P1.2

A STATISTICAL PROCEDURE TO FORECAST WARM SEASON LIGHTNING OVER PORTIONS OF THE FLORIDA PENINSULA

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

1. INTRODUCTION

In a typical year, cloud-to-ground (CG) lightning exceeds both tornadoes and hurricanes in causing weather related fatalities across the United States (U.S.) (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 will be helpful 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 U.S. 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 deservedly has been labeled the "lightning capital" of the U.S.

Figure 1 shows the spatial distribution of CG lightning in Florida for May-September during the 14-year period 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 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 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 (e.g., López and Holle 1987, Camp et al. 1998, Lericos et al. 2002).

Lightning related 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 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, since summertime convection and lightning over Florida often 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 (MLR), binary logistic

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

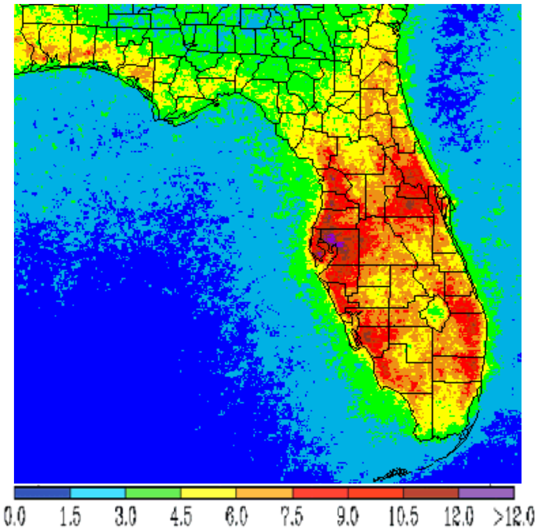


FIG. 1. Map of the spatial distribution of warm season CG lightning (flashes km^{-2} warm season^{-1}) for the state of Florida during a 14-year period from 1989-2002. Map obtained from http://bertha.met.fsu.edu/~jstroupe/flclimo.html.

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 for warm season afternoon and evening lightning for eleven areas of the Florida Peninsula that are serviced by FP&L. 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.

2. DATA

a. *Study areas*

Statistical guidance was developed for eleven coastal areas of the Florida peninsula that are serviced by FP&L (Figure 2). These irregularly shaped areas were specified by FP&L 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 U.S. The NLDN is owned and operated by Vaisala Inc., providing both real-time and historical data. 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 U.S. ranged from 65%-85%, with location accuracies between 8 km-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 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 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

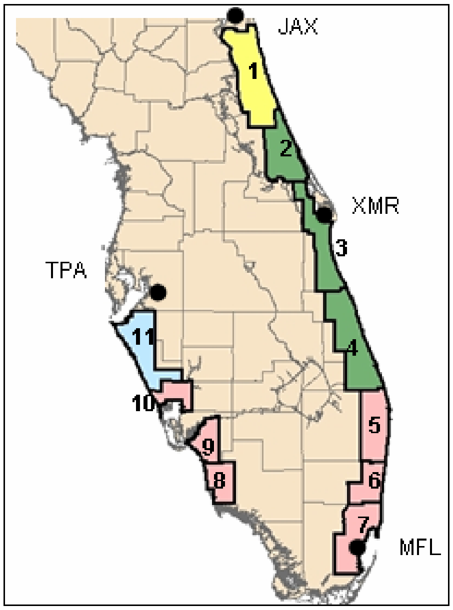


FIG. 2. Map of the eleven forecast areas grouped by sounding: 1) Flagler/ St. Johns, 2) Volusia, 3) Brevard, 4) Martin/ St. Lucie /Indian River, 5) Palm Beach, 6) Broward, 7) Miami-Dade, 8) Collier, 9) Lee, 10) Charlotte, and 11) Manatee/ Sarasota. Red shading denotes areas where the MFL sounding was used, blue shading TPA, green XMR, and yellow shading is JAX.

single flash by retaining the first flash's time and location and adding the multiplicities.

The number of daily CG flashes in each forecast area (FA) (Fig. 2) was counted over the period of interest, NM local time (LT) (1600-0359 UTC). A morning flash count for 0600-1159 LT (1000-1559 UTC) also was calculated as a potential predictor of afternoon lightning.

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 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 (http://raob.fsl.noaa.gov).

