STATISTICAL ANALYSIS AND MODELING OF LIGHTNING

R. D. Wooten and C. P. Tsokos
Department of Mathematics and Statistics
University of South Florida
Tampa, Florida 33620, USA

ABSTRACT. Florida is the lightning capital of the United States. Lightning strikes occur when electrostatic energy within storm conditions is unbalanced and ephemeral discharges of static electricity are set off to help the system find equilibrium. Lightning, meaning the number of lightning strikes per month, is characterized by relative humidity, sea level pressure, sea surface temperature, rain, precipitable water and the outgoing long-wave radiation. In the present study we use real data to identify the probability distribution that characterizes the behavior of the number of lightning strikes, develop a statistical model that identifies that key attributable variables to the subject strikes along with attributing interactions and proceed to estimate the number of lightning strikes with an acceptable degree of confidence. The result of the present study can be effectively used for strategic protection planning, among others.

Keywords: Parametric Analysis, Modeling and Least-Squares Regression: Linear and Non-linear.

1. INTRODUCTION

Lightning in the state of Florida is a significant event or phenomenon that we must make every effort to monitor and understand. Although this paper concentrates on the state of Florida, similar methodology, modeling and procedures can be applied to other states and generalized regions where lightning is a factor. Lightning causes deaths, wildfires (Rorig, Ferguson, Werth and Goodrick, 2005), power outages, among others.

The purpose of the present study is to first use parametric analysis to identify the most appropriate probability distribution that characterizes the behavior of the number of lightning that occurs in the state of Florida. Having identified the probability distribution we will be in a position to probabilistically characterize the behavior of this random phenomenon; that is, probabilistically characterize the number of lightning strikes per month. Furthermore, we will obtain confidence limits on the true (unknown) number of lightning in a given month, among others.

Next, the primary objective of this study is to use historical data collected by the National Lightning Detection Network (NLDN) in conjunction with other meteorological phenomenon to determine the true number of cloud-to-ground lightning over the state of Florida. To identify and rank the contributing entities or explanatory variables (independent variables) which cause lightning strikes to occur (Fieux, Paxton, Stano and DiMarco, 2005). Of special importance is identifying the interaction of the attributing variables to the subject matter.

This will be accomplished by developing statistical models and making inferences through hypothesis testing to identify and rank the various entities such as **temperatures** T_i at various levels (hPa), **relative humidity** (rh_i), etc., that contribute to lightning. In ancient Romans times writing by Lucretius (58 B.C.) discusses the coexistence of lightning and precipitation. However, few studies have statistically examined the relationship between the **number of lightning strikes** (N) and the amount of **precipitable water** (pw) in the atmosphere.

Recent studies (Sheridan, Griffiths and Orville, 1997) modeling cloud-to-ground lightning strikes (CG) with sounding parameters: convective potential energy (CAPE), precipitable water, lifted index (LI), Showalter index (SI), K – index (KI) and total totals index (TT). These indexes are empirically based and are not governed by physics and statistical inference. The Showalter index is the difference in temperature measured at 500mb and the temperature of a parcel lifted dry adiabatically for 850mb to its condensation level and moist adiabatically to 500mb. This assumes no moisture and no heat transfer as well as assuming that the temperatures between 850mb and 500mb are the most significant and does not measure the true range of temperatures within a column of air. Its advantage is that it is a single function and not the maximum of the temperature differences at various levels. This type of definition is similar for all the

indexes measured. Hence, in the present study, we will consider these contributing temperatures by level, including the maximum range of temperatures within a column of atmosphere in conjunction with precipitable water.

In the present study, we first perform parametric inferential analysis on the subject response, namely lightning; that is, the number of lightning strikes per month. Using statistical goodness-of-fit methods along with maximum likelihood estimates (MLE) we determine that the Weibull probability distribution best characterizes the probabilistic behavior of the subject phenomenon. Having knowledge of the probabilistic nature of the number of lightning strikes, enables us to estimate return period for the peak number of lightning strikes in a given month as well as estimates of the expected number of strikes along with the standard error and confidence intervals at an acceptable level of confidence. Such information could be used as measures of lightning detection, protection and planning.

