit rate) was chosen as the final prediction equation.

After final equations were obtained for the three outcomes, a decision tree was constructed to determine the predicted lightning quartile using probability thresholds for the three equations. To produce an unbiased scheme, the thresholds were chosen so an equal number of training cases were partitioned to the left and right at each split of the decision tree. This guarantees that the scheme will not have a prediction bias toward any one quartile (i.e., a tendency to forecast a particular quartile more often than another). Further details about the decision tree and its implications are discussed in the results section.

To assess the degree to which the final prediction scheme generalizes to data that were not used to derive the equations, a k-fold cross-validation procedure was followed, whereby each warm season of data was individually set aside to be used as an independent data set for cross-validating the prediction scheme derived from the remaining thirteen warm seasons (used as the "training" data). In other words, the procedure described in the preceding paragraphs was repeated a total of 14 times, once for each warm season of data.

To obtain the three logistic regression equations that will be implemented operationally by FP&L (three for each domain, a total of six), one final stepwise screening was performed using the entire fourteen-year data set (i.e., all days, all years). The predictors now entered for screening were only those "best" predictors that were identified during the k-fold cross-validation procedure described above. Again, this assures that only the predictors that are most likely to generalize to independent data are selected. As before, a sequence of candidate models was produced, and the model with the highest percentage of correctly classified days among the sequence was chosen as the final model. Finally, decision trees were constructed to obtain the final operational prediction schemes for each domain.

4. RESULTS

4.1 *Final Prediction Equations*

Since the final equations and results are very similar for the two study regions, this section will only present results for the Miami-Dade domain. Table 5 displays the final operational logistic regression equations giving the conditional probability of a Q1 lightning event, an event in the upper two quartiles, and a Q4 lightning event for

Table 5. Final logistic regression models for the Miami-Dade domain.

Model for the probability of a Q1 lightning event (< 8 flashes)

Predictor	Coefficient (B)	Std. Error	Wald	P-value	Odds ratio
UPERP	-0.709	0.121	34.565	0.000	0.492
$UPERP^2$	0.117	0.048	5.850	0.016	1.124
$UPERP^3$	0.099	0.031	10.436	0.001	1.104
KI^2	0.130	0.033	15.650	0.000	1.139
PDQRT	-0.187	0.052	13.073	0.000	0.830
CLIQ3Q4*	-2.105	0.279	57.043	0.000	0.122

Model for the probability of a Q3 or Q4 lightning event (≥ 40 flashes)

Predictor	Coefficient (B)	Std. Error	Wald	P-value	Odds ratio
UPERP	0.951	0.114	69.695	0.000	2.587
$UPERP^2$	-0.217	0.046	22.488	0.000	0.805
$UPERP^3$	-0.138	0.033	17.136	0.000	0.871
KI^2	-0.136	0.036	14.143	0.000	0.873
MLI	-0.268	0.065	17.039	0.000	0.766
PDQ3Q4	0.680	0.105	42.219	0.000	1.974

Model for the probability of a Q4 lightning event (> 125 flashes)

Predictor	Coefficient (B)	Std. Error	Wald	P-value	Odds ratio
UPERP	0.874	0.117	56.026	0.000	2.396
$UPERP^2$	-0.477	0.087	30.286	0.000	0.620
KI^2	-0.238	0.067	12.715	0.000	0.789
MLI	-0.335	0.075	19.964	0.000	0.716
PDQ3Q4	0.710	0.150	22.479	0.000	2.034
CLIQ3Q4*	-2.091	0.253	68.249	0.000	0.124

* $CLIQ3Q4 = 0.562 - 0.063*(DAY^2)$, where DAY is a z-score

the Miami-Dade domain. This table lists the predictors in each of the equations, their coefficients (B) and standard errors for the coefficients, as well as other statistics that indicate the significance of each term and its relative predictive importance. The Wald statistic (calculated by $(B / \text{Std. error})^2$) and its p-value test the null hypothesis that there is no relationship between the independent variable and the log-likelihood of observing a "1" in the binary dependent variable (also called the "log-odds") (Hosmer and Lemeshow 1989). For example, a p-value of 0.001 indicates that there is only a 1/1000 chance that the relationship found in the sample would not also be true in the population, indicating that the parameter has "statistical significance" (Wilks 1995). The p-values in Table 5 show that all of the coefficients exceed the 95% significance level, and all but one exceed the 99% level, providing strong evidence that the parameters are significant and belong in the equations.

Also shown in the last column of Table 5 are odds ratios for each of the parameters (calculated by exp(B)). The odds ratio is defined as the ratio of the odds of observing the outcome of interest (e.g., a Q4 day) to the odds of not observing the outcome (e.g., the odds of observing a Q3 or lesser event). The values shown in Table 6 indicate the change (increase > 1, decrease < 1) in the odds of observing the outcome of interest for every one unit change in the predictor variable, holding all other variables in the equation constant (Hosmer and Lemeshow 1989).

It is informative to discuss the importance of each parameter in the equations (Table 5) and their relationships to lightning activity. First, it is clear that the three prediction equations generally are a variation on the same theme. The dominant effect in each equation is the component of the wind perpendicular to the coastline (UPERP). This is not surprising, since it is well known that the magnitude and direction of the prevailing low-level flow has a significant influence on the strength and inland penetration of the sea-breeze front (López and Holle 1987; Reap 1994; Lericos et al. 2002).