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, including wind direction and speed, moisture, temperature, and stability. Pressure weighted layer averages of these variables also were calculated (Table 1).

The sounding closest to each FA generally was used under the assumption that the closest sounding is 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 were better correlated with lightning in the Charlotte and Lee County areas than parameters from the closer TBW sounding. Since the sub-tropical ridge axis usually is located north of MFL, the low-level flow in these areas generally 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.

The sounding used for each forecast area is indicated in Fig. 2. 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 onwards would be used for that area. In addition, due to poor availability of XMR soundings prior to 1992, only the 1000-1200 UTC data from 1992 onwards were used for the Volusia, Martin/ St. Lucie/ Indian River, and Brevard FAs. For all other areas, MFL 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 discarding any day whose 1000-700 hPa layer average wind speed was greater than 3 standard deviations from the climatological mean. However, this simple procedure does not guarantee that every synoptically disturbed day was removed.

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

Table 1. Radiosonde-derived parameters used as candidate predictors. Cross-shore and along-shore wind components are with respect to an average coastline orientation.

Data for each 25 hPa level (38 levels):	Stability & Moisture Parameters:
Temperature (°C)	Height of the freezing level (meters)
Dew point (°C)	Height of the wet bulb zero level (meters)
Relative humidity (percent)	K-index (°C)
Wind speed (knots)	Vertical Totals (°C)
Height (meters)	Cross Totals (°C)
Cross-shore wind component (knots)	Total Totals (°C)
Along-shore wind component (knots)	Severe Weather Threat Index (SWEAT)
	Convective temperature (°C)
Pressure-weighted layer averages^{+}	CAPE (J/kg) *
SIN (layer average wind direction)	Modified CAPE (J/kg) **
Wind speed (knots)	Temperature at modified EL (°C)
Cross-shore wind component (knots)	Precipitable water (cm)
Along-shore wind component (knots)	Lifted index (°C) *
Relative humidity (percent)	Modified lifted index (°C) **
Layer temperature lapse rate (°C / km)	Showalter Stability index (SSI) (°C)
Layer thickness (meters)	

$^{+}$ 45 possible layers (e.g., 1000-900, 1000-800,…, 900-800, 900-700,…, etc.)
* Based on unaltered surface parcel
** Based on modified parcel heated to the convective temperature

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. 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.

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, i.e., 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 MLR 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).

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

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:

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

$$f(X_K) = b_0 + b_1X_1 + . . . + b_KX_K. \tag{2}$$

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 then is 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 MLR, 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 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 comprise 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 four to six groups were formed through this process. These groups contained parameters that described 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 non-redundant predictors are retained in the dataset for possible selection by the BLR procedure.

The most common wind parameters resulting from the PCA for each FA 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), and mid-level temperature lapse rate 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

Contingency tables using a persistence forecast (not shown) indicated that persistence is a powerful predictor of lightning during the warm season in Florida, and must be included as a candidate predictor. We included a same day morning (0600-1159 LT) and 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. Also, since persistence typically produces a more accurate forecast than climatology, persistence will be the standard of reference for assessing the overall skill of the equations derived in this study.

Non-linear and interaction effects were included in the candidate predictor pool by computing standardized 2-way cross products and power terms up to the fourth degree (e.g., Neumann and Nicholson 1972; Reap 1994). The power terms were calculated only for the PCA-selected physical variables, while the 2-way cross products were calculated between all first order parameters including persistence.

e. Model building

Four logistic regression equations were derived for each of the eleven FAs, using the PCA selected sounding predictors, as well as the non-linear and interaction terms. 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 the occurrence at least one flash (described in section 3a). Rather than having one equation for each quartile (4 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 for selection into each model.

A procedure combining forward stepwise screening and cross-validation was used to derive each of the four equations. Variables were selected for inclusion in each equation using a "forward conditional" stepwise selection algorithm, with a test for backward elimination. The stepwise screening and cross-validation procedures are described in detail in Shafer and Fuelberg (2005), and will not be repeated here.

The stepwise screening and cross-validation procedure was performed to identify the combination of predictors that most likely generalizes to independent data, and does not over-fit the dependent sample. The "best" predictors identified through this process then were re-entered for stepwise screening on the entire working dataset 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 are given in section 4.