In introducing the subject analysis, we will use the technique of bootstrapping; implementing this technique, we will obtain a more reliable estimate of the true (unknown) average number of lightning strikes. This technique also allows us to monitor the mean estimate as the same size increases; that is, determine the convergence of the sample estimate, $\bar{x}$ as the sample size, n, increase and simultaneously reduce the standard error.

Secondly, we use real data gathered by National Lightning Detection Network (NLDN) to develop linear and non-linear statistical models for the **number of lightning** (response) as a function of the contributable variables: **precipitable water**, **tropical storm wind total**, **sea level pressure anomaly**, **tropical storm winds anomaly**, and **Bermuda high average**. In addition to the **relative humidity** at various levels, **rain** in various counties, **sea surface temperature**, **precipitation anomaly district one**, **temperatures at various levels**, **temperature range**, **Pacific Decadal Oscillation (PDO) standard anomaly**, **Solar Flux standard anomaly**, **Pacific-North America Index (PNA) standard anomaly**, **precipitation anomaly district four**, **day of year**, **Arctic Oscillation (AO) standard anomaly** and **Outgoing Long Wave Radiation (OLR)**. To our knowledge, this is the first statistical model of its kind that includes the contributing interaction among the attributing variable to the response.

The developed statistical model enables us to identify and rank the contributing entities (explanatory variables/independent variables and interactions) that explain under

what conditions lightning strikes occur (Fieux, Paxton, Stano, and DiMarco, 2005) and estimate the response variable; namely, the number of lightning strikes in a given month. The quality of this model was determined using the following criteria: the coefficient of determination, R^2, in conjunction with R^2_{adj}, test for significance (p-value), Mallow's C(p) statistic and the F-statistics. All criteria uniformly support the high quality of the developed statistical model. Finally, using the developed model we performed surface response analysis; that is, we determined the values of the contributing variables that either maximize or minimize the response variable with an acceptable (pre-specified) level of confidence.

Lightning in the State of Florida is a significant phenomenon that we must make every effort to monitor and understand. Although this study concentrates on the State of Florida, similar methodology, and procedures are applicable to other states and further generalized to other regions where lightning is a factor. Moreover, with the networks established to continue collecting lightning data so that we can continually update the proposed statistical models and thus increase the occurrence and their effective use.

In our present study we will address the following questions, among others:

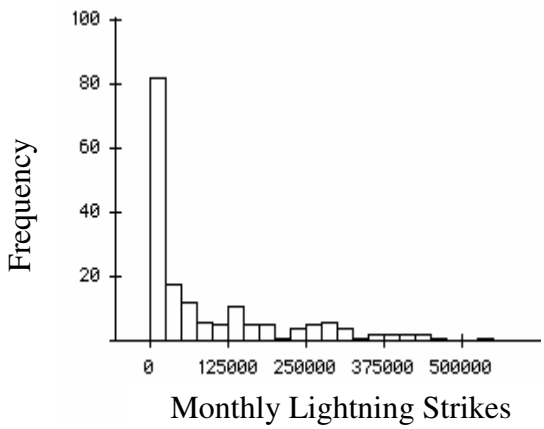
1. *Identify the region to be addressed and why?*
2. *What constitutes Lightning?*
3. *What is the best-fit probability distribution?*
4. *What is the expected number of lightning strikes?*
5. *What is the primary contributor to the number of lightning strikes?*
6. *What are the significant interactions that contribute to lightning?*

2. DESCRIPTION OF VARIOUS CONTRIBUTING ENTITIES

The data used in the present study were obtained from several sources listed below. The main data set consist of monthly total lightning strikes for a period of 16 years. Monthly records previously compiled with **relative humidity**, **temperature** including **Bermuda highs**, **tornadoes**, **waterspouts**, **hail**, among others. The actual sources where our information was obtained are:

1. NOAA, National Environmental Satellite, Data, and Information Service (NESDIS).
2. Monthly cloud-to-ground lightning data over Florida from 1989 to 2004 for May through September, NLDN.
3. Total monthly rainfall collected by sixteen counties in the state of Florida 1989 to 2004 from Southwest Florida Water Management District Hydrologic Data.