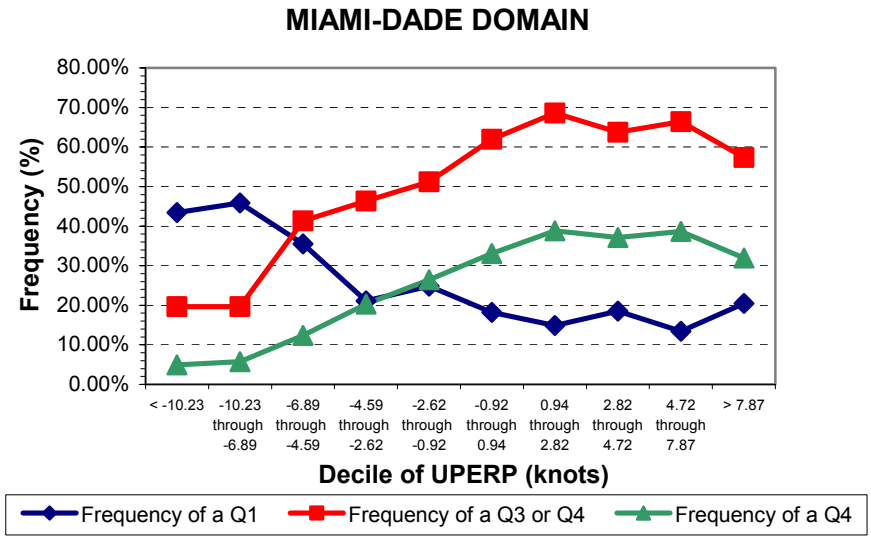
The magnitude of UPERP's effect on the likelihood of each outcome can be assessed by examining the signs of the coefficients and their odds ratios (Table 5). Since the first-order terms for each of the physical variables are standardized quantities (or z-scores), their respective coefficients indicate the change in the log-odds ($\ln[\pi/(1-\pi)]$) of the dependent variable for every one standard deviation (1σ) change in the independent variable, holding all other variables constant (Hosmer and Lemeshow 1989). Since UPERP is positive for offshore flow and negative for onshore flow, a 1σ increase in UPERP (~ 7 knots) signifies a stronger offshore (or weaker onshore) component, while a 1σ decrease signifies a stronger onshore (or weaker offshore) component. For example, in the equation for the probability of a Q1 event the coefficient for UPERP is negative, while it is positive for the other two equations. If we examine the corresponding odds ratios it is clear that a 1σ increase in UPERP (greater offshore component) decreases the odds of a Q1 event by a factor of 1 / 0.492 ~ 2 and increases the odds of a Q4 event by a factor of 2.396.

The signs of the coefficients of UPERP and the odds ratios are consistent with physical concepts. Specifically, offshore flow restricts the Atlantic coast sea breeze to near the coastline and allows a greater eastward penetration of the Gulf coast sea breeze, which increases the likelihood of enhanced convective activity and lightning in the two study regions. Conversely, a strong onshore flow tends to produce a less vigorous Atlantic coast sea breeze that migrates further inland, typically confining most convection and lightning to the western peninsula, away from the two study areas. Of course, there can be exceptions when outflow boundaries propagate into the two regions and trigger new convection. Since this particular situation is not handled by the prediction equations, it sometimes can produce an incorrect forecast.

There is a significant non-linear relationship between UPERP and lightning activity, as evident by the higher order terms included in the equations (Table 5). An illustration of the non-linearity is shown in Fig. 5a. This figure shows a clear increase in the frequency of events in the upper two quartiles, and a decrease in the frequency of Q1 events, as the flow becomes more offshore (as UPERP increases), reaching a maximum at an offshore speed of ~ 1-3 knots. Arritt (1993) found that similar flows are associated with the most intense sea breezes and greatest vertical velocities, which leads to more intense thunderstorm and lightning activity. As the flow becomes even more strongly offshore, the frequency of events in the upper two quartiles begin to decline, and Q1 events become more likely. This may be due to the strong opposing flow inhibiting the sea breeze front from advancing as far inland, or even remaining just offshore (Arritt 1993).

The K index is another important predictor in each of the equations (Table 5). It is interesting

a)



b)

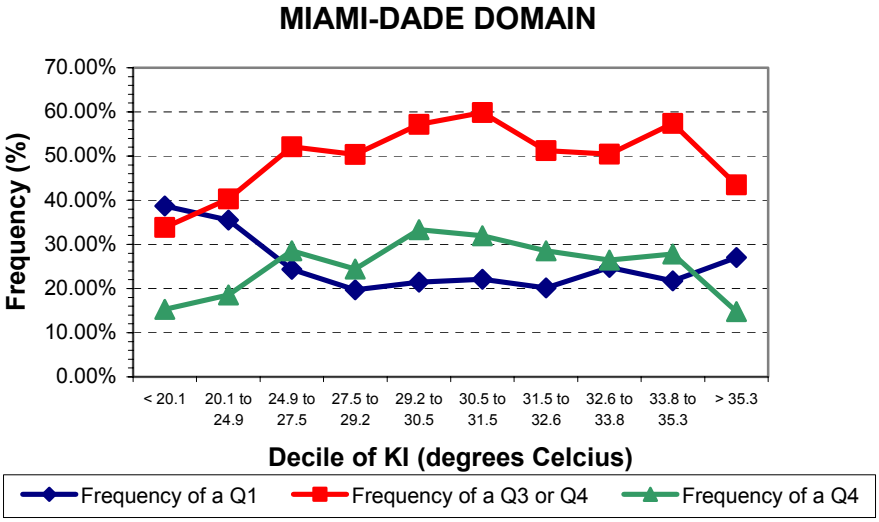


FIG. 5. The frequency of a Q1, the upper two quartiles, and a Q4 for each decile of (a) UPERP (knots) and (b) K-index for the Miami-Dade domain.

---

that this effect appears only as a second-order term (with no linear term). Fig. 5b shows that the frequency of events in the upper two quartiles increases with increasing K index until a peak is reached between 30.5 °C and 31.5 °C, after which the frequencies decline while the frequency of Q1 events increase. Since K index increases with more unstable mid-level lapse rates and greater middle-tropospheric moisture, it is reasonable that convection and lightning will increase with increasing K index. The decline in lightning activity for K index values larger than ~ 35 °C is not entirely clear, but may be due to excess mid-level moisture and cloud cover from early morning convection (i.e., at or near the sounding time). This would tend to suppress surface heating and strong afternoon activity.

Also appearing as a first order term in two of the three prediction equations is the modified lifted index (MLI). A decrease in MLI indicates greater afternoon instability, and vice versa. The coefficients and odds ratios in the Q3/Q4 and Q4 equations (Table 5) indicate that the odds of an upper two quartile lightning event increase for

every 1σ decrease in MLI (about 1.4 C°). Fuelberg and Biggar (1994) found a form of modified lifted index (SLI, i.e., a surface lifted index based on conditions at 1100 EDT) to be a useful stability index for discriminating between categories of afternoon convective development over the Florida panhandle.

