A Statistical Procedure to Forecast Warm Season Lightning over Portions of the Florida Peninsula

PHILLIP E. SHAFER* AND HENRY E. FUELBERG

Department of Meteorology, The Florida State University, Tallahassee, Florida

(Manuscript received 24 August 2005, in final form 19 December 2005)

ABSTRACT

Sixteen years of cloud-to-ground lightning data from the National Lightning Detection Network and morning radiosonde-derived parameters are used to develop a statistical scheme to provide improved forecast guidance for warm season afternoon and evening lightning for 11 areas of the Florida peninsula serviced by Florida Power and Light Corporation (FPL). Logistic regression techniques are used to develop equations predicting whether at least one flash will occur during the noon–midnight period in each area, as well as the amount of lightning that can be expected during this same period, conditional on at least one flash occurring. For the amount of lightning, the best results are achieved by creating four quartile categories of flash count based on climatology, and then using three logistic equations and a decision tree approach to determine the most likely quartile. A combination of forward stepwise screening and cross validation are used to select the best combination of predictors that are most likely to generalize to independent data. Results show the guidance equations to be superior to persistence on both the dependent dataset and during cross validation. The greatest skill scores are achieved for predicting whether at least one flash will occur, as well as predicting the number of flashes to within one quartile of that observed. These results demonstrate that the equations possess forecast skill and will provide useful guidance for the probability and amount of lightning in each of the 11 FPL service areas.

1. Introduction

Over the past 30 yr, cloud-to-ground (CG) lightning has exceeded both tornadoes and hurricanes in causing 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. Improved forecasts of the timing and location of thunderstorms and associated 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–September. Many studies examining lightning patterns across the contiguous United States have found that Florida receives more CG 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.

Figure 1 shows the spatial distribution of CG lightning in Florida for May–September during the 14-yr period 1989–2002 (Lericos et al. 2002; Stroupe 2004). 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 Miami on the east coast. These regions of enhanced flash density are due to many complex factors, including irregularly shaped and protruding coastlines, and thermal circulations such as the sea breeze and lake/river breezes (e.g., López and Holle 1987; Arritt 1993; Lericos et al. 2002).

During the warm season, absent of synoptic or tropical disturbances, the Atlantic and Gulf of Mexico sea breezes act as the primary triggering mechanism for

* Additional affiliation: NOAA/National Weather Service Forecast Office, Tallahassee, Florida.

Corresponding author address: Phillip E. Shafer, Dept. of Meteorology, The Florida State University, Tallahassee, FL 32306-4520.
E-mail: pshafer@met.fsu.edu

© 2006 American Meteorological Society

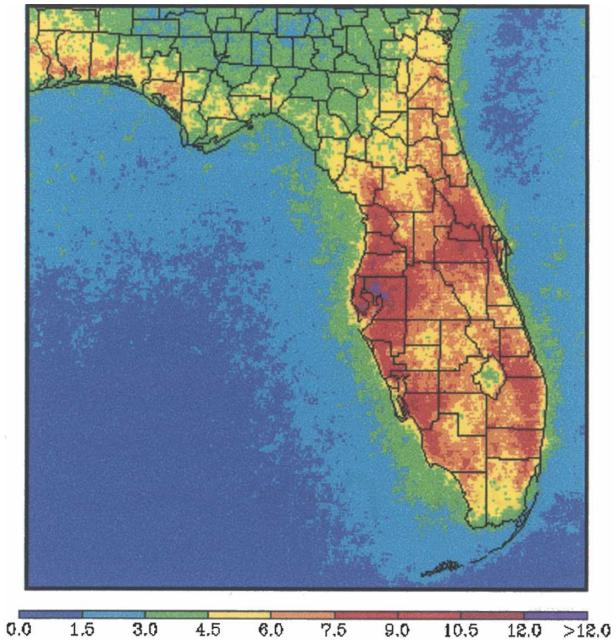

FIG. 1. Map of the spatial distribution of warm season CG lightning (flashes per square kilometer per warm season) for the state of Florida during a 14-yr period from 1989 to 2002. This figure is similar to that shown in Lericos et al. (2002), but has been updated with four additional warm seasons of lightning data.

afternoon convection in Florida. If adequate moisture and instability are present, the location and amount of afternoon thunderstorms are governed primarily by the strength and inland extent of the sea breeze, which previous studies have shown to be highly dependent on the magnitude and direction of the prevailing low-level wind. A synoptic-scale offshore (opposing) flow results in a stronger land–sea thermal gradient and greater convergence along the sea-breeze front. Opposing flow refers to winds with a component that is opposite the direction of propagation of the sea breeze. This scenario creates enhanced thunderstorm and lightning activity along the offshore coast. Conversely, onshore flow results in a weaker thermal contrast and a sea-breeze circulation that penetrates farther inland. In this case convection and lightning are suppressed along the onshore coast. In the case of the Florida peninsula, which has two coastlines, a large-scale southeasterly flow results in enhanced convection and lightning along the west (offshore) coast and suppressed activity along the east (onshore) coast, with the opposite being true for a large-scale southwesterly wind (López and Holle 1987; Camp et al. 1998; Lericos et al. 2002).

Figure 1 indicates that many heavily populated areas along the east and west coasts of Florida are vulnerable to intense lightning. Consequently, the risks for casualties, damage, and disruptions to power and communications are much greater in these areas. 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 and Light Corporation (FPL) must determine well ahead of time whether lightning is likely during the late afternoon and evening within their service areas. If a high lightning threat is perceived, extra crews must be retained after normal business hours to deal with potential disruptions. If this threat is misjudged, the company either will not be able to respond effectively to outages, or, conversely, resources could be wasted on a threat that does not occur.

The development of a lightning forecast procedure is a difficult problem. Despite the regular and predictable forcing produced by the sea breeze, summertime convection and lightning over Florida exhibit considerable spatial and temporal variability (López et al. 1984). Even if one could pinpoint the exact locations that will experience convection on a particular day, these areas may not experience the most lightning, since lightning is governed by cloud microphysical processes that are poorly resolved by numerical models. Nevertheless, 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 likelihood and amount of afternoon and evening lightning, based on past events under similar atmospheric conditions.

Many studies have found statistical models to be useful for predicting warm season thunderstorms and lightning. Some of the statistical methods that have been used include multiple linear regression, binary logistic regression, and classification and regression trees (CART) (e.g., Livingston et al. 1996; Mazany et al. 2002; Burrows et al. 2004; Brenner 2004; Lambert et al. 2005). These methods attempt to quantify the relationship between a set of predictors and the outcome of interest such as thunderstorm probability or lightning frequency (e.g., Neumann and Nicholson 1972; Reap 1994).

The present study develops a statistical scheme that provides improved forecast guidance of warm season afternoon and evening lightning for 11 areas of the Florida peninsula serviced by FPL. Logistic regression techniques are used to develop equations predicting whether at least one CG flash will occur during the noon–midnight (NM) period in each area, as well as the amount of lightning that can be expected, conditional on at least one flash occurring. The equations are derived for the warm season (June–August) when the sea breeze generally is the dominant forcing mechanism for convection and lightning. Candidate predictors for the

regression models include various wind, stability, and moisture parameters calculated from morning radiosonde data at Miami, Cape Canaveral, Jacksonville, and Tampa. Previous day persistence and same day morning lightning also are used as candidate predictors of afternoon lightning. Although the equations derived in this study are applicable to specific coastal areas of Florida, the statistical methods employed also should be useful anywhere that convective forcing is rather cyclic, such as other coastal areas having sea breezes as well as areas where diurnal topographic forcing is important.

2. Data

a. Study areas

Statistical guidance was developed for 11 coastal areas of the Florida peninsula that are serviced by FPL. The name of each forecast area (FA) and its boundaries are shown in Figs. 2a–d. These irregularly shaped areas were specified by FPL based primarily on the location and number of customers, although some meteorological factors also were considered.

b. 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, Inc., and provides both real-time and historical data to government, educational, and commercial users. 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 June–August for the years 1989–2004. 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 United States ranged from 65% to 85%, with location accuracies between 8 and 16 km. A system upgrade in 1995 allowed a greater number of flashes to be detected, as well as improved location accuracy. Since the upgrade, the NLDN has a location accuracy of ~0.5 km over most of the United States, 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 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 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 dataset. In addition, multiple flashes occurring during the same second and within 10 km of each other were assumed to be duplicate flashes, and were combined into a single flash by retaining the first flash's time and location and adding the multiplicities.