A descriptive display of the number of lightning strikes is shown by the histogram, along with the basic statistics, in Figure 1. The data is extremely skewed to the right with a monthly mean number of lightning strikes per month, $\bar{x} = 93{,}691.33$. Moreover, the sample standard deviation, $s = 120{,}871.21$ which dominates the sample mean with a coefficient of variation, $CV = \frac{s}{\bar{x}} \approx 1.29$; that is, 129% of the sample mean, and standard error $\varepsilon = \frac{s}{\sqrt{n}} \approx 9137.00$.



Statistic	Estimated Monthly
Count	175
Mean	93,691.33
Median	34,256
Standard Deviation	120,871.21
Standard Error	9,137.00
Minimum	102
Maximum	529,981
Range	529,879
Skewness	1.48
Kurtosis	1.35

Figure 1: Histogram for the number of lightning strikes per month including descriptive statistics for the number of lightning strikes per month

With such large coefficients of variation and standard errors, the interpretation of the sample mean could be misleading. We will address this issue in the next section, which we will introduce the bootstrapping procedure to reduce the standard error and obtain a more realistic estimate of the sample mean of the number of lightning strikes. In

this section, we are concerned with the best-fit probability distribution that characterizes the behavior of the subject data.

3. PARAMETRIC ANALYSIS

As illustrated in the above histograms, the data is extremely skewed. Thus we are certain that the data does not follow the normal (Gaussian) probability distribution; the skewness is measured to be 1.485 for the monthly distribution indicating that the data is significantly asymmetric.

3.1 BEST-FIT PROBABILITY DISTRIBUTION

It is clear that we have a peak in our distribution, which is more "curved" than the symmetric normal distribution. Furthermore, kurtosis measuring zero is normal and negative kurtosis indicates a flat distribution, the present data has a positive kurtosis; namely, 1.35 for the monthly distribution and even greater 3.04 for the daily distribution. Using standard statistical goodness-of-fit methods, first for the monthly number of lightning strikes, the only probability distribution that failed to be rejected at the 0.01 level of significance, as shown in Table 1, is the two- and three-parameter Weibull probability distribution, using four useful and commonly used tests to determine the goodness-of fit; namely, Kolmogorov – Smirnov, Cramer – von Miser, Anderson Darling and Chi-squared test statistics.

Table 1: Best-fit probability distribution for the number of lightning strikes per month

Test	**Weibull**
Kolmogorov – Smirnov	<0.001
Cramer - von Mises	<0.001
Anderson - Darling	<0.001
Chi-Squared	0.015

Furthermore, when considering the empirical cumulative probability distribution compared to each of the individual best-fit distributions using the coefficient of determination, R^2 and the adjusted statistic R^2_{adj}. Where the estimate of R^2 is given by

equation 1, increases as n increases whereas the estimate of R^2_{adj}, given by equation 2, does not increase when additional design parameters are added to the regression model and p is a constant; this statistic penalizes the inclusion of insignificant model terms and therefore is a better indicator of how well the model explains the behavior of the response.

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{1}$$

and

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right) \tag{2}$$

More specifically, R^2 and R^2_{adj} give a measure of the percent of the variance (deviation) explained by the developed model.

As shown in Table 2, the three-parameter Weibull has the highest R^2 and R^2_{adj} indicating the three-parameter Weibull is the best-fit probability distribution among the individual distributions tested.

Table 2: Estimated parameters for the number of lightning strikes per month

Distribution	**Parameters**	R^2	R^2_{adj}
Weibull (2)	$\hat{\alpha} = 0.6229, \hat{\lambda} = 67086, \hat{\theta} = 0$	99.0%	99.0%
Weibull (3)	$\hat{\lambda} = 102, \hat{\beta} = 187.847, \hat{\alpha} = 0.464$	99.5%	99.5%

However, since these models are for the most part the same and the argument can be made that the number of lightning strikes as shown in the sample data could be as few as one we will employ the law of parsimony and continue our study using the two-parameter Weibull.