The dominant non-physical predictor in the guidance equations is persistence (Table 5). Except for the Q1 equation, the most important persistence variable is the yes/no binomial indicator stating whether or not the previous day was in the upper two quartiles (PDQ3Q4). The odds ratios in Table 5 indicate that the odds of observing a Q4 lightning day increase by a factor of ~ 2 if the value of PDQ3Q4 is a "1" (i.e., if the previous day was in the upper two quartiles). Day number enters the Q1 and Q4 equations through the climatic predictor (CLIQ3Q4, shown at the bottom of Table 5). Other studies also have found the day number (in conjunction with other parameters) to be a useful predictor of warm season thunderstorm probabilities over Florida (e.g., Neumann and Nicholson 1972; Reap 1994).

4.2 *Results for Dependent Data*

To assess the predictive skill of each logistic model on the 14 warm seasons of dependent data, 2 x 2 contingency tables were constructed showing the number of observed "1s" and "0s" versus the number predicted for each outcome (Table 6). Various statistics also were calculated to assess the accuracy of each model. These include the probability of detection (POD), hit rate, false-alarm ratio (FAR), the bias, and the critical success index (CSI). Table 7 shows a sample contingency table used in computing the scores shown in Table 6. The POD is the ratio of the number of "1s" correctly predicted by the model to the total number of observed "1s" in the sample: POD = x / (x + y). The hit rate is the most direct measure of the accuracy of categorical (yes/no) forecasts, indicating the percentage of correctly classified "1s" and "0s": hit rate = (x + w) / (w + x + y + z). The FAR is measure of the forecast events ("1s") that fail to materialize: FAR = z / (x + z). The bias (B) indicates the degree of over-forecasting (B>1) or under-forecasting (B<1) the outcome: B = (x + z) / (x + y). Finally, the CSI is a frequently used alternative to the hit rate when the event to be forecast occurs less frequently than its nonoccurrence (e.g., the Q1 and Q4 days), and combines attributes of the POD and FAR: CSI = x / (x + y + z) (Reap 1994; Mazany et al. 2002).

Examination of the computed statistics for each equation (Table 6) reveals that the greatest accuracy is achieved (i.e., the highest CSI and lowest FAR) with the Q3/Q4 model, which distinguishes upper two lightning quartile events from lower two quartile events. The Q1 and Q4 equations generally perform well at differentiating Q1 days from other days, and Q4 days from other days (POD ranges from 65-70%). However, the Q1 and Q4 events are considerably over-forecast, as evident by the relatively large bias and FAR. This indicates that Q1 days are not easily distinguishable from Q2 days, and Q4 days are not easily distinguishable from Q3 days. This also may be due to the choice of the cut point between the quartile categories (i.e., < 8 flashes or > 125 flashes).

Once probabilities are obtained from the three logistic equations in Table 5 (giving prob(Q1), prob(Q3/Q4), and prob(Q4)), the next step is to determine the most likely quartile for the day (i.e., Q1, Q2, Q3, or Q4). Since each regression equation does not contain the same exact set of predictor variables, one cannot simply solve for the probability of each quartile by using the output from the three equations. Instead, as described in section 3, the best results for predicting the quartile were obtained by constructing a decision tree using probability thresholds for the three equations. The thresholds were set so an equal number of cases in the dependent data set were partitioned to the left and right at each branch to eliminate any prediction bias toward any one quartile. The resulting decision tree for the Miami-Dade domain is shown in Table 8. The first branch to the left or right depends on the probability output by the Q3/Q4 model, since this model distinguishes the lower two quartiles from the upper two. For example, if the probability of ≥ 40 flashes is ≥ 0.51654, the right branch is taken and that day is predicted to be either a Q3 or Q4 event. Then, output from the Q4 equation is used to determine which of these two quartiles is more likely. If the probability of a Q4 day (> 125 flashes) is ≥ 0.39178, a Q4 lightning day is forecast, otherwise that day is predicted to be a Q3 event. Conversely, if the probability of ≥ 40 flashes is less than the threshold of 0.51654, the left branch is taken and either a Q1 or Q2 event is more likely, in which case output from the Q1 model determines which to predict. If the probability of a Q1 (< 8 flashes) is ≥ 0.33085, a Q1 is predicted; otherwise Q2 is forecast to occur.

The overall accuracy of the prediction scheme can be easily assessed by examining a 4 x 4 contingency table for the number of observed days

Table 6. 2 x 2 contingency tables for each logistic regression model for the Miami-Dade domain. The results shown are for the 14 years of dependent data. The probability cut values shown beneath each table were chosen to maximize the hit rate for comparison only.

Model for the probability of a Q1 lightning event (< 8 flashes)

	PREDICTED					
	Q1	> Q1	Total	% Correct		
OBSERVED					CSI:	0.322
Q1	**199**	110	309	64	FAR:	0.608
> Q1	309	**591**	900	66	POD:	0.644
					Hit rate:	0.653
Total	508	701	1209	65	Bias	1.644

Probability cut value = 0.25

Model for the probability of a Q3 or Q4 lightning event (≥ 40 flashes)

	PREDICTED					
	Q3/Q4	Q1/Q2	Total	% Correct		
OBSERVED					CSI:	0.531
Q3/Q4	**422**	182	604	70	FAR:	0.312
Q1/Q2	191	**414**	605	68	POD:	0.699
					Hit rate:	0.691
Total	613	596	1209	69	Bias	1.015

Probability cut value = 0.51

Model for the probability of a Q4 lightning event (> 125 flashes)

	PREDICTED					
	Q4	< Q4	Total	% Correct		
OBSERVED					CSI:	0.368
Q4	**209**	95	304	69	FAR:	0.558
< Q4	264	**641**	905	71	POD:	0.688
					Hit rate:	0.703
Total	473	736	1209	70	Bias	1.556

Probability cut value = 0.30

Table 7. Sample 2 x 2 contingency table for computing skill scores.

	PREDICTED		
	Yes	No	Total
OBSERVED			
Yes	*x*	*y*	*x* + *y*
No	*z*	*w*	*z* + *w*
Total	*x* + *z*	*y* + *w*	*w* + *x* + *y* + *z*

Table 8. Probability decision tree used to determine the predicted lightning quartile for the Miami-Dade domain. Also shown is a 4 x 4 contingency table for the number of observed days in each quartile versus the number predicted using the decision tree. These results are for the 14 years of dependent data.