4. RESULTS

a. Final logistic equations

The final equations for the eleven FAs generally are a variation on the same theme; therefore, this section only presents results for the Miami-Dade area. Table 3 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-

Table 3. Final logistic regression equations for the Miami-Dade forecast area.

Eq. 1: Probability of at least one CG flash

Predictor	B	S.E.	Wald	p-value
SSI	-0.170	0.028	35.761	0.000
SINDIR*	-1.001	0.124	65.064	0.000
MNSPD**	-0.094	0.018	25.969	0.000
$(\text{MNSPD**})^2$	0.220	0.061	12.978	0.000
Morning yes/no	1.240	0.183	46.055	0.000
Previous day yes/no	0.919	0.154	35.814	0.000
(Prev day yes/no) x (T925)	0.297	0.104	8.191	0.004
Constant	1.325	0.229	33.575	0.000

Eq. 2: Conditional probability of a Q1 event

Predictor	B	S.E.	Wald	p-value
SINDIR*	0.659	0.091	52.435	0.000
KI^2	0.174	0.059	8.634	0.003
Morning yes/no	-0.436	0.169	6.632	0.010
Previous day quartile	-0.222	0.061	13.424	0.000
(Prev. day quartile) x (MNSPD**)	0.175	0.044	15.628	0.000
(Morning yes/no) x (MNSPD**)	-0.391	0.152	6.644	0.010
Constant	-0.603	0.154	15.407	0.000

Eq. 3: Conditional probability of upper two quartile event

Predictor	B	S.E.	Wald	p-value
KI^2	-0.176	0.069	6.515	0.011
SSI	-0.119	0.035	11.186	0.001
SINDIR*	-1.715	0.245	49.014	0.000
$(\text{SINDIR*})^3$	0.338	0.088	14.688	0.000
MNSPD**	-0.063	0.016	15.532	0.000
Prev. day Q3 or Q4 yes/no	0.786	0.147	28.413	0.000
Constant	0.788	0.202	15.192	0.000

Eq. 4: Conditional probability of a Q4 event

Predictor	B	S.E.	Wald	p-value
KI^2	-0.310	0.114	7.335	0.007
SINDIR*	-1.707	0.268	40.645	0.000
$(\text{SINDIR*})^3$	0.340	0.090	14.208	0.000
Previous day quartile	0.263	0.067	15.567	0.000
(Prev. day quartile) x (SSI)	-0.104	0.044	5.575	0.018
(Prev. day quartile) x (MNSPD**)	-0.131	0.034	14.960	0.000
Constant	-1.390	0.204	46.248	0.000

* 1000-800 hPa layer ** 1000-900 hPa layer

Dade domain. The predictors in each equation are given along with their coefficient (B) and standard error, as well as other statistics that indicate the significance of each term and its relative predictive importance. Hosmer and Lemeshow (1989) give detailed descriptions of these statistics. The p-values in Table 3 indicate that all of the coefficients exceed the 95% significance 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 3) 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 has a significant influence on the strength and inland penetration of the sea-breeze, and thus, on the location and amount of CG lightning (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 3). 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 m s^{-1} 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.