The number of daily CG flashes in each FA (Figs. 2a–d) was counted over the period of interest, NM eastern daylight time (EDT) (1600–0359 UTC). A morning flash count for 0600–1159 EDT (1000–1559 UTC) also was calculated as a potential predictor of afternoon lightning.

Figure 3 shows the hourly distribution of flash count for the Miami–Dade domain for all warm season days (June–August) during the 16-yr period. A diurnal peak occurs between 1400 and 1500 EDT, with a rapid decrease after 1900 EDT. Similar diurnal variations have been observed in previous studies (e.g., Maier et al. 1984; Reap and MacGorman 1989; Livingston et al. 1996; Lericos et al. 2002; Mazany et al. 2002). Distributions for the other 10 areas are very similar, with only slight variations in the time of peak lightning. The NM forecast period considered in this study accounts for 85%–95% of the daily total in each FA. Although early afternoon is the most active lightning period, the evening hours were included so that FPL officials can better plan whether to retain extra crews after normal business hours.

c. Radiosonde data

Morning radiosonde data for Miami (MFL), Jacksonville (JAX), Tampa (TBW), and Cape Canaveral (XMR) were used to calculate various wind, moisture, and stability parameters to 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 (2000–2004) were obtained directly from FSL's "Radiosonde Database Access" Web site (FSL 2004).

A total of 597 parameters was calculated from the radiosonde data (Table 1), many of which have been found in previous studies to be useful predictors of thunderstorms and lightning during the warm season. They include variables describing wind direction and speed, moisture, temperature, and stability. Pressure-weighted layer averages of these variables also were

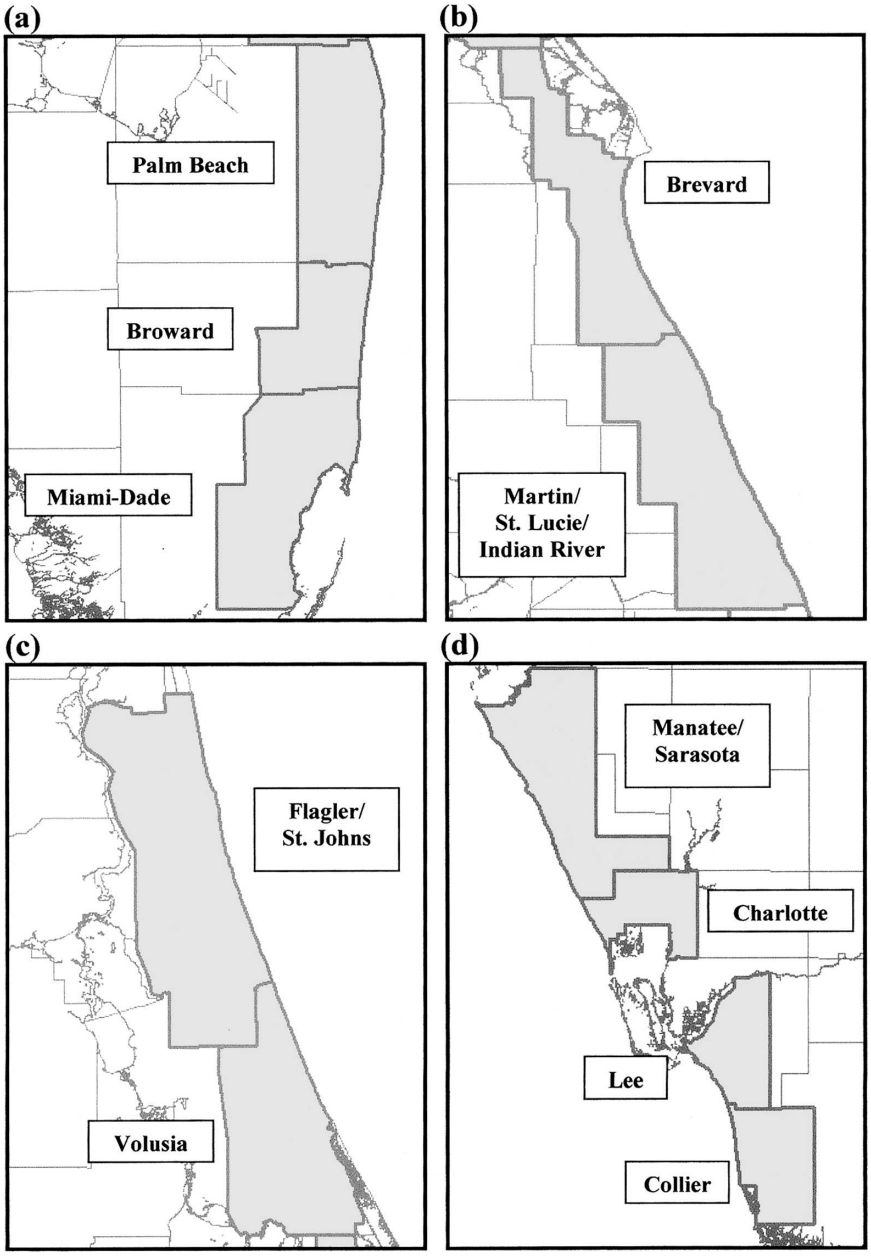

FIG. 2. Maps of each forecast area: (a) southeast coast, (b) east coast, (c) northeast coast, and (d) west/southwest coast.

calculated, as well as layer thickness and temperature lapse rate. This list of potential predictors will be greatly reduced in subsequent procedures described in section 3.

The sounding closest to each FA generally was used under the assumption that the closest sounding would be most representative of the conditions in that area. Correlations between the sounding parameters and lightning showed that this was indeed the case, with only a few exceptions. Specifically, parameters calculated from the MFL sounding generally were better correlated with lightning activity in the Charlotte and Lee areas than parameters from the closer TBW sounding. Since the subtropical ridge axis usually is located north of MFL, the low-level flow in these areas usually is from the southeast. As a result, atmospheric properties in the Charlotte and Lee areas tend to be more similar to MFL than TBW. Thus, better results would be achieved using the MFL sounding for Charlotte and Lee instead of TBW.

Figure 4 shows the sounding used for each forecast

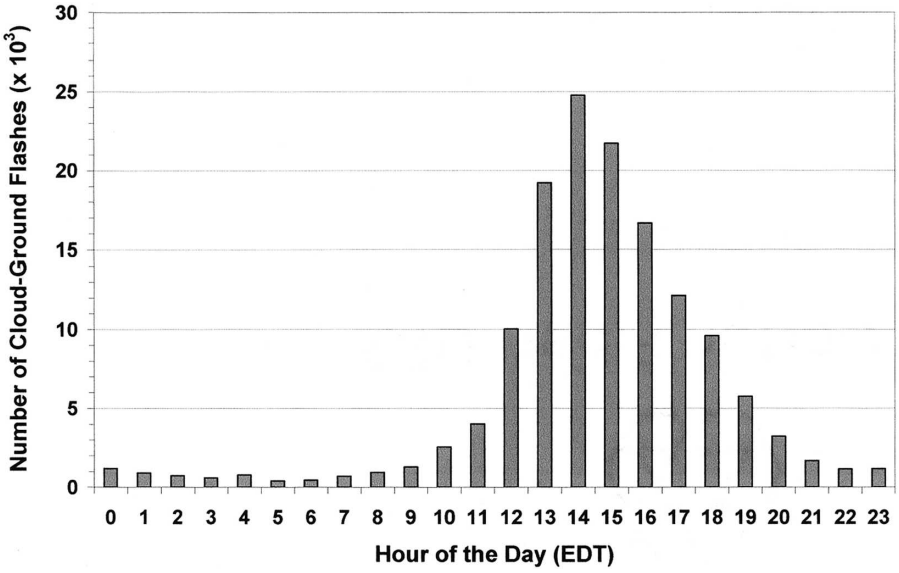

FIG. 3. Hourly distribution of CG flash count for the Miami–Dade forecast area for all warm season days (June–August) during the period 1989–2004.

area. Prior to 1995, the JAX site was located in Waycross, GA (AYS). Since the JAX soundings are more representative of conditions in the Flagler–St. Johns area than AYS, it was decided that only JAX data from 1995 onward would be used for that area. In addition, due to the poor availability of XMR soundings prior to 1992, only the 1000–1200 UTC data from 1992 onward were used for the Volusia, Martin–St. Lucie–Indian River, and Brevard FAs. For all other areas, MFL–PBI or TBW data for 1989–2004 were used.