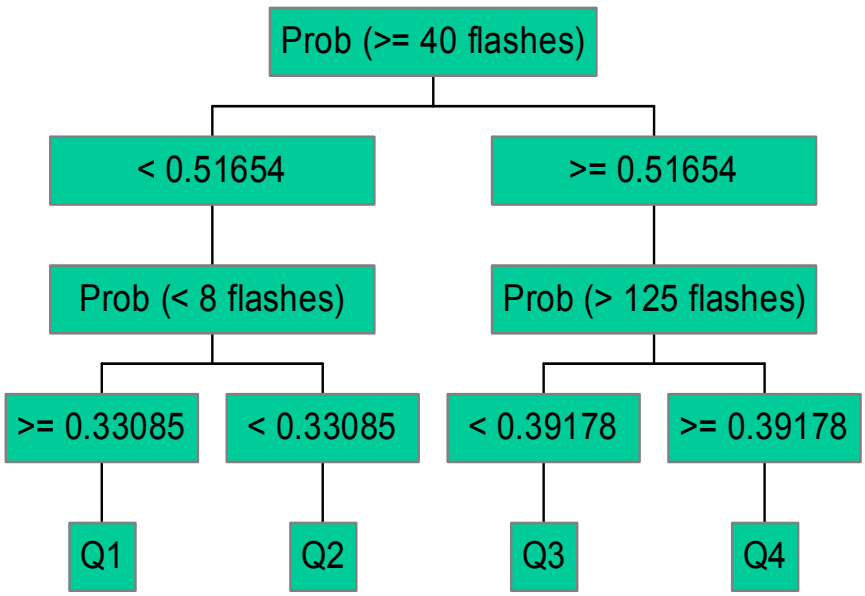


Table 8 (continued).

Results for all 14 warm seasons of dependent data
(Accuracy measures have been rounded to the nearest percent)

	PREDICTED						
	Q1	Q2	Q3	Q4	Total	% Correct	% Within 1Q
OBSERVED							
Q1	**135**	93	46	35	309	44	74
Q2	91	**99**	64	42	296	33	86
Q3	51	81	**92**	76	300	31	83
Q4	25	29	100	**150**	304	49	82
Total	302	302	302	303	1209	**39**	**81**

Persistence:	34	73
SS_{pers}:	**6**	**8**
Climatology:	25	63
SS_{clim}:	**14**	**19**

in each quartile versus the number predicted from the decision tree. The results for the Miami-Dade scheme are shown in Table 8 (beneath the corresponding decision tree) for all 14 warm seasons of dependent data. It is encouraging that the number of days observed for each quartile, and the number of days predicted, is maximized along the diagonal. The scheme best forecasts Q1 and Q4 events, with a hit rate of ~ 44% and ~ 49%, respectively. It again appears that Q2 days are not easily distinguishable from Q1 days, and Q3 days are not easily distinguished from Q4 days. Thus, the hit rates for the Q2 and Q3 quartiles are somewhat smaller (between 30-33%). This may occur because many cases in the data set were assigned predicted probabilities that were very close to the thresholds for being partitioned to the left or right at each branch of the decision tree. In addition, flash counts on many days straddle the cut point between the quartile categories, which also could contribute to the lack of discernability for the Q2 and Q3 days. Of course, the probability thresholds in the decision tree can be adjusted to increase the detection for any quartile of choice (e.g., the Q4s), but not without creating a bias toward that quartile. In this case, a more meaningful measure of accuracy is the percentage of time the scheme predicts to within one quartile of the observed. For example, when Q1 is observed, the Miami-Dade scheme predicts either Q1 or Q2 ~ 74% of the time, and when a Q4 is observed, the scheme predicts either Q3 or Q4 ~ 82% of the time. The percentages for the Q2 and Q3 events are somewhat higher since there are three possible predictions that can be within one quartile of the observed in this case.

Considering all quartiles together, the Miami-Dade scheme forecasts the correct quartile ~ 39% of the time and is correct to within one quartile of the observed ~ 81% of the time. Shown beneath the 4 x 4 table is the percentage of correctly classified days and the percentage correct to within one quartile of the observed for the reference forecasts (i.e., persistence and climatology). Also shown is the difference in forecast accuracy (i.e., a skill score expressed as a percentage point difference) between the model and that achieved by persistence and climatology, denoted by SS_{pers} and SS_{clim}, respectively. The positive scores show that the two schemes are superior to both climatology and persistence, and

thereby represent real forecast skill. For the percentage of days correctly forecast, the scheme is a ~ 6 percentage point improvement over persistence and a ~ 14 percentage point improvement over climatology (a random guess would correctly forecast the quartile by chance ~ 25% of the time). In terms of the percentage of forecasts correct to within one quartile of the observed, the scheme is a ~ 8 percentage point improvement over persistence and ~ 19 percentage point improvement over climatology.

4.2 *Results for K-fold Cross-Validation*

The results presented in the previous section (and those shown in Tables 6 and 8) are for all 14 warm seasons of "dependent" data. That is, the results show the predictive accuracy of the scheme on the same data that was used to derive it. These results do not fairly depict how well the guidance equations will predict cases that were not involved in the model development. Thus, the k-fold cross-validation procedure described in section 3 was conducted to give a better estimate of how the prediction scheme will perform when implemented operationally.

Figures 6a and b, Fig. 7 and Table 9 show the performance of the scheme when each year is treated independently (i.e., when each warm season is withheld from the data set for cross-validating the scheme derived from the remaining 13 warm seasons of data). Figure 6a is a plot of the cross-validated hit rate compared to persistence for each independent warm season for the Miami-Dade domain. It is clear that a persistence forecast is very difficult to beat during any particular year. The scheme either matches or beats persistence on 10 of the 14 warm seasons, and underperforms persistence on 4 of the years. All but one of the independent tests shows greater skill than climatology. It also is clear from Fig. 6a that the cross-validated hit rate is fairly stable, with most years ranging between 30-40%. When considering the percentage correct to within one quartile of the observed, the scheme demonstrates even greater skill over persistence and climatology (Fig. 6b), with most years beating persistence by 8-12 percentage points (to as much as 15) and climatology by 15-20 percentage points.