Thus, the probability density function that characterize the probabilistic behavior of the number of lightning in a given month is given by the two-parameter Weibull with maximum likelihood estimates (MLEs) of the parameter as follows: the shape parameter $\hat{\alpha} = 0.6229$ and scale parameter $\hat{\lambda} = 67{,}086$ and the threshold set to zero

$(\hat{\theta} = 0)$; and therefore, the associated probability density function and cumulative probability distribution function given by equation 3.3 and 3.4, respectively.

$$f(x) = \begin{cases} \dfrac{0.6229}{67086}\left(\dfrac{x}{67086}\right)^{0.6229-1} \exp\left[-\left(\dfrac{x}{67086}\right)^{0.6229}\right], x > 0 \\ 0 \qquad , \text{otherwise} \end{cases} \tag{3}$$

and

$$F(x) = P(X \le x) = \begin{cases} 1 - \exp\left[-\left(\dfrac{x}{67086}\right)^{0.6229}\right], x > 0 \\ 0 \qquad , \text{otherwise.} \end{cases} \tag{4}$$

Table 3 gives the percentiles for both the observed data and the estimated values. According to the observed data, given that lightning has occurred, there is a 1% chance that there were less than 213 strikes. There is a 50% chance that given lightning occurred, there could be up to 34,256 strikes. There is also a 1% chance that given lightning occurs that there could be more than 451,786 strikes. According to the estimation, there is a 1% chance that given lightning occurs that there could be more than 778,687 strikes. Additional estimates can be made using the cumulative probability distribution is shown in Figure 2. In example, to estimate the 80% percentile using the Weibull probability distribution graph, we project backwards to find that approximately 150,000 strikes; that is, there is a 20% chance that given that lightning occurs that there will be at least 150,000 strikes. As Table 3 and Figure 2 both show a comparison of the estimated number of lightning strikes using the two-parameter Weibull and that measured in the data.

3.2 BOOTSTRAPPING

The technique of bootstrapping is a re-sampling procedure used when the estimates of the unknown, such as the sample mean, are such that the statistical error is significantly larger than the estimate of the true value. In the present study we are experiencing such a behavior and we believe that is due to the sample size not being large enough. Thus, we wish to study the stability of the sample mean of the number of

lightning strikes by increasing the sample size and thus reducing the standard error to an acceptable level.

Bootstrapping is a procedure that involves choosing random samples, *with replacement,* from the given data set and analyzing each sample in the same manner; that is, estimating the sample mean $\bar{x}$ and the standard error, $\frac{s}{\sqrt{n}}$. This procedure of resampling as a means of acquiring more information about the uncertainty of statistical estimators; allowing us to test the reliability of the estimates and assess whether stochastic effects have an influence on the probabilistic distribution which characterizes the phenomenon under study by reducing the standard error.

The original data set is considered and the following statistics computed: the sample mean, standard deviation and the standard error. Recall that the sample mean number of lightning strikes per month is $\bar{x} = 93{,}691.33$ with a standard deviation of $s = 120{,}871.21$, and standard error of $\varepsilon = \frac{s}{\sqrt{n}} \approx 9137.00$. There the variance significantly dominates the sample mean. Bootstrapping generated a data set of size five hundred; and the above statistics were calculated again. Then, independent of the above set, bootstrapping generated a data set of one thousand, then again we generated an independent data set of fifteen hundred and finally we increased the data set to ten thousand. See the appropriate statistics in Table 4, given below, along with 95% confidence limits of the true mean.

Table 4: Sample size, mean, standard deviation, standard error and the 95% confidence interval for the true mean number of lightning strikes per month

Data	**Number** n	**Mean** $\bar{x}$	**Std. Dev.** s	**Standard Error** $\frac{s}{\sqrt{n}}$	**Lower Bound**	**Upper Bound**
Original	175	93691.33	120871.2	9137.00	757820.80239	111599.8576
BS 1	500	91832.02	116924.7	5229.03	81583.11816	102080.9218
BS 2	1000	93197.93	120531.9	3811.55	85727.28544	100668.5746
BS 3	1500	89095.06	113873.5	2940.20	83332.26857	94857.85421
BS 4	10000	91138.73	118640.2	1186.40	88813.38208	93464.07792

This procedure is consistent with the sequence of sample means,

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}; n = 1,2,...,10000,$$

the sequence of standard deviations,

$$s_n = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n-1}}; n = 1,2,...,10000,$$

and sequence of sample standard error: $\varepsilon_n = \frac{s_n}{\sqrt{n}}; n = 1,2,...,10000$ generated using the fifth data set described above. The convergence of the mean, the convergence of the standard deviation and standard error are illustrated in Figures 3 thru 5.