The equations being derived are for situations when the sea breeze is the dominant forcing mechanism for convection; they are not meant for days when large-scale forcing leads to thunderstorms. Therefore, an effort was made to remove these synoptically influenced days before equation development. This was done by deleting any day whose 1000–700-hPa layer average wind speed was greater than 3σ from the climatological mean. This simple procedure does not guarantee that every synoptically disturbed day was removed. Approximately 2% of the days during the June–August period were removed by this procedure.

TABLE 1. List of radiosonde-derived parameters used as candidate predictors. Cross-shore and alongshore wind components are with respect to an average coastline orientation. Sounding data for each 25-hPa level (38 levels) also were submitted as candidate predictors.

Stability and moisture parameters	Pressure-weighted layer averages[a]
Height of the freezing level (m)	SIN (layer-averaged wind direction)
Height of the wet-bulb zero level (m)	Wind speed (kt)
K index (°C)	Cross-shore wind component (kt)
Vertical totals (°C)	Alongshore wind component (kt)
Cross totals (°C)	Relative humidity (%)
Total totals (°C)	Layer temperature lapse rate (°C km^{-1})
Severe weather threat index (SWEAT)	Layer thickness (m)
Convective temperature (°C)	
CAPE (J kg^{-1})[b]	
Modified CAPE (J kg^{-1})[c]	
Temperature at modified EL (°C)	
Precipitable water (cm)	
Lifted index (°C)[b]	
Modified lifted index (°C)[c]	
Showalter stability index (SSI) (°C)	

[a] Taken from 45 possible layers (e.g., 1000–900 hPa, 1000–800, . . . , 900–800, 900–700, . . . , etc.).
[b] Based on an unaltered surface parcel.
[c] Based on a modified parcel heated to the convective temperature.

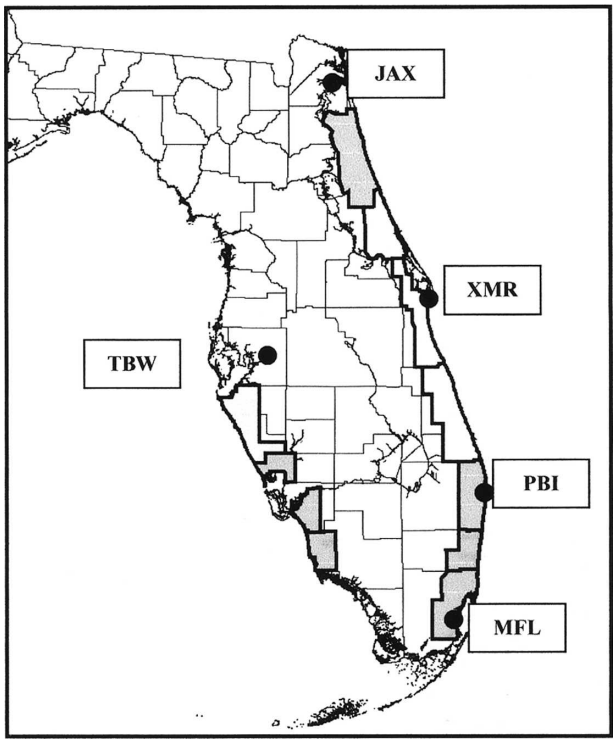

FIG. 4. Map of the Florida peninsula with each forecast area grouped by sounding. Filled circles indicate the locations of each sounding site used in the study.

d. Statistical software

Two statistical software packages were used. Most of the exploratory work was done using S-PLUS, version 6.1 for Windows, distributed by Insightful Corporation. 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. Equation development

a. Predictands

The first objective of the study was to develop statistical guidance to predict whether at least one CG flash would occur during the NM period in each FA. The output from this equation is intended to aid FPL personnel in determining which areas are most likely to experience at least some lightning during the NM period, so that appropriate preparedness measures can be taken. Since a forecast of "yes" or "no" was sought, a binary indicator was assigned to each day in the dataset; "1" if at least one CG flash was observed during NM anywhere within each area, or "0" if no activity. This binary indicator served as the predictand for the yes–no equations.

TABLE 2. The four quartiles of flash count for each forecast area.

Forecast area	Q1	Q2	Q3	Q4
Miami–Dade	1–11	12–52	53–166	>166
Broward	1–7	8–43	44–137	>137
Palm Beach	1–11	12–67	68–235	>235
Martin–St. Lucie–Indian River	1–18	19–113	114–378	>378
Brevard	1–11	12–77	78–268	>268
Volusia	1–16	17–99	100–324	>324
Flagler–St. Johns	1–18	19–115	116–404	>404
Manatee–Sarasota	1–20	21–88	89–269	>269
Charlotte	1–8	9–31	32–103	>103
Lee	1–12	13–46	47–137	>137
Collier	1–13	14–44	45–124	>124

The second objective was to develop equations to estimate the amount of lightning that would occur during the NM period, conditional on at least one flash occurring. A major decision was to determine the form of the predictand, that is, whether to forecast the actual flash count or to transform the counts into discrete categories and predict a range of counts. Our initial efforts focusing on eastern Miami–Dade and Broward Counties in south Florida showed that using multiple linear regression to estimate a flash count produced comparatively poor results. Instead, results indicated that predicting a range of flash counts was the best option. Therefore, the flash counts in each FA were grouped into four quartile categories based on climatology, with the quartiles used as the predictand. Flash count ranges for each quartile are shown in Table 2.

Rather than developing one model to forecast the quartile, the best results were achieved using separate equations to distinguish the lowest quartile of activity (Q1) from all other days, the highest quartile (Q4) from other days, and an equation to differentiate the upper two quartiles (Q3, Q4) from the lower two (Q1, Q2). Again, we sought a "yes" or "no" forecast for each of these outcomes, so three binary indicators were assigned to each lightning day (days with one or more flashes). That is, 1 was assigned to Q1 lightning events and 0 otherwise, 1 for Q4 events and 0 otherwise, and 1 for events in the upper two quartiles (Q3 or Q4) and 0 otherwise. These three equations then could be used to forecast the most likely quartile (explained in more detail in section 4).

b. Binary logistic regression

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 occurrence of at least one CG flash). BLR relates this probability to a linear combination of predictor variables, X_K, by the following relations:

$$\ln[\pi/(1 - \pi)] = f(X_K) \quad \text{and} \tag{1}$$

$$f(X_K) = b_0 + bX_1 + \ldots + b_K X_K. \tag{2}$$

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

$$\pi = \exp[f(X_K)]/\{1 + \exp[f(X_K)]\}, \tag{3}$$

and the probability of a failure (i.e., not observing at least one CG flash) is $1 - \pi$.

BLR has less stringent assumptions than linear regression. Unlike multiple linear regression, BLR does not assume a linear relationship between the independent variables and the dependent (binary) outcome. Rather, the logit function in (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 of (2). In addition, the form of (3) guarantees that BLR will always produce probability estimates bounded between zero and one inclusive (Hosmer and Lemeshow 1989).

c. Principal component analysis

It is clear that several of the parameters in Table 1 are highly correlated; in other words, they contain redundant information. For example, precipitable water is closely related to the 1000–500-hPa-layer average relative humidity, and the mean cross-shore wind component in a layer is highly correlated with the sine of the mean wind direction in that layer. Wilks (1995) cautions that estimates of the coefficients and standard errors can become unreliable and model performance can be adversely affected when highly correlated predictors compose the model. Thus, a method was needed to reduce the candidate predictors to only the most important variables without much loss of information. This was accomplished by performing a principal component analysis (PCA; Wilks 1995) on all potential sounding predictors (Table 1) using the SPSS software. PCA is a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components (PCs). In this study, the PCs were used as a classification method to cluster the highly correlated predictors into groups having physical meaning. As described in Wilks (1995), only components with eigenvalues >1 were extracted, and the sounding parameters having the greatest weights (or "loadings") on each component were grouped together.

A total of five groups generally were formed through this process. The five groups contained parameters that described either wind direction, wind speed, moisture, or stability, with a final "miscellaneous" group containing variables that were not highly correlated with those in any other group. Finally, to determine which predictors to retain for the regression analysis, one parameter was chosen from each group that was the most physically relevant and had the greatest correlation with each of the four binary predictands (described in section 3a). This procedure ensures that only the most important and nonredundant predictors are retained in the dataset for possible selection by the BLR procedure.