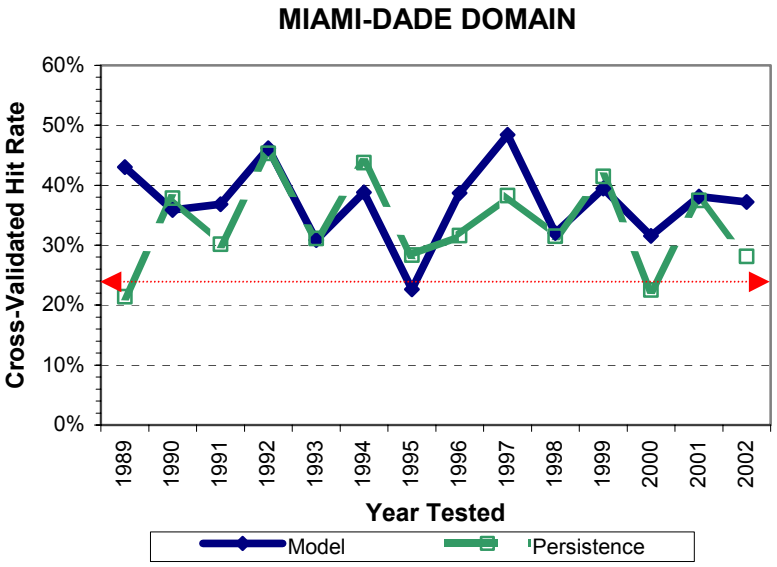
It is informative to assess the performance of the guidance equations by month. Fig. 7 plots monthly values of cross-validated hit rate and the percentage correct to within one quartile of the observed for all independent test seasons combined for the Miami-Dade domain. The monthly variations in accuracy generally are small; however, the greatest accuracy is evident during June through August (the hit rate peaks in July), with somewhat less skill during May and September. This is expected since June-August is the period when the sea breeze is most likely to be the dominant forcing mechanism for convection and lightning in South Florida, unlike May and September, which often contain days with synoptic or tropical influences. Some effort was made to remove these days from the analysis (by removing days with MNSPD greater than 3σ from the climatological mean value). However, it cannot be assumed that all such days were removed, which undoubtedly plays a role in the slight reduction in accuracy during May and September. Nevertheless, the skill scores for all months generally are within $\pm$ 10% of each other. This demonstrates that it is not necessary to have separate prediction schemes for each month, and also shows the benefit of including a climatic predictor in the equations.

Table 9 shows a 4 x 4 contingency table for the number of observed days in each quartile versus the number predicted for all independent years combined (i.e., the 14 individual tables for each cross-validated warm season were summed into a single table). The computed skill scores show that the overall cross-validated percentage of correctly forecast days is a ~ 4 and ~ 12 percentage point improvement over persistence and climatology, respectively, and the percentage correct to within one quartile of the observed is an improvement over persistence and climatology by ~ 6 and ~ 17 percentage points, respectively. Although there are year-to-year variations (Figs. 6a, b), the overall cross-validation results show only a 1-2 percentage point reduction in skill from what was obtained for the 14 years of dependent data (Table 8). This demonstrates that the prediction scheme is statistically "robust." Thus, there is a high degree of confidence that the guidance equations will achieve similar results when implemented operationally by FP&L.

4.3 *Upper Quartile (Q4) Lightning Events*

The 4 x 4 contingency table (Table 9) shows that the scheme correctly predicts Q4 lightning events on ~ 48% of the cases for the 14 independent tests, and is correct to within one quartile (i.e., either a Q3 or Q4 was predicted) ~ 79% of the time. These numbers are quite good. It is interesting to see how the frequency of correct predictions of Q4 events varies as the magnitude of the Q4 event increases (i.e., for

a)



b)

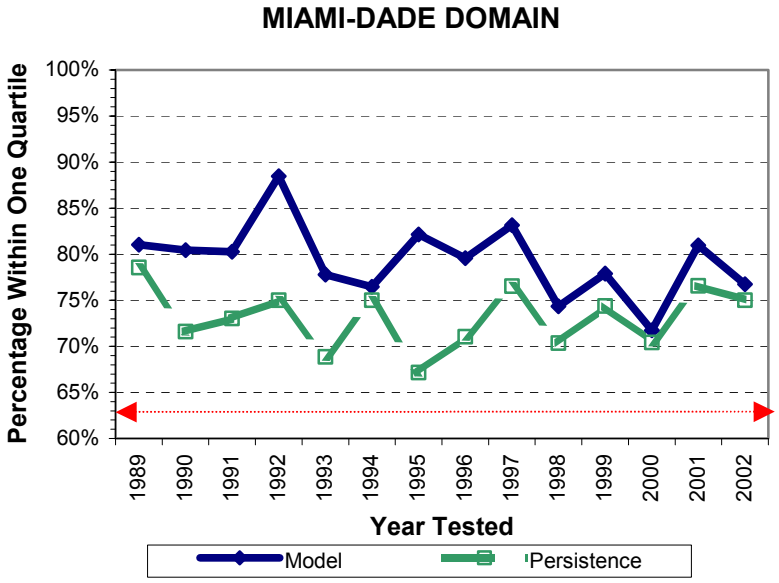


FIG. 6. (a) Cross-validated hit rate and (b) percentage correct to within one quartile of the observed for each year treated independently for the Miami-Dade domain. The red horizontal line on each plot denotes the climatology percentages, which serves as the zero reference level for forecast skill.

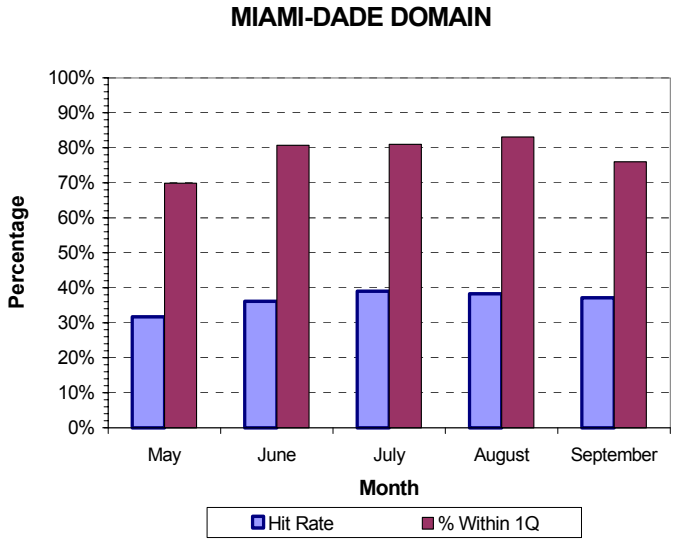


FIG. 7. Cross-validated hit rate and the percentage correct to within one quartile of the observed by month for all independent test seasons combined for the Miami-Dade domain.