Note that, as $s_n \to s$ as $n \to \infty$ therefore for $\varepsilon_n < 1$, this implies $n > s^2$. In general, to reduce the standard error by a factor of nine, we would need eight-one times as much data.

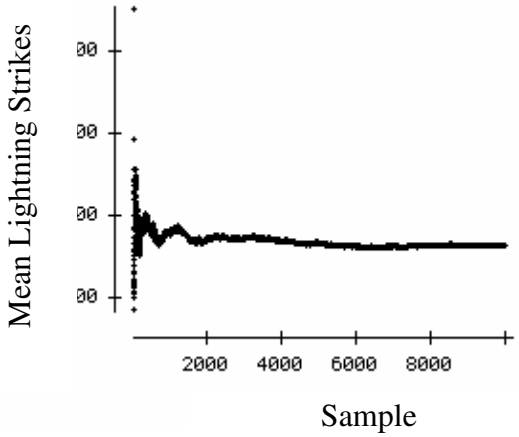


Figure 3: Convergence of the sample mean $\bar{x}$

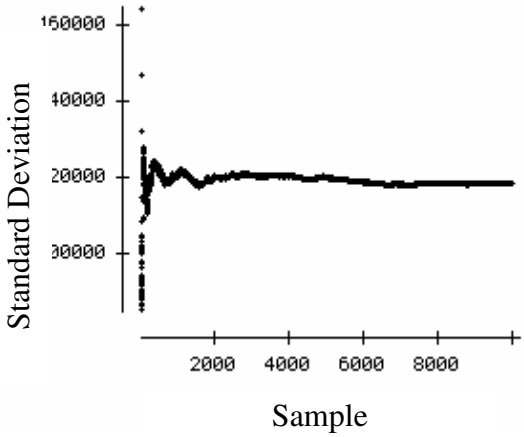


Figure 4: Convergence of the sample standard deviation s_n

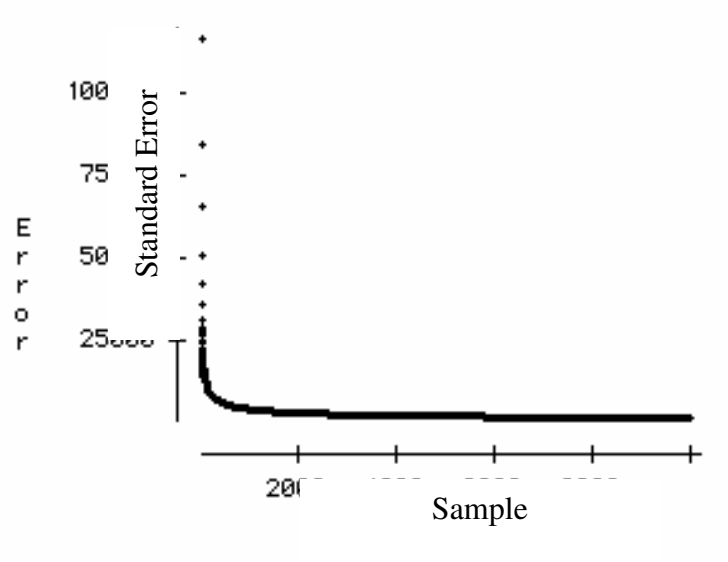


Figure 5: Convergence of the sample standard error ε_n

Furthermore, if we consider the percent change in the sample mean, $p_i = \frac{\bar{x}_i - \bar{x}_{i-1}}{\bar{x}_{i-1}}; i = 3,4,...,10000$ and the percent change in the standard error, $q_i = \frac{\varepsilon_i - \varepsilon_{i-1}}{\varepsilon_{i-1}}; i = 3,4,...,10000$, we have a very good convergence as shown in Figures 6 and Figure 7.