Table 3 lists the final set of candidate physical predictors for each FA resulting from the PCA. The most common wind parameters are the sine of the layer-averaged wind direction, the layer-averaged cross-shore wind component, and the layer-averaged speed in the low levels. The *K* index (KI) often was the most important moisture-related parameter, while total totals (TT), Showalter stability index (SSI), layer temperature lapse rate, and low-level temperature were the most important stability parameters. The physical relevance of these parameters to lightning occurrence will be discussed in section 4a.

d. Additional candidate predictors

Table 4 (top) is a 2×2 contingency table for the number of days when at least one flash was observed in the Miami–Dade domain versus what occurred the previous day. A similar 4×4 contingency table is shown for the quartiles (bottom). Nearly 82% of lightning events are correctly forecast by persistence in the Miami–Dade area, while persistence correctly forecasts the quartile nearly 36% of the time. These results show that persistence is a powerful predictor of lightning during the warm season in Florida and must be included as a candidate predictor. In addition to the parameters listed in Table 3, we included a same day morning (0600–1159 EDT) and a previous day NM indicator of at least one flash, as well as the previous day's lightning quartile and an indicator for the upper two or lower two quartiles. Since persistence typically produces a more accurate forecast than climatology, persistence will be used as the standard of reference for assessing the overall skill of the prediction equations derived in this study.

To incorporate possible nonlinear and interaction effects, power terms up to the fourth degree and two-way

TABLE 3. Final candidate physical predictors resulting from the PCA for each forecast area. Power terms up to the fourth degree and two-way cross products of these parameters as well as various forms of persistence also were included in the final predictor pool. The parameters wdir, speed, and uperp refer to the vector-averaged wind direction, speed, and component of the wind perpendicular to the coastline, respectively, in the indicated layer.

Miami–Dade	Broward	Palm Beach	Martin–St. Lucie–Indian River
Sin(1000–800 wdir)	Sin(1000–800 wdir)	Sin(1000–500 wdir)	1000–500 uperp
1000–900 speed	1000–900 speed	1000–900 speed	KI
KI	KI	KI	SSI
SSI	SSI	SSI	900–700 lapse rate
925-hPa temp	925-hPa temp	925-hPa temp	
Brevard	**Volusia**	**Flagler–St. Johns**	**Manatee–Sarasota**
1000–500 uperp	1000–900 uperp	Sin(1000–700 wdir)	Sin(900–800 wdir)
KI	500–100 speed	KI	KI
SSI	KI	SSI	TT
900–600 lapse rate	SSI	Modified LI	Lifted index
900-hPa temp	900–600 lapse rate	900–600 lapse rate	900–600 lapse rate
Charlotte	**Lee**	**Collier**	
Sin(1000–600 wdir)	Sin(1000–600 wdir)	Sin(1000–600 wdir)	
900–100 speed	800–300 speed	800–700 speed	
TT	TT	TT	
Lifted index	Lifted index	Lifted index	
500–100 lapse rate	525-hPa temp	525-hPa temp	

cross products were included as additional candidate predictors (e.g., Neumann and Nicholson 1972; Reap 1994). The power terms were calculated only for the PCA-selected physical variables, while the two-way cross products were calculated between all first-order parameters including persistence. To avoid collinearity problems among these terms, the physical variables first were normalized (i.e., subtract the mean and divide by the standard deviation) before raising them to a power or computing cross products (Wilks 1995).

e. Model building

Four logistic regression equations were derived for each of the 11 FAs. The first gave the probability of at least one CG flash occurring during the NM period in each area. Three additional logistic equations were derived to determine the most likely quartile of lightning, conditional on at least one flash occurring. One equation gave the probability of an event in the lowest quartile (Q1), one for the probability of the upper two quartiles (Q3 or Q4), and a final equation for the probability of the greatest quartile (Q4). Rather than having one equation for each quartile (four total), this three equation approach combined with a decision tree (described later) produced the best results. The logistic regression algorithm in SPSS was used to derive the equations and screen the variables (Table 3) for selection into each model.

A procedure combining forward stepwise screening and cross validation was used to derive each of the four equations. The process began by randomly dividing the dataset into two parts. One set, containing 75% of the cases, served as a "learning" sample for screening the predictors for selection. The remaining 25% were reserved as an "evaluation" sample to test the model each time a new variable 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. The first independent variable selected is that which produces the greatest reduction in the residual sum of squares (or residual deviance, RD), that

TABLE 4. A 2 × 2 contingency table (top) for the number of days when at least one CG flash was observed vs what occurred the previous day for the Miami–Dade forecast area. A similar 4 × 4 contingency table (bottom) is shown for the quartiles for cases when at least one CG flash occurred the previous day

	Previous day			
Obs	Yes	No	Tot	% correct
Yes	813	184	997	81.5
No	176	197	373	52.8
Tot	989	381	1370	73.7

	Previous day quartile					
Obs	Q1	Q2	Q3	Q4	Tot	% correct
Q1	56	58	37	24	175	32.0
Q2	51	64	39	42	196	32.6
Q3	43	49	66	55	213	31.0
Q4	30	31	64	104	229	45.4
Tot	180	202	206	225	813	35.7

is, the variable that explained the most variation in observed lightning. The algorithm then selected the next variable, which, together with the first, further reduced the RD by the greatest amount. At each step, the algorithm performed a backward check to determine if the addition of a variable caused any previously selected variable to become insignificant (i.e., if the *p* value of the variable became >0.1), in which case the insignificant variable was removed. This process continued until the RD could no longer be reduced by a significant amount, or until no other variables remained.

At each step in the sequence, a 2 × 2 contingency table was produced to show the percentage of correctly classified days for both the 75% learning sample and the 25% evaluation sample. The predictors composing the model at the step with the highest percentage of correctly classified days in the 25% evaluation sample were noted. A new random sample then was generated, and the stepwise screening was repeated. This resampling and rescreening continued until the combination of predictors was identified that most likely generalizes to independent data, and does not overfit the dependent sample. These "best" predictors identified by the resampling then were reentered for stepwise screening on the entire working dataset (no 25% and 75% random sampling) to obtain the final four logistic equations.

After final equations were obtained for the probability of the lowest quartile, upper two quartiles, and highest quartile, a decision tree was constructed to determine the most likely quartile using probability thresholds for the three equations. To produce an unbiased scheme, the thresholds were chosen so an equal number of cases was 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 scheme are given in section 4.

TABLE 5. Final logistic regression equations for the Miami–Dade forecast area. The predictors in each equation are shown with their corresponding coefficients (B), standard errors (S.E.s), and *p* values. The *p* value indicates the statistical significance of each term, or the probability that the relationship found in the sample would not also be true in the population.

Predictor	B	S.E.	*p* value
Eq. (1): Probability of at least one CG flash			
SINDIR*	−1.001	0.124	0.000
MNSPD**	−0.094	0.018	0.000
(MNSPD**)^2	0.220	0.061	0.000
SSI	−0.170	0.028	0.000
Morning yes–no	1.240	0.183	0.000
Previous day yes–no	0.919	0.154	0.000
(Previous day yes–no) × (T925)	0.297	0.104	0.004
Const	1.325	0.229	0.000
Eq. (2): Conditional probability of a Q1 event			
SINDIR*	0.659	0.091	0.000
KI^2	0.174	0.059	0.003
Morning yes–no	−0.436	0.169	0.010
Previous day quartile	−0.222	0.061	0.000
(Previous day quartile) × (MNSPD**)	0.175	0.044	0.000
(Morning yes–no) × (MNSPD**)	−0.391	0.152	0.010
Const	−0.603	0.154	0.000
Eq. (3): Conditional probability of an upper two quartile event			
SINDIR*	−1.715	0.245	0.000
(SINDIR*)^3	0.338	0.088	0.000
MNSPD**	−0.063	0.016	0.000
SSI	−0.119	0.035	0.001
KI^2	−0.176	0.069	0.011
Previous day Q3 or Q4 yes–no	0.786	0.147	0.000
Const	0.788	0.202	0.000
Eq. (4): Conditional probability of a Q4 event			
SINDIR*	−1.707	0.268	0.000
(SINDIR*)^3	0.340	0.090	0.000
KI^2	−0.310	0.114	0.007
Previous day quartile	0.263	0.067	0.000
(Previous day quartile) × (SSI)	−0.104	0.044	0.018
(Previous day quartile) × (MNSPD**)	−0.131	0.034	0.000
Const	−1.390	0.204	0.000

* For the 1000–800-hPa layer.
** For the 1000–900-hPa layer.