Table 9. 4 x 4 contingency table for the number of observed days in each quartile versus the number predicted for all independent years combined for the Miami-Dade domain.

All independent test years combined
(Accuracy measures have been rounded to the nearest percent)

	PREDICTED						
	Q1	Q2	Q3	Q4	Total	% Correct	% Within 1Q
OBSERVED							
Q1	**124**	95	56	34	309	40	71
Q2	92	**89**	67	48	296	30	84
Q3	46	84	**94**	76	300	31	85
Q4	23	42	94	**145**	304	48	79
Total	285	310	311	303	1209	**37**	**79**

Persistence:	34	73
SS_{pers}:	**4**	**6**
Climatology:	25	63
SS_{clim}:	**12**	**17**

larger flash events within the Q4 category). Figure 8 subdivides the Q4 lightning events into four (quartile) groups based on CG flash count (with ~ 75 cases in each group), and plots the percentage of cases that were correctly predicted to be a Q4 event (i.e., > 125 flashes) for each subgroup of flashes. Also shown is the percentage of cases that were correctly predicted to be in the upper two quartiles (i.e., ≥ 40 flashes) for each subgroup. This figure shows that correct predictions of Q4 events generally increase with increasing flash count within the Q4 category. For the lowest range of flashes within this category, the scheme correctly predicts a Q4 event ~ 36% of the time and an upper two quartile event ~ 74% of the time. In contrast, for cases when at least 381 flashes were observed during the noon-midnight period (the upper one-sixteenth of all events), the scheme flags the event as a Q4 day ~ 66% of the time and is within one quartile nearly 95% of the time! This demonstrates that the scheme correctly identifies the bigger flash events as a Q4 day most of the time, and will only rarely forecast the lower two quartiles when a large Q4 event occurs.

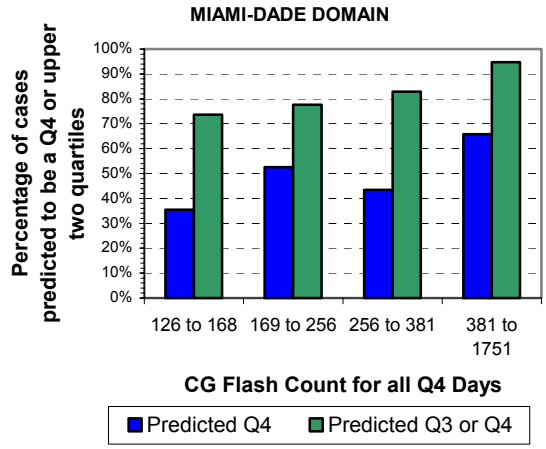


FIG. 8. Percentage of cases predicted to be a Q4 event or the upper two quartiles (Q3 or Q4) for different ranges (quartiles) of CG flash count for all Q4 events for the Miami-Dade domain.

4.3 *Test on Days with No Observed Activity*

Although the prediction equations were derived for days when at least one flash occurs in either of the two study areas, it is interesting to determine what the scheme would have predicted on days when no lightning was observed. This was investigated by testing the final set of equations (Table 5) on those days in the original data set when no activity was observed during the noon-midnight time period in each domain (these data were never used during model development). The decision tree in Table 8 was used to predict a quartile based on the output from the three probability equations.

Out of 778 days with no activity in the Miami-Dade region, the scheme predicted a Q1 event 579 times (~ 74%), and predicted either a Q1 or Q2 event on 700 occasions (~ 90%), conditional on at least one flash occurring. This indicates that the scheme only rarely predicts the upper two quartiles of activity on days when no lightning is observed in the two study areas. This is a desirable result, since one hopes that the scheme will not often give a false alarm of this magnitude. Nevertheless, there were 78 days in the Miami-Dade area (~ 10% of the non-lightning days) when the scheme did predict either a Q3 or Q4 (conditional on at least one flash occurring) but no lightning was observed in the study area.

An analysis of these major "false alarm" days revealed that the prediction was warranted on physical grounds, i.e., conditions were very favorable for an event in the upper two quartiles. Specifically, close to 90% of the days had either weakly onshore or offshore flow, 75% had a K-index > 25, nearly 80% had a MLI < -5, and 45% had either a Q3 or a Q4 event the previous day (which represents a drastic departure from persistence that the scheme did not capture). Of the 20 days (~ 2.5%) when a Q4 was incorrectly predicted for the Miami-Dade region, 8 days had some activity in the Broward domain. Similarly, of the 34 cases (~ 4.2%) when a Q4 was incorrectly predicted in the Broward region, 17 of those days had activity in the Miami-Dade domain. This suggests that the prediction was not completely wrong since there was nearby lightning activity. Nonetheless, it did not occur within the boundaries of the domain being forecast. Analysis conducted in the companion study to this one (Winarchick and Fuelberg 2005, *Conference on Meteorological Applications of Lightning Data*) has shown that this type of situation occurs frequently. That is, intense lightning can occur in close proximity to, but not within the domain areas (i.e., one county away or just on the other side of Krome Avenue or U.S. Route 27). Such situations often lead to a busted forecast, even though conditions were favorable for an active lightning day.

These findings illustrate the inherent difficulty in attempting to forecast lightning and thunderstorm activity for very small regions and for

a specific time period. Of course, one would expect forecast accuracy to increase with the size of the domain area. To assess whether improved results would be achieved if the forecast domain was enlarged, a set of prediction equations was developed for a larger area of South Florida encompassing the entire two county region. The results (not presented here) showed that indeed more accurate predictions are made when the forecast area is enlarged. However, the prediction scheme for the larger area was an improvement over persistence by nearly the same proportionate amount as the scheme for the smaller domain. Thus, from the point of view of FP&L, there would be little to gain (with respect to improvement over persistence) from expanding the forecast area. This is especially true since the majority of the population is located in the eastern halves of the two counties (i.e., to the east of Krome Avenue and U.S. Route 27), and lightning strikes over the water conservation areas to the west usually do not cause power outages and are of minimal concern to FP&L officials.