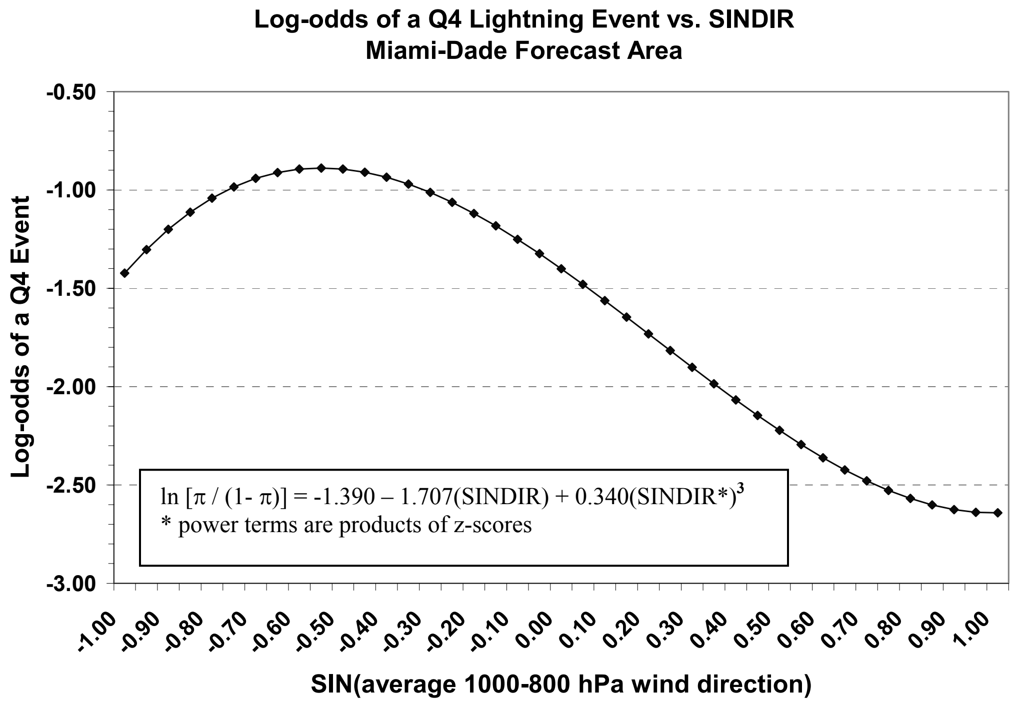
It is interesting that a non-linear (cubic) term with a positive coefficient was selected for SINDIR in Eqs. 3 and 4 (Table 3), in addition to the first order term. This relationship is depicted in Fig. 3a, which plots the log-odds of a Q4 lightning event versus SINDIR if only the first order and cubic terms in Eq. 4 are considered, i.e., setting all other variables in the equation equal to zero. The figure indicates that the log-odds of a Q4 event are 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 in South Florida during June-August. Therefore, SW flow likely is the greatest contributor to the maximum in the log-odds of Q4 events. A SW (offshore) flow transports sub-tropical 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 K-index (KI) appears only as a quadratic term in the three quartile equations. Figure 3b 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 25-30°C. Then, the likelihood of a Q4 event decreases for larger values of KI. Since KI increases with more unstable mid-level 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 mid-level 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.

As expected, persistence was selected as a predictor of afternoon lightning in the Miami-Dade area (Table 3). 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 0600-1159 LT in the morning.

(a)



(b)

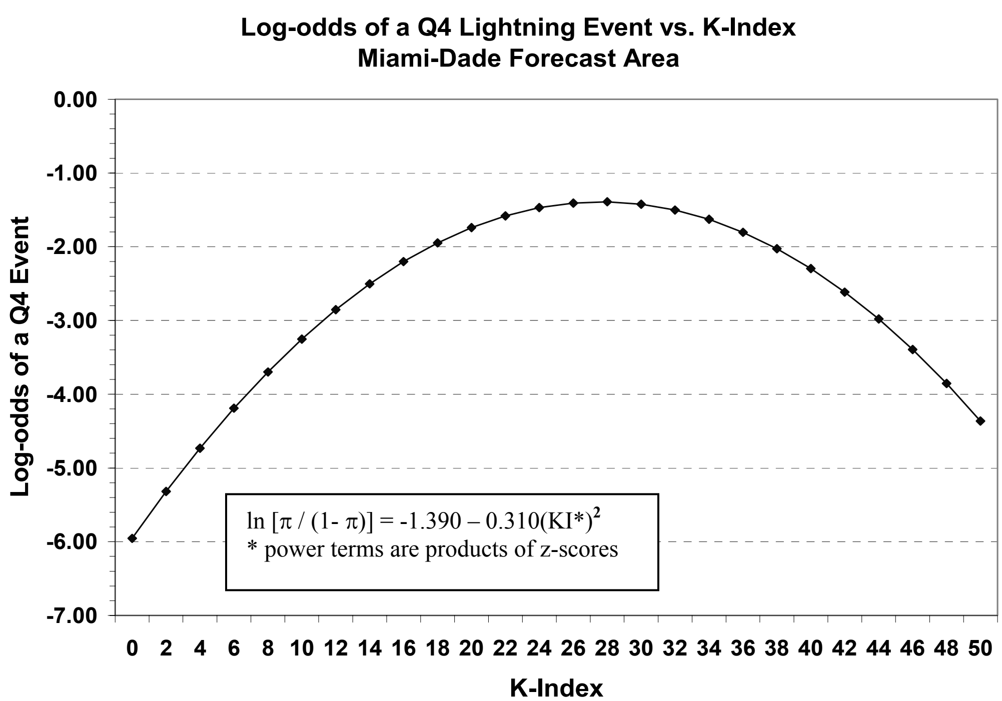


FIG. 3. Log-odds of a Q4 lightning event as a non-linear function of a) the sine of the 1000-800 hPa layer average wind direction and b) K-index, for the Miami-Dade forecast area.