4. Results

a. Final logistic equations

The final equations for the 11 FAs generally are a variation on the same theme; therefore, this section only presents results for the Miami–Dade area. Table 5 displays the final equations giving the probability of at least one CG flash [Eq. (1)] and the conditional probability of the lowest quartile [Eq. (2)], the upper two quartiles [Eq. (3)], and the greatest quartile [Eq. (4)] for the Miami–Dade domain. The predictors in each equation are given along with their coefficient (B) and standard error (S.E.), as well as the *p* value, which indicates the significance of each term and its relative predictive importance. The *p* value tests 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 a few exceed the 99% level, providing strong evidence that the parameters are significant and belong in the equations.

It is informative to describe the physical significance of each parameter in the equations (Table 5) and their relationships to lightning activity. The most often selected physical predictors are the sine of the vector-averaged wind direction in the 1000–800-hPa layer (SINDIR) and the average wind speed in the 1000–900-hPa layer (MNSPD). Their selection is not surprising, since previous studies have documented that the magnitude and direction of the prevailing low-level wind with respect to the coastline have a significant influence on the strength and inland penetration of the sea breeze (e.g., López and Holle 1987; Reap 1994; Lericos et al. 2002).

In all equations except Eq. (2) the coefficient of SINDIR is negative (Table 5). Since the sine of angles between 180° and 360° is negative, an offshore, low-level wind increases the probability of afternoon lightning and increases the likelihood of a Q3 or Q4 event in the Miami–Dade area. Conversely, the positive coefficient in Eq. (2) indicates that Q1 events are less likely for offshore flow (SINDIR < 0) and more likely for onshore flow (SINDIR > 0). The coefficients for MNSPD in Eqs. (1) and (3) also are negative, suggesting that as the low-level wind speed increases, the probability of at least one flash and the likelihood of upper two quartile events decreases. This result is consistent with Camp et al. (1998) and Arritt (1993) who found that onshore wind speeds exceeding several meters per second and offshore speeds greater than 11 m s^{-1} suppress sea-breeze development in areas near the coastline. Conversely, weak offshore flow produces a strong sea-breeze circulation whose leading edge remains near the coastline. In eastern Miami–Dade County this offshore scenario can produce extensive, slow-moving thunderstorms and high flash count events if adequate moisture and instability are present.

It is interesting that a nonlinear (cubic) term with a positive coefficient was selected for SINDIR in Eqs. (3) and (4) (Table 5), in addition to the first-order term. This relationship is depicted in Fig. 5a, which plots the log-odds of a Q4 lightning event versus SINDIR if only the first-order and cubic terms in Eq. (4) are considered, that is, setting all other variables in the equation equal to zero. The figure indicates that the log-odds of a Q4 event is maximized for SINDIR between −0.65 and −0.45, with diminishing log-odds as SINDIR increases. This maximum corresponds to wind directions between 205°–220° (SW) and 320°–335° (NW). Northwest flow is uncommon during June–August in south Florida. Therefore, SW flow likely is the greatest contributor to the maximum in the log-odds of Q4 events. A SW (offshore) flow transports subtropical moisture northward into south Florida and opposes the sea breeze, producing enhanced convergence along the sea breeze and widespread thunderstorm and lightning activity in eastern Miami–Dade County.

The Showalter stability index (SSI) was selected in Eqs. (1) and (3). SSI is similar to the lifted index except the parcel is lifted from 850 hPa instead of the surface, with values becoming more negative as instability increases. The negative coefficients indicate that as instability increases, the likelihood of at least one flash and a Q3 or Q4 event increases. Studies by Livingston et al. (1996) and Lambert et al. (2005) also found SSI to be a useful predictor of afternoon lightning.

The KI appears only as a quadratic term in the three quartile equations. Figure 5b plots this quadratic relationship between the log-odds of a Q4 event and KI for the Miami–Dade domain, if all other parameters in Eq. (4) are set to zero. Clearly, the likelihood of a Q4 event increases with increasing KI until a peak is reached between 25° and 30°C. Then, the likelihood of a Q4 event decreases for larger values of KI. Since KI increases with more unstable midlevel lapse rates and greater middle-tropospheric moisture, it is reasonable that convection and lightning also will increase. The reason for decreasing log-odds for KI values greater than ~30°C is uncertain, but may be due to excess midlevel moisture and cloud cover from early morning convection (i.e., at or near the sounding time), which would tend to suppress surface heating and strong afternoon activity. The decreasing odds for larger KI values may also be due to lightning being the predictand, as opposed to updraft speed or rainfall, which generally increase with instability.

As expected, persistence was selected as a predictor of afternoon lightning in the Miami–Dade area (Table 5). For the probability of at least one flash [Eq. (1)], both the morning and previous day indicators were chosen. The positive coefficients suggest that the likelihood of at least one flash during the NM period increases if at least one flash occurred the previous day, or if at least one flash occurred from 0600 to 1159 EDT in the morning. The morning persistence indicator also appears in Eq. (2) for the probability of a Q1 event. The negative coefficient implies that Q1 events become less likely, and thus Q2 or greater events become more likely, if there was at least one flash during the morning. The previous day quartile indicator was selected in Eqs. (2) and (4), and the persistence indicator for the upper two or lower two quartiles was chosen for Eq. (3). The signs of their respective coefficients indicate that higher

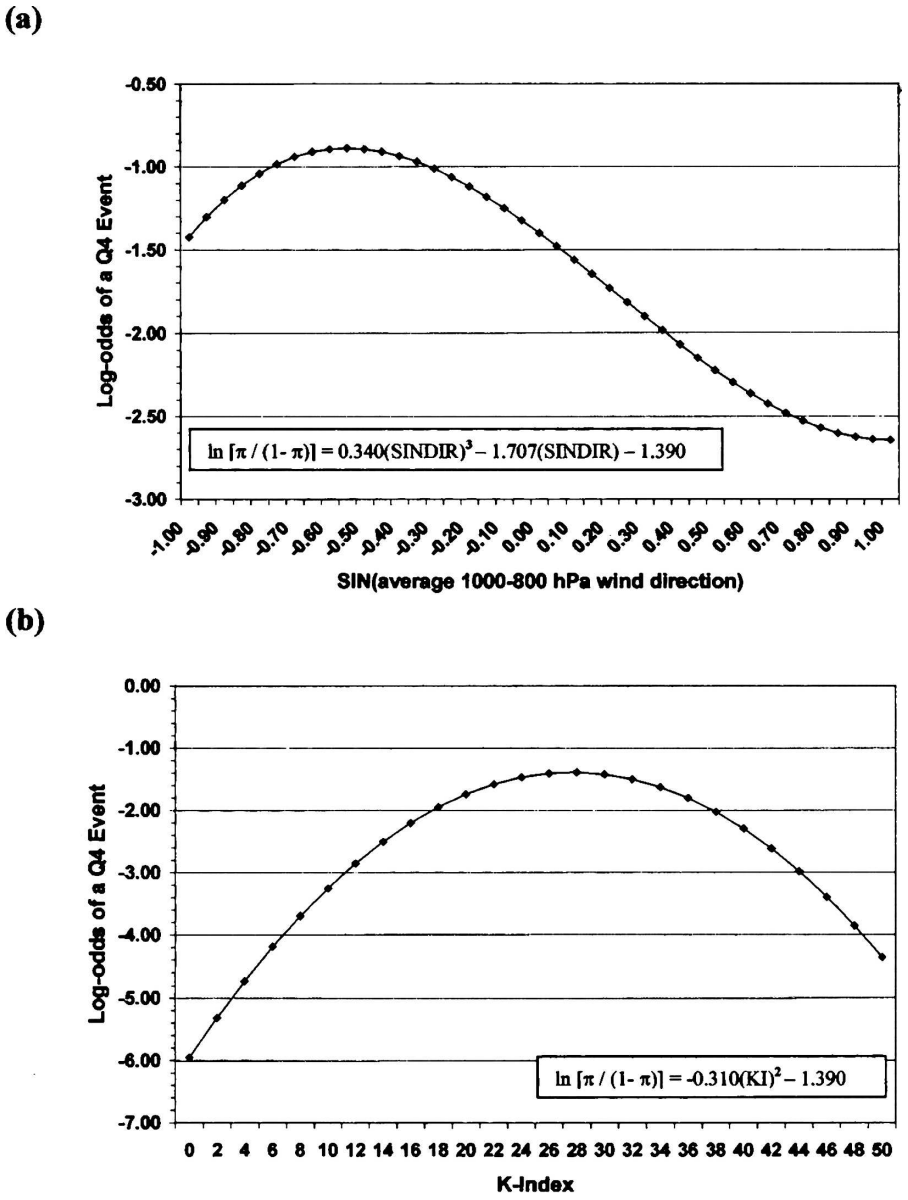

FIG. 5. Log-odds of a Q4 lightning event as a nonlinear function of (a) the sine of the 1000–800-hPa-layer average wind direction and (b) the *K* index, for the Miami–Dade forecast area. The log-odds is equivalent to the logit value obtained from (1).

flash count events are more likely if the previous day also had a high flash count. Meteorological conditions in south Florida during the warm season often change little from day to day. Thus, if conditions were favorable for lightning on the previous day, conditions on the current day often are similar. Lightning activity during the morning suggests that outflow boundaries may be present during the afternoon. These boundaries can enhance low-level convergence by interacting with the sea-breeze circulation.