5. RESULTS FOR SUMMER 2004

The lightning guidance equations described in the previous sections (Table 5) were run daily during the 2004 warm season (May-September) to assess the predictive accuracy of the scheme on a completely independent data set. The output from the equations and the predicted lightning quartile (Table 8) were sent to Florida Power & Light each morning for their consideration. Table 10 presents the results for June-August 2004 for the Miami-Dade domain for all days when there was at least one flash observed during the noon-midnight period (the results for Broward were comparable and are not shown). Results for May are not included here since the month was primarily dry, containing only 2 days with observed lightning within the study area. The month of September, on the other hand, was highly anomalous due to several tropical systems that affected Central and South Florida during the month. These highly disturbed conditions violated the basic assumption behind our statistical guidance, that the sea breeze is the dominant forcing mechanism for convection and associated lightning.

Overall, the prediction scheme performed well during the June-August 2004 period. The values in the 4 x 4 contingency table (Table 10) indicate that the correct quartile was forecast ~ 38% of the time, which is very close to the percentage obtained from the 14-year k-fold cross-validation (Table 9). The guidance was correct to within one quartile of the observed ~ 88% of the time, which is actually better than the cross-validation results by ~ 9 percentage points. This again demonstrates that the scheme does best predicting to within one quartile of the observed, as well as differentiating the upper two quartile events from the lower two quartiles. This is further evident by the scores shown in the 2 x 2 contingency table (bottom of Table 10) for the Q3/Q4 equation. Recall that this equation gives the conditional probability of the upper two quartiles ($\geq$ 40 flashes in Miami-Dade domain), and its output determines the first branch in the decision tree for predicting the quartile (Table 8). Overall, the model correctly distinguished upper two quartile events from lower two quartile events on ~ 69% of the days during the period, with a CSI of 49%, a POD of 64%, and a FAR of 32%.

6. SUMMARY AND CONCLUSIONS

This study utilized 14 warm seasons of data (1989-2002) from the National Lightning Detection Network, and morning radiosonde releases from Miami and West Palm Beach, to develop statistical guidance equations describing the amount of lightning that will occur during the noon-midnight period over eastern Miami-Dade and Broward Counties in South Florida. A total of 54 sounding parameters were calculated that have been found in previous studies to be useful predictors of thunderstorms and lightning over Florida. These parameters describe wind direction and speed in various layers, as well as moisture, temperature, and stability.

Several statistical techniques that were attempted initially were found to yield undesirable results, including multiple linear and Poisson log-linear regression, as well as Classification and Regression Trees. The best results were obtained by creating four quartile groups of flash count based on climatology, and then using binary logistic regression to develop three prediction equations for each domain, one giving the conditional probability of the lowest quartile of flashes, another for the probability of an upper two quartile event, and a third equation giving the probability of an event in the highest quartile.

Results for the Miami-Dade domain were presented. The three probability equations generally were found to be a variation on the same theme. The dominant effect in each of the equations was the cross-shore wind component (UPERP), which was found to have a significant non-linear relationship with lightning activity. The peak likelihood of Q3 and Q4 events was found for

Table 10. 4 x 4 contingency table for the number of observed days in each quartile versus the number predicted during June-August 2004 for the Miami-Dade domain. Also shown is a 2 x 2 contingency table for the model that differentiates upper two vs. lower two quartile events.

Results for June-August 2004

(Accuracy measures have been rounded to the nearest percent)

	PREDICTED						
	Q1	Q2	Q3	Q4	Total	% Correct	% Within 1Q
OBSERVED							
Q1	**11**	4	1	1	17	65	88
Q2	5	**9**	5	3	22	41	86
Q3	1	8	**1**	1	11	9	91
Q4	0	3	13	**6**	22	27	86
Total	17	24	20	11	72	**38**	**88**

Model for the probability of a Q3 or Q4 lightning event (≥ 40 flashes)

	PREDICTED					
	Q3/Q4	Q1/Q2	Total	% Correct		
OBSERVED					CSI:	0.488
Q3/Q4	**21**	12	33	64	FAR:	0.323
Q1/Q2	10	**29**	39	74	POD:	0.636
					Hit rate:	0.694
Total	31	41	72	**69**	Bias	0.939

Probability cut value = 0.51654

light offshore flow of between 1-3 knots. This type of flow is associated with the most intense sea breezes and vertical velocities, which leads to enhanced convection and lightning in the two study areas. Conversely, a strong onshore flow produces a weaker sea breeze that migrates further inland, typically confining most convection and lightning to the west of the two study regions. Other important variables were found to be the K-index and modified Lifted Index. Day number and persistence also were selected as important indicators for the amount of afternoon lightning that will occur in the study region.

The accuracy of the prediction scheme generally was found to be superior to persistence and climatology for both the dependent data and during k-fold cross-validation. Thus, they possess real forecast skill. Overall, the cross-validation results showed only a 1-2 percentage point reduction in skill from that obtained for the fourteen years of dependent data, demonstrating that the two schemes are statistically robust, and can be expected to achieve similar results when implemented operationally by FP&L.

The guidance equations that were derived in this study utilized parameters calculated from the morning 1200 UTC sounding at Miami/West Palm Beach. This approach was based on several assumptions that are not valid on all days. For example, it was assumed that atmospheric

conditions do not vary significantly from the sounding time through the end of the forecast period. This assumption is approximately valid most of the time over South Florida during the warm season, but sometimes is violated if advection of a different air mass occurs, such as the passage of a synoptic scale system or tropical disturbance. It also was assumed that atmospheric conditions at the radiosonde site are representative of those for the entire domain area, which may not necessarily be true, even during the warm season. Whenever these assumptions are not met, there will be errors in the lightning forecast. It also is clear that factors not considered in this study have an important influence on the degree of afternoon convection and lightning that occurs over eastern Miami-Dade and Broward Counties. These include outflow boundaries from pre-existing storms, and the interaction of smaller scale circulations such as lake/river breezes with the sea breeze. These processes often aid in forming new convection in areas that otherwise would not be favored because of the speed and direction of the prevailing low-level flow. Cloud microphysical processes also were not considered in this study.

Despite these limitations, the current results show how remarkably well one can predict afternoon convection and lightning for areas as small as the eastern halves of two counties with input from just a morning sounding. Future work will seek to improve the current results by incorporating mesoscale model output into the equations. The model data will be more location and time specific than just a static 1200 UTC sounding at one location, and likely will result in considerable improvement in forecast skill. Nevertheless, the current scheme will provide useful guidance about the degree of afternoon and evening lightning activity that will occur over the heavily populated areas of eastern Miami-Dade and Broward Counties in South Florida.