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 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 for the Miami-Dade area (Table 3). 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 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 x 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 non-lightning events correctly forecast (defined in Reap 1994 and Mazany et al. 2002). Table 4 shows a sample contingency table with formulas used in computing these scores.

Table 4. Sample 2 x 2 contingency table and formulas for computing skill scores.

	Predicted		
Observed	Yes	No	Total
Yes	x	y	$x + y$
No	z	w	$z + w$
Total	$x + z$	$y + w$	$w + x + y + z$

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 non-events: $w / (z + w)$

Figure 4 shows how these statistics vary using different thresholds for eastern Miami-Dade County. Except for CSI, HR, and the percentage of non-events 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 non-events 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 ten forecast areas ranged from 45% to 60%.

Table 5 shows a 2 x 2 contingency table and statistics for all 16 warm seasons of dependent data for eastern Miami-Dade County, using the optimum probability threshold of 50%. The scores

Statistics for Yes/No Equation: Miami-Dade Area

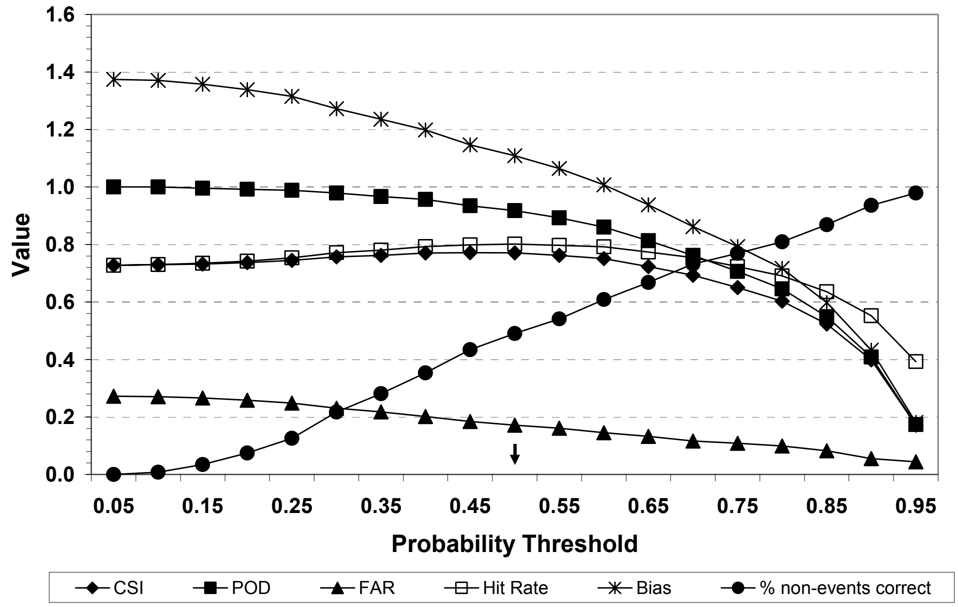


FIG. 4. Values of CSI, POD, FAR, hit rate, bias, and the percentage of non-events correctly forecast for varying probability thresholds for the Miami-Dade forecast area. The optimum probability threshold is marked with an arrow.

==

are quite good, with a CSI of 77%, POD of 92%, a bias near 1, and a low FAR of 17%. Also shown is a skill score (SS_{mod}) calculated from the model CSI and the persistence CSI:

$$SS_{mod} = [(CSI_{mod}–CSI_{pers}) / (1–CSI_{pers})]*100\% \qquad (4)$$

The skill score is positive (25.4%), indicating that forecasts made by the model are superior to persistence. The skill scores for the remaining ten FAs (not shown) also 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, size of the FA, and the proximity of the radiosonde site being used.