Note that the rate at which both the sample mean and the sample standard error converges is only significant when $n < 500$. Thus, the sample mean is a stable estimate of the unknown true number of lightning strikes; we proceed to obtain confidence limits

of the true mean of the number of lightning strikes. Table 5 gives 90%, 95%, and 99% confidence limits of the true number of lightning based on our sampled data.

Hence, based on the actual sample of $n = 175$, the bootstrapping methodology was very helpful in establishing the significant importance of the key estimate of the unknown number of lightning strikes.

3.3 PERCENTAGE CHANGE IN THE MEAN AND STANDARD ERROR:

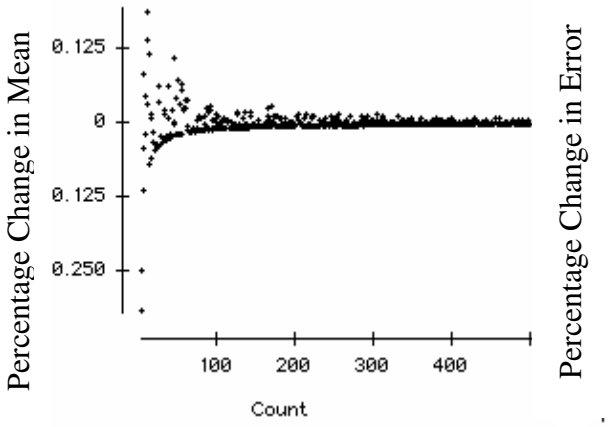


Figure 6: Convergence of the standard error p_n

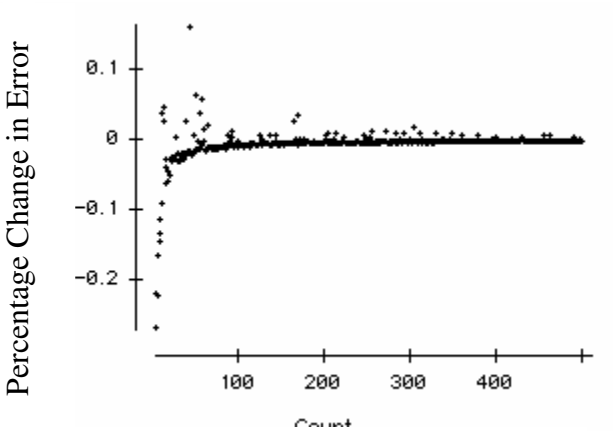


Figure 7: Convergence of the standard error q_n

Table 5: Confidence intervals for the true mean using the two-parameter Weibull

Confidence Level	Lower	Upper
99%	69,895	117,488
95%	75,658	111,725
90%	78,582	108,801

That is, we are at least 99% confident that the true mean number of lightning strikes per month is between 69,895 and 117,488, similarly we are at least 95% confident that the true mean number of lightning strikes per month is between 75,658 and 111,725, etc.

4. MULTIVARIATE STATISTICAL MODEL OF THE NUMBER OF LIGHTNING STRIKES

In this section, we will use historical data collected by the National Lightning Detection Network (NLDN) in conjunction with other meteorological phenomena to identify the contributing entities or explanatory variables (independent variables) that cause lightning to occur.

Several additional variables were tested but were not found to significantly contribute to the number of lightning strikes – such variables as **tornados** by category, **waterspouts**, **hail** by size, **North Atlantic Oscillation** (NAO), **Madden/Julian Oscillation**, etc. To identify and rank the contributing variables to the number of lightning strikes, a statistical model was developed using forward regression analysis. We retained the variables that were found to be significant at the 0.01 level that resulted in increasing the quality of the statistical model by maximizing the increase to R^2 and R^2_{adj}, the key criteria used in determining the quality of the model. Of significance importance in the development of the subject model was the contribution to the response variable due to the interaction of the contributing variables.