Interaction terms were selected in three of the four guidance equations (Table 5). Such terms appear when the effect that one independent variable has on the response (i.e., lightning) is modulated by changes in another independent variable. For example, in Eq. (4) the effect of persistence (previous day quartile) on the likelihood of Q4 events is modulated by MNSPD and SSI. The negative coefficients suggest that decreasing values of MNSPD and SSI reinforce the positive relationship between persistence and the likelihood of Q4 events. Conversely, an increase in MNSPD or SSI counteracts the positive effect of persistence. Thus, these

TABLE 6. Sample 2 × 2 contingency table and formulas for computing skill scores.

Obs	Predicted: Yes	Predicted: No	Tot
Yes	x	y	$x + y$
No	z	w	$z + w$
Tot	$x + z$	$y + w$	$w + x + y + z$
Probability of detection	POD = $x/(x + y)$		
Overall hit rate	HR = $(x + w)/(w + x + y + z)$		
False alarm ratio	FAR = $z/(x + z)$		
Bias	$B = (x + z)/(x + y)$		
Critical success index	CSI = $x/(x + y + z)$		
Hit rate nonevents	$w/(z + w)$		

interaction terms serve to prevent persistence from having undue influence on the forecast if current atmospheric conditions are unfavorable for a Q4 event, or enhance its contribution to the forecast if conditions are favorable.

b. Results for dependent data

1) YES–NO EQUATIONS

The BLR equations provide a probability ranging between zero and one. To forecast whether at least one CG flash will occur during the NM period, a threshold probability must be determined. Then, if the calculated probability exceeds this threshold, at least one flash is forecast to occur; otherwise, no lightning is forecast. The optimum threshold was determined using verification scores from a 2 × 2 contingency table giving the number of days when at least one flash was observed compared to the number predicted using varying trial thresholds. These scores include the probability of detection (POD), hit rate (HR), false alarm ratio (FAR), bias, critical success index (CSI), and the percentage of nonlightning events correctly forecast. The POD is the ratio of the number of events correctly predicted by the model to the total number of observed events in the sample. The HR is the most direct measure of accuracy for categorical (yes–no) forecasts, indicating the percentage of correctly predicted events and nonevents. The FAR is a measure of the forecast events that fail to occur. The bias B indicates the degree of overforecasting ($B > 1$) or underforecasting ($B < 1$) an event. Finally, the CSI combines attributes of the POD and FAR, and can be viewed as a HR for the event being forecast after removing correct no forecasts from consideration (Wilks 1995). These quantities are further defined in Reap (1994) and Mazany et al. (2002). Table 6 shows a sample contingency table with formulas used in computing these scores.

Figure 6 shows how these statistics vary using different thresholds for eastern Miami–Dade County. Except for CSI, HR, and the percentage of nonevents correctly forecast, values decrease as the threshold is increased. Based on Reap (1994), we sought to maximize the CSI and POD while minimizing the FAR and capturing as many of the nonevents as possible. This latter consideration was used because results showed that the BLR scheme better forecast days with observed lightning than days without lightning. We found that the hit rate was improved by sacrificing some accuracy forecasting days with lightning in order to improve the forecasts of days without lightning. Based on the above considerations, a threshold of 50% was chosen for the Miami–Dade model. Thresholds for the other 10 forecast areas (not shown) ranged from 45% to 60%.

Table 7 shows a 2 × 2 contingency table (top) and statistics (bottom) for all 16 warm seasons of dependent data for eastern Miami–Dade County, using the optimum probability threshold of 50%. The scores are quite good, with a CSI of 77%, POD of 92%, a bias near 1, and a low FAR of 17%. Also shown is the CSI based on a persistence forecast (Table 3), along with a skill score (SS_{mod}) calculated from the model CSI and the persistence CSI,

$$SS_{mod} = [(CSI_{mod} - CSI_{pers})/(1 - CSI_{pers})] \times 100\%. \quad (4)$$

The skill score is positive (25.4%), indicating that forecasts made by the model are superior to persistence. Table 8 contains CSI and skill scores for all 11 FAs. All of the skill scores are positive, ranging from 15.6% in the Charlotte FA to 31.9% in the Volusia area. Thus, all model forecasts based on the dependent data are superior to persistence. The variations in skill score can be attributed to factors such as differences in the skill of persistence in each FA, the size of each FA, and the proximity of the radiosonde site being used.

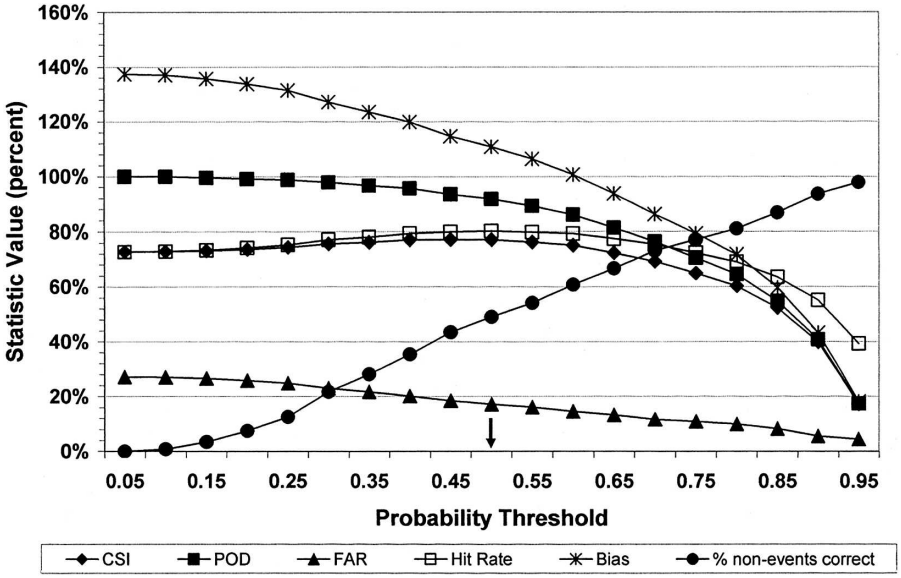

FIG. 6. Values of CSI, POD, FAR, hit rate, bias, and the percentage of nonevents correctly forecast for varying probability thresholds. These statistics are derived from the logistic model giving the probability of at least one CG flash in the Miami–Dade forecast area. The optimum probability threshold for forecasting a lightning event is marked with an arrow.

2) QUARTILE SCHEME

Once probabilities are obtained from the three quartile equations (Table 5), one must determine which quartile to forecast. Since the equations do not contain the same set of parameters, one cannot simply solve for the probability of each quartile using output from the three equations. Instead, the best results were achieved by creating a decision tree using the probability thresholds described in section 3e (e.g., Burrows et al. 2004).

TABLE 7. A 2 × 2 contingency table (top) for the number of days when at least one CG flash was observed vs the number predicted for all 16 warm seasons of dependent data for the Miami–Dade forecast area. Skill scores for the Miami–Dade area also are shown (bottom).