2) *Quartile scheme*

Once probabilities are obtained from the three quartile equations (e.g., Eqs. 2-4 in Table 3), one must determine which quartile to forecast. Since the equations do not contain the same parameters, one cannot simply solve for the probability of each quartile using output from the three equations. Instead, best results were achieved by creating a decision tree using the probability thresholds described in section 3e (e.g., Burrows et al. 2004).

The decision tree for the Miami-Dade FA and its resulting 4 x 4 contingency table are shown in Figure 5 and Table 6, respectively. The first branch in the tree depends on the probability from Eq. 3, i.e., 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 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 x 4 contingency table for all 16 warm seasons of dependent data (Table 6). It is encouraging that the number of observed days in each quartile versus the number predicted is

Table 5. 2 x 2 contingency table for the number of days when at least one CG flash was observed and the number predicted for all 16 warm seasons of dependent data for the Miami-Dade area. Also shown are the CSI and skill scores for all eleven forecast areas, sorted by skill score.

	Predicted		
Observed	Yes	No	Total
Yes	**915**	82	997
No	190	**183**	373
Total	1105	265	1370

Probability of detection:	0.92
Overall hit rate:	0.80
False alarm ratio:	0.17
Bias:	1.10
Critical success index:	0.77
Hit rate non-events:	0.49
SS_{mod}:	**25.4%**

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 Q4s), but not without creating a bias toward that quartile.

Another measure of accuracy is the percentage of time that the scheme correctly predicts to within one quartile of the observed (Table 6). 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 6 shows the percentage of correctly classified events and the percentage correct to within one quartile using persistence, as well as the skill score computed from (4). Both SS_{mod} 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. The same is true for the remaining ten FAs (not shown). In forecasting the correct quartile, all SS_{mod} scores are positive, ranging from 6.4% in the Charlotte area to 20% in Flagler/ St. Johns. However, scores for the percentage correct to within one quartile 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 5 and 6) 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. This involved withholding one warm season of data at a time for testing, while using the remaining 15 warm seasons to re-derive the equations (following the same procedure mentioned 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 eleven areas, while Charlotte was one of the least skillful, 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 (Tables 7) for both areas produce only a slight reduction in SS_{mod} of between 0.6%-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 8) 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

Quartile Decision Tree: Miami-Dade Forecast Area

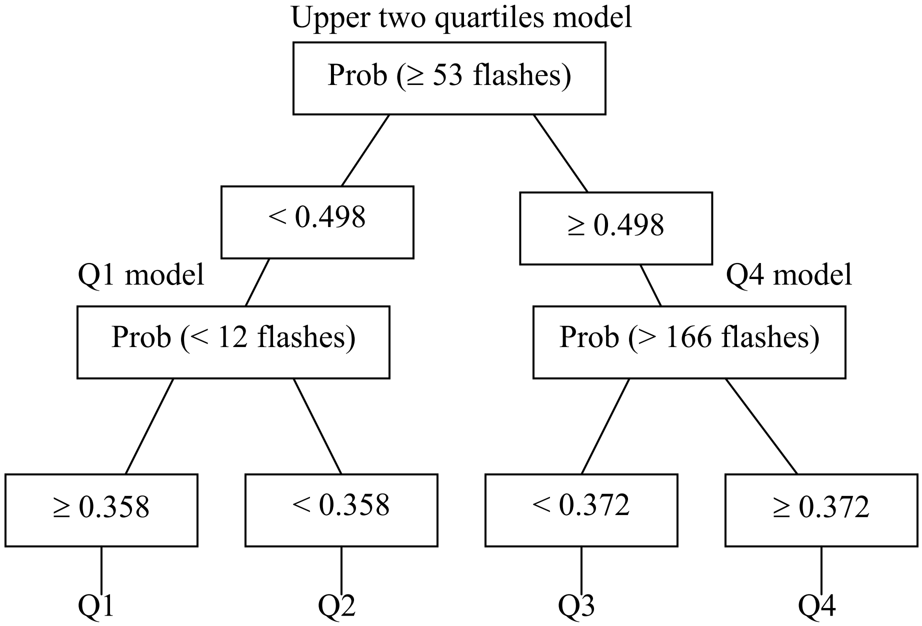


FIG. 5. Probability decision tree used to determine the predicted lightning quartile for the Miami-Dade domain.

Table 6. 4 x 4 contingency table for the number of observed days in each quartile and the number predicted using the decision tree in Fig. 5. These results are for all 16 warm seasons of dependent data for the Miami-Dade area.