7. ACKNOWLEDGMENTS

This research was funded by Florida Power & Light Corporation. Special thanks go to Dr. James Elsner of the Florida State University Department of Geography and Justin Winarchick of the Department of Meteorology for their assistance and contributions to the project. Collaborations with the National Weather Service were fostered by the NOAA CSTAR grant NA03NWS4680018. Appreciation is extended to Irv Watson of the National Weather Service in Tallahassee, FL, and to Rusty Pfost and Dr. Pablo Santos of the National Weather Service in Miami, FL, for their ideas and suggestions. Lastly, a special thanks goes to Paul Hebert from Florida Power & Light Corporation for contributing his extensive knowledge of the South Florida sea breeze and summertime weather patterns.

8. REFERENCES

American Meteorological Society, 2000: *Glossary of Meteorology*, Second Edition.

Arritt, R. W., 1993: Effects of the large-scale flow on characteristic features of the sea breeze. *J. Appl. Meteor.*, 32, 116-125.

Blanchard, D. O., and R. E. López, 1985: Spatial patterns of convection in south Florida. *Mon. Wea. Rev*., 113, 1282-1299.

Brenner, I. S., 2004: The relationship between meteorological parameters and daily summer rainfall amount and coverage in West-Central Florida. *Wea. Forecasting*, 19, 286-300.

Burrows, W. R., Price, C., and Wilson, L. J., 2004: Statistical models for 1-2 day warm season lightning prediction for Canada and the Northern United States. Preprints, *17th Conference on Probability and Statistics in the Atmospheric Sciences*, Seattle, WA, Amer. Meteor. Soc.

Camp, J. P., A. I. Watson, and H. E. Fuelberg, 1998: The diurnal distribution of lightning over north Florida and its relation to the prevailing low-level flow. *Wea. Forecasting*, 13, 729-739.

Cummins, K. L., M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer, 1998: A combined TOA/MDF technology upgrade of the U.S. National Lightning Detection Network. *J. Geophys. Res.*, 103, 9035-9044.

Curran, E. B., R. L. Holle, and R. E. López, 1997: Lightning fatalities, injuries and damage reports in the United States from 1959-1994. NOAA Tech. Memo. NWS SR-193, 64 pp.

Estoque, M. A., 1962: The sea breeze as a function of the prevailing synoptic situation. *J. Atmos. Sci.*, 19, 244-250.

Forecast Systems Laboratory, 2004: *Radiosonde Database Access*. Available [http://raob. fsl. noaa.gov].

FSL and NCDC, 1999: *Radiosonde Data of North America1946-1999*. CD-ROM, Version 1.0. [Available from DOC/NOAA/OAR, Forecast Systems Laboratory, R/FSL, 325 Broadway, Boulder, CO 80305.]

Fuelberg, H. E., and D. G. Biggar, 1994: The preconvective environment of summer thunderstorms over the Florida Panhandle. *Wea. Forecasting*, 9, 316-326.

Gentry, R. C., and P. L. Moore, 1954: Relation of local and general wind interaction near the sea coast to time and location of air-mass showers. *J. Atmos. Sci*, 11, 507–511.

Hosmer, D. W., and S. Lemeshow, 1989: *Applied Logistic Regression*. John Wiley & Sons, Inc., 307 pp.

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

Livingston, E. S., J. W. Nielson-Gammon, and R. E. Orville, 1996: A climatology, synoptic assessment, and thermodynamic evaluation for cloud-to-ground lightning in Georgia: A study for the 1996 Summer Olympics. *Bull. Amer. Meteor. Soc.*, 77, 1483-1495.

López, R. E., P. T. Gannon, Sr., D. O. Blanchard, and C. C. Balch, 1984: Synoptic and regional circulation parameters associated with the degree of convective shower activity in South Florida. *Mon. Wea. Rev.*, 112, 686-703.

-----, and R. L. Holle, 1987: The distribution of summertime lightning as a function of low-level wind flow in central Florida. NOAA Tech. Memo. ERL ESG-28, National Severe Storms Laboratory, Norman, OK, 43 pp.

Maier, L. M., E. P. Krider, and M. W. Maier, 1984: Average diurnal variation of summer lightning over the Florida peninsula. *Mon. Wea. Rev.*, 112, 1134-1140.

Mazany, R. A., Businger, S., Gutman, S. I., and Roeder, W., 2002: A lightning prediction indext that utilizes GPS integrated precipitable water vapor. *Wea. Forecasting*, 17, 1034-1047.

Neumann, C. J., and J. R. Nicholson, 1972: Multivariate regression techniques applied to thunderstorm forecasting at the Kennedy Space Center. Preprints, *International Conference on Aerospace and Aeronautical Meteorology*, Washington, D.C., Amer. Meteor. Soc., 6-13.

Orville, R. E., and A. C. Silver, 1997: Lightning ground flash density in the contiguous United States: 1992-95. *Mon. Wea. Rev.*, 125, 631-638.

-----, G. R. Huffines, W. R. Burrows, R. L. Holle, and K. L. Cummins, 2002: The North American Lightning Detection Network (NALDN)—first results: 1998-2002. *Mon. Wea. Rev.*, 130, 2098-2109.

Reap, R. M., 1994: Analysis and prediction of lightning strike distributions associated with synoptic map types over Florida. *Mon. Wea. Rev.*, 122, 1698-1715.

-----, and D. R. MacGorman, 1989: Cloud-to-ground lightning: Climatological characteristics and relationships to model fields, radar observations, and severe local storms. *Mon. Wea. Rev.*, 117, 518-535.

Stroupe, J. R., 2003: *1989-2002 Florida Lightning Climatology*. Available [http://bertha. met. fsu. edu/~jstroupe/flclimo.html].

United States Census Bureau, 2004: *Population Estimates*. Available [http://www.census.gov].

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* International Geophysics Series, Vol. 59, Academic Press, 464 pp.

Winarchick, J. M., and H. E. Fuelberg, 2005: Developing a statistical scheme to predict the occurrence of lightning in South Florida. *Conference on Meteorological Applications of Lightning Data.* San Diego, CA, Amer. Meteor. Soc.