	Predicted		
Obs	Yes	No	Tot
Yes	915	82	997
No	190	183	373
Tot	1105	265	1370

Statistic	Value
CSI	0.77
POD	0.92
FAR	0.17
HR	0.80
Bias (B)	1.10
Nonevents	0.49
Persistence CSI	0.69
SS_{mod}	25.4%

The decision tree for the Miami–Dade FA and its resulting 4 × 4 contingency table are shown in Fig. 7 and Table 9, respectively. The first branch to the left or right depends on the probability obtained from Eq. (3), that is, distinguishing between upper two and lower two quartile events. For example, if the probability of an upper two quartile event (≥53 flashes) exceeds the threshold of 0.498, the right branch is taken and either a Q3 or Q4 event is forecast. Then, Eq. (4) is used to determine which of these two quartiles is most likely. If the probability of a Q4 event (>166 flashes) exceeds 0.372, then a Q4 event is forecast; otherwise, that day is

TABLE 8. CSI for yes–no regression model, persistence CSI, and skill score computed from (4) for all 11 forecast areas. These results are for all 16 warms seasons of dependent data.

Forecast area	Model CSI	Persistance CSI	Skill score (%)
Volusia	0.75	0.64	32.0
Flagler–St. Johns	0.77	0.66	31.6
Brevard	0.78	0.69	27.0
Collier	0.77	0.69	25.7
Miami–Dade	0.77	0.69	25.4
Palm Beach	0.76	0.68	25.3
Broward	0.74	0.65	24.5
Manatee–Sarasota	0.77	0.71	23.6
Lee	0.75	0.68	21.4
Martin–St. Lucie–Indian River	0.80	0.75	20.5
Charlotte	0.72	0.66	15.6

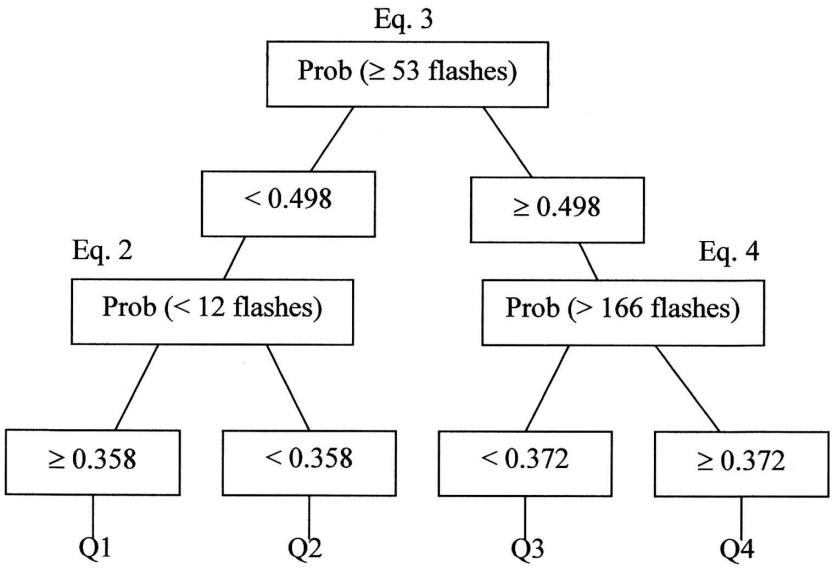

FIG. 7. Probability decision tree used to determine the predicted lightning quartile for the Miami–Dade domain.

predicted to be a Q3 event. Conversely, if the probability of the upper two quartiles is less than the threshold of 0.498, the left branch is taken and the lower two quartiles are most likely, in which case Eq. (2) determines which to predict, using a threshold of 0.358.

The overall accuracy of the quartile scheme for the Miami–Dade area can be assessed from the 4 × 4 contingency table for all 16 warm seasons of dependent data (Table 9). It is encouraging that the number of observed days in each quartile versus the number predicted is maximized along the diagonal. The scheme best forecasts Q1 and Q4 events, with hit rates of 47% and 48%, respectively. The table also reveals that Q2 events are not easily distinguished from Q1 events, and Q3 days are not easily distinguished from Q4 days. Thus, hit rates for the Q2 and Q3 quartiles are somewhat worse (31%–33%). This may be due to many days having probabilities that are very near the thresholds for being partitioned left or right at a branch of the decision tree. In addition, flash counts on many days straddle the cut point between quartiles. The probability thresholds could be adjusted to increase the detection for any quartile of choice (e.g., the Q4 events), but not without creating a bias toward that quartile.

Another measure of accuracy is the percentage of events that the scheme correctly predicts to within one quartile of the observed (Table 9). For example, when a Q1 event was observed, the scheme predicted either a Q1 or a Q2 event 78% of the time, and when a Q4 event was observed, the scheme predicted either a Q3 or a Q4 79% of the time. Considering all quartiles together, the Miami–Dade scheme correctly forecasts the quartile 40% of the time using the dependent data, and is correct to within one quartile of the observed 82% of the time.

The bottom of Table 9 shows the percentage of correctly classified events and the percentage correct to within one quartile using persistence (Table 3), as well as the skill score computed from (4). Both scores are positive, indicating that the quartile scheme is more skillful than persistence in correctly forecasting the quartile, and much more skillful than persistence at predicting to within one quartile of the observed. Tables 10 and 11 show the hit rate and percentage correct to within one quartile, respectively, for the model and persistence for all 11 FAs, as well as skill scores. In forecasting the correct quartile (Table 10), all scores are positive, ranging from 6.4% in the Charlotte area to

TABLE 9. A 4 × 4 contingency table for the number of observed days in each quartile and the number predicted using the decision tree in Fig. 7. These results are for all 16 warm seasons of dependent data for the Miami–Dade area.

	Predicted quartile					Hit	Within
Obs	Q1	Q2	Q3	Q4	Tot	rate	1 Q
Q1	117	77	34	22	250	0.47	0.78
Q2	74	82	61	33	250	0.33	0.87
Q3	39	57	77	73	246	0.31	0.84
Q4	20	33	77	121	251	0.48	0.79
Tot	250	249	249	249	997	0.40	0.82
					Persistence	0.36	0.75
					SS_{mod}	6.4%	28.7%

TABLE 10. Quartile HR for the prediction scheme, quartile HR for persistence, and skill scores for all 11 forecast areas. Results are for all 16 warm seasons of dependent data.

Forecast area	Quartile HR model (%)	Quartile HR persistence (%)	Skill score (%)
Flagler–St. Johns	46.4	32.9	20.0
Brevard	40.2	30.1	14.4
Martin–St. Lucie–Indian River	39.8	30.6	13.3
Manatee–Sarasota	39.9	32.3	11.2
Collier	39.4	31.8	11.2
Lee	38.4	30.9	10.8
Volusia	37.8	31.4	9.2
Palm Beach	39.9	33.9	9.1
Broward	38.0	31.8	9.1
Miami–Dade	39.8	35.7	6.4
Charlotte	35.3	30.8	6.4

20% in Flagler–St. Johns. However, scores for the percentage correct to within one quartile (Table 11) are much greater, ranging from 19.1% in the Lee FA to 33.8% in the Brevard area.

c. *Cross validation*

The results presented in section 4b (and those in Tables 7–11) are for all 16 warm seasons of dependent data. That is, the results show the predictive accuracy of the equations when applied to the same data that were used to derive them. These results do not assess how well the guidance equations will predict cases that were not involved in equation development. To estimate the performance of the equations on independent data, a k-fold cross-validation (CV) procedure was followed.

TABLE 11. Percentage correct to within one quartile for the prediction scheme, percentage within one quartile for persistence, and skill scores for all 11 forecast areas. Results are for all 16 warm seasons of dependent data.

Forecast area	% within 1 Q model	% within 1 Q persistence	Skill score (%)
Brevard	78.2	67.0	33.8
Martin–St. Lucie–Indian River	81.7	72.3	33.7
Flagler–St. Johns	80.4	71.1	32.3
Palm Beach	81.8	74.3	29.4
Miami–Dade	81.9	74.5	28.7
Charlotte	76.0	67.4	26.4
Volusia	78.3	70.7	26.0
Broward	77.5	70.7	23.2
Manatee–Sarasota	81.0	75.5	22.3
Collier	77.1	70.8	21.7
Lee	76.2	70.6	19.1

TABLE 12. A 2 × 2 contingency tables for the number of days when at least one CG flash was observed vs the number predicted during cross validation for the (top) Flagler–St. Johns and (bottom) Charlotte forecast areas.