4 x 4 Contingency Table & Stats for Miami-Dade Area

		Predicted					
Observed	Q1	Q2	Q3	Q4	Total	Hit rate	Within 1Q
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
Total	250	249	249	249	997	**0.40**	**0.82**
					Persistence:	0.36	0.75
					SS_{mod}:	**6.4%**	**28.7%**

Table 7. 2 x 2 contingency tables for the number of days when at least one CG flash was observed and the number predicted during cross-validation for the Flagler/St. Johns (top) and Charlotte forecast areas (bottom).

==

	Predicted		
Observed	Yes	No	Total
Yes	**546**	47	593
No	119	**146**	265
Total	665	193	858

SS_{mod}: **31.0%**
Dependent data: **31.6%**
Difference: **-0.6%**

	Predicted		
Observed	Yes	No	Total
Yes	**780**	82	862
No	235	**165**	400
Total	1015	247	1262

SS_{mod}: **14.4%**
Dependent data: **15.6%**
Difference: **-1.2%**

==

3.5%-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%-5.9%. These results are surprisingly good for independent data and are likely a consequence of the random sampling and testing methodology that was used to derive the original equations (section 3e). The CV results suggest that the guidance equations are statistically robust, and can be expected to provide useful guidance when implemented operationally by FP&L.

5. SUMMARY & CONCLUSIONS

This study has 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 eleven areas of the Florida peninsula serviced by FP&L. A total of 597 sounding parameters that previous studies have found to be useful indicators of thunderstorms and lightning during the warm season in Florida 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 at least one flash occurring. Flash counts were sub-divided into four quartile categories 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 was found to be the dominant factor 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 K index and Showalter Stability Index, as well as morning and previous day persistence. Non-linear and interaction effects also were found to be important. An important result is that forecasts for all eleven 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 amount.

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 (8 AM LT) through the end of the forecast period (midnight). This assumption 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 pre-existing 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 in this study.

Despite these limitations, the current results show how remarkably well one can predict afternoon lightning over Florida for areas as small as half a county using input from just a morning sounding. Future work will seek to improve the current results by utilizing mesoscale model output to create spatial forecast fields of lightning probability and amount for the entire FP&L service territory (to include the entire state of Florida). 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 accurate than those described here.

6. ACKNOWLEDGEMENTS

This research was funded by Florida Power & Light Corporation. We appreciate the assistance of Justin Winarchick and Geoffrey Stano at Florida State University. Appreciation also is extended to Irv Watson (NWS Tallahassee) and Dr. Pablo Santos (NWS Miami) for their ideas and

Table 8. 4 x 4 contingency tables for the number of observed days in each quartile and the number predicted during cross-validation for the Flagler/St. Johns (top) and Charlotte forecast areas (bottom).

Cross-validation results for Flagler/St. Johns area: Quartile scheme

	Predicted						
Observed	Q1	Q2	Q3	Q4	Total	Hit rate	Within 1Q
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
Total	163	132	154	144	593	**0.44**	**0.79**

SS_{mod}:	**16.5%**	**26.4%**
Dependent data:	**20.0%**	**32.3%**
Difference:	**-3.5%**	**-5.9%**

Cross-validation results for Charlotte area: Quartile scheme

	Predicted						
Observed	Q1	Q2	Q3	Q4	Total	Hit rate	Within 1Q
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
Total	215	240	201	206	862	**0.32**	**0.74**

SS_{mod}:	**1.4%**	**21.7%**
Dependent data:	**6.4%**	**26.4%**
Difference:	**-5.0%**	**-4.6%**

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.

7. 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., 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.

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

FSL and 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 & Sons, Inc., 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, *Conference on Meteorological Applications of Lightning Data*, San Diego, CA, Amer. Meteor. Soc.

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.

Mazany, R. A., Businger, S., Gutman, S. I., and Roeder, W., 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, *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.

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.

Shafer, P. E., and H. E. Fuelberg, 2005: A statistical procedure to forecast the daily amount of warm season lightning in South Florida. Preprints, *Conference on Meteorological Applications of Lightning Data*, San Diego, CA, Amer. Meteor. Soc.

Stroupe, J. R., 2003: *1989-2002 Florida Lightning Climatology*. Available [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.