Cross-validation results: Flagler–St. Johns			
	Predicted		
Obs	Yes	No	Tot
Yes	546	47	593
No	119	146	265
Tot	665	193	858
		SS_{mod}	31.0%
		Dependent data	31.6%
		Diff	−0.6%

Cross-validation results: Charlotte			
	Predicted		
Obs	Yes	No	Tot
Yes	780	82	862
No	235	165	400
Tot	1015	247	1262
		SS_{mod}	14.4%
		Dependent data	15.6%
		Diff	−1.2%

This involved withholding one warm season of data at a time for testing, while using the remaining 15 warm seasons to rederive the equations (following the same procedure outlined in section 3e). The process was repeated 16 times, once for each warm season. Since the CV procedure is both tedious and time consuming, it was performed only for the Flagler–St. Johns (FSJ) and Charlotte FAs. Since the FSJ models achieved one of the best skill scores of the 11 areas, while those for Charlotte were among the least skillful (Tables 10 and 11), it is reasonable to assume that the CV skill scores for the remaining nine areas will lie somewhere in between.

For the yes–no equations, the CV results (Table 12) for both areas produce only a slight reduction in SS_{mod} of between 0.6% and 1.2% compared to the dependent data. These scores range from 31.0% in the FSJ area to 14.4% in the Charlotte area. The quartile equations (Table 13) exhibit a somewhat larger reduction in SS_{mod} compared to the dependent data. For the hit rate, skill scores range from 16.5% in FSJ to only 1.4% in the Charlotte area, a reduction of between 3.5% and 5.0%. For the percentage correct to within one quartile of the observed, skill scores range from 26.4% in the FSJ area to 21.7% in Charlotte, a reduction ranging between 4.6% and 5.9%. These results are surprisingly good for independent data and are likely a consequence of the random sampling and testing procedure that was used to derive the original equations (section 3e). The CV

TABLE 13. Two 4 × 4 contingency tables for the number of observed days in each quartile vs the number predicted during cross-validation for the (top) Flagler–St. Johns and (bottom) Charlotte forecast areas.

Cross-validation results: Flagler–St. Johns							
	Predicted quartile						
Obs	Q1	Q2	Q3	Q4	Tot	Hit rate	Within 1 Q
Q1	82	24	22	22	150	0.55	0.71
Q2	42	47	36	21	146	0.32	0.86
Q3	27	39	57	26	149	0.38	0.82
Q4	12	22	39	75	148	0.51	0.77
Tot	163	132	154	144	593	0.44	0.79
					SS_{mod}	16.5%	26.4%
					Dependent data	20.0%	32.3%
					Diff	3.5%	−5.9%
Cross-validation results: Charlotte							
	Predicted quartile						
Obs	Q1	Q2	Q3	Q4	Tot	Hit rate	Within 1 Q
Q1	79	68	43	32	222	0.36	0.66
Q2	71	58	49	31	209	0.28	0.85
Q3	44	65	50	56	215	0.23	0.80
Q4	21	49	59	87	216	0.40	0.68
Tot	215	240	201	206	862	0.32	0.74
					SS_{mod}	1.4%	21.7%
					Dependent data	6.4%	26.4%
					Diff	−5.0%	−4.6%

results suggest that the guidance equations are statistically robust, and can be expected to provide useful guidance when implemented operationally.

5. Summary and conclusions

This study utilized 16 warm seasons of NLDN data (1989–2004) together with morning radiosonde releases from Miami, Cape Canaveral, Jacksonville, and Tampa to develop statistical lightning guidance equations for 11 areas of the Florida peninsula serviced by FPL. A total of 597 sounding parameters that previous studies have found to be useful indicators of thunderstorms and lightning during the warm season were considered as candidate predictors. These parameters describe wind direction and speed in various layers, as well as moisture, temperature, and stability. Persistence and same day morning lightning also were used as candidate predictors. A combination of stepwise screening and cross validation was used to derive logistic regression equations to predict whether at least one CG flash would occur during the NM period, as well as the amount of lightning that could be expected, conditional on the occurrence of at least one flash. Flash counts were subdivided into four quartiles based on climatology, and a decision tree scheme was used to determine the most likely quartile.

Results for the Miami–Dade domain were presented in detail. The speed and direction of the prevailing low-level wind were found to be the dominant effects in each of the guidance equations. This wind has a significant influence on the strength and inland extent of the afternoon sea-breeze circulation. Other important predictors were the *K* index and Showalter stability index, as well as morning and previous day persistence. Nonlinear and interaction effects also were found to be important. An important result is that forecasts for all 11 FAs were superior to those from persistence for both the dependent data and during cross validation. The greatest skill scores were achieved predicting whether at least one flash will occur and predicting to within one quartile category of the observed. The equations have been implemented operationally by FPL during the warm season months of June–August, and have provided useful guidance regarding the probability and amount of afternoon lightning in each of their 11 service areas.

The guidance equations derived in this study utilized parameters calculated from an appropriate morning sounding. This approach was based on several assumptions that are not valid on all days. For example, we assumed that atmospheric conditions do not vary significantly from the sounding time (0800 EDT) through the end of the forecast period (2359 EDT). This as-

sumption is approximately valid most of the time over Florida during the warm season, but sometimes is violated if a different air mass is advected into the area. We also assumed that atmospheric conditions at the radiosonde site are representative of those in the entire FA, which may not be true, even during the warm season. Whenever these assumptions are not met, errors in the lightning forecast will result. It also is clear that factors not considered in this study have an important influence on the likelihood and amount of lightning in each area. These include outflow boundaries from preexisting storms, and the interaction of smaller-scale circulations such as lake/river breezes with the sea breeze (e.g., Laird et al. 1995; Rao and Fuelberg 2000). 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.

Future work will seek to address the above limitations by utilizing mesoscale model output to create spatial forecast fields of lightning probability and amount for the entire Florida peninsula and panhandle. The model forecast data will be more location and time specific than a static morning sounding at one location. The incorporation of model-derived cloud microphysics hopefully can be related to charging mechanisms and lightning occurrence. The forecasts resulting from these improvements are expected to be more useful than those described here.

Despite the limitations, the results show how remarkably well one can predict afternoon lightning for areas as small as half a county by using input from only a morning sounding. Although the statistical guidance derived in this study was developed for specific areas of Florida for a commercial user, similar techniques could be used to develop guidance for any coastal region where (and when) the sea breeze provides the dominant forcing mechanism for thunderstorms. In addition, these techniques could be used in areas such as the western United States where diurnal topographic forcing is important for storm initiation.

Acknowledgments. This research was funded by Florida Power and Light Corporation. We appreciate the assistance of Justin Winarchick and Geoffrey Stano at The Florida State University. Appreciation is also extended to Irv Watson (NWS Tallahassee) and Dr. Pablo Santos (NWS Miami) for their ideas and suggestions. Finally, Paul Hebert and Ira Brenner at Florida Power & Light Corporation deserve special thanks for their many suggestions and their extensive knowledge of summertime sea-breeze weather patterns in Florida.

REFERENCES

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

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., C. Price, and L. J. Wilson, 2004: Statistical models for 1–2 day warm season lightning prediction for Canada and the northern United States. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 1.5.

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.

FSL, cited 2004: Radiosonde database access. [Available online at http://raob.fsl. noaa.gov.]

FSL–NCDC, 1999: *Radiosonde Data of North America 1946–1999.* CD-ROM, Version 1.0. [Available from DOC/NOAA/OAR, Forecast Systems Laboratory, R/FSL, 325 Broadway, Boulder, CO 80305.]

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

Laird, N. F., D. Kristovich, R. Rauber, H. Ochs III, and L. Miller, 1995: The Cape Canaveral sea and river breezes: Kinematic structure and convective initiation. *Mon. Wea. Rev.,* **123,** 2942–2956.

Lambert, W. C., M. Wheeler, and W. Roeder, 2005: Objective lightning forecasting at Kennedy Space Center and Cape Canaveral Air Force Station using cloud-to-ground lightning surveillance system data. Preprints, *Conf. on Meteorological Applications of Lightning Data,* San Diego, CA, Amer. Meteor. Soc., CD-ROM, 4.1.

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., 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.

——, 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.

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., S. Businger, S. I. Gutman, and W. Roeder, 2002: A lightning prediction index 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, *Int. Conf. on Aerospace and Aeronautical Meteorology,* Washington, DC, 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.

Rao, P. A., and H. E. Fuelberg, 2000: An investigation of convection behind the Cape Canaveral sea-breeze front. *Mon. Wea. Rev.,* **128,** 3437–3458.

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., cited 2004: 1989–2002 Florida lightning climatology. [Available online at http://bertha.met.fsu.edu/~jstroupe/flclimo.html.]

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