4.1 STATISTICAL MODEL

The initial statistical model we developed for the response variable (**number of lightning strikes**) as a function of the **month** in the year, **perceptible water**, **precipitation** by district, **sea level pressure**, **Bermuda highs**, **relative humidity** at various levels in the atmosphere, **temperatures** at various levels in the atmosphere, **sea surface (water) temperature**, the **Pacific Decadal Oscillation** (PDO), the **Pacific/North America Oscillation** (PNA), the **Artic Oscillation** (AO), the **Outgoing Long-wave Radiation** (OLR), **Solar Flux** as an average or an anomaly; and **rainfall** in sixteen counties enumerated in the state as follows:

Enumeration of Counties: 1 Levy, 2 Marion, 3 Citrus, 4 Sumter, 5 Hernando, 6 Lake, 7 Pasco, 8 Polk, 9 Pinellas, 10 Hillsborough, 11 Manatee, 12 Hardee, 13 Highlands, 14 Sarasota, 15 Desoto and 16 Charlotte

Table 6 gives the mathematical notations of the attributing variables that are used in the present study along with a brief description.

Table 6: Variables of interest to investigate in estimating the mean number of lightning strikes per month in the State of Florida

Variable	**Description (possible contributing variable)**
m	Month of year
pw	Perceptible water
p_d	Precipitation anomaly by district; $d = 1,2,3,4,5,6,7$
P_w	Sea level pressure: average
$\dot{P}_w$	Sea level pressure: anomaly
w	Tropical storm winds: total
$\dot{w}$	Tropical storm winds: anomaly
T_B	Bermuda high: average
$\dot{T}_B$	Bermuda high: anomaly
rh_{mb}	Relative humidity: Levels: 1000mb, 850mb, 700mb, 600mb, 500mb, 400mb 300mb, 200mb, 100mb
T_{mb}	Temperature: Levels: 1000mb, 850mb, 700mb, 600mb, 500mb, 400mb 300mb, 200mb, 100mb
T_{range}	The maximum range between temperatures at various levels
T_w	Sea surface temperature
PDO	Pacific Decadal Oscillation index standard anomaly
PNA	Pacific/North America Oscillation index standard anomaly
AO	Artic Oscillation
orl	Outgoing Long-wave Radiation
sf	Solar flux std anomaly
r_i	Rainfall: total monthly rainfall collected by sixteen counties in the state of Florida

Note that not all of the variables listed in Table 5, were found significant; therefore, only significantly contributing variables will be considered. Moreover, we consider interaction between these terms to better understand the underlying relationships between the contributing variables, but first we identify and rank the contributing statistically significant variables listed in Table 7. The criteria used to determine which variables remained in the model, using Mallow's $C(p)$ statistic.

Mallow's $C(p)$ statistic measure the significance of including additional parameters and is computed using Equation 5 where s^2 (sample variance) is the mean square error for the full model, SSE_p is the sum of square errors for the model with p parameters. If we have identified the right model, then the statistic estimates of the number of parameters (p) required in the model, will result in $C(p)$ converging to p, with

$$C(p) = \frac{SSE_p}{s^2} - (N - 2p). \tag{5}$$

Hence, only the first seventeen variables should be included; that is, this number of parameter yields the least difference between $C(p) = 16.9$ and $p = 17$. Given below, Table 7 is the contributing variables ranked by according to their percent contribution into the number of lightning strikes. The key statistical factors in developing this important table are based on the R^2 values.

Table 7: Statistical ranking of the attributing variables to the number of lightning strikes

Rank	**Variable**	R^2
1	Precipitable water	61.00
2	Tropical storm wind total	71.44
3	Sea level pressure Anomaly	78.94
4	Tropical storm winds Anomaly	89.99
5	Bermuda high average	92.34
6	Relative humidity (1000mb)	94.76
7	Rain in Hernando county	95.53
8	Sea surface temperature	96.10
9	Temperature range	96.61
10	Precipitation Anomaly district one	96.90
11	Relative humidity (850mb)	97.10
12	Relative humidity (500mb)	97.40
13	Temperature (850mb)	97.80
14	PDO standard Anomaly	98.00
15	Rain in Hillsborough county	98.10
16	Rain in Highlands county	98.20
17	Solar Flux standard Anomaly	98.30

4.3 INTERACTION

While the contributing variables given in Table 7 explain 98.3% of the variation in the response, to improve the quality of the model we tested for possible contribution to the number of lightning strikes by various interacting contributing variables. After an extensive study of all possible interaction we have found the following interacting variables to statistically contributing to the response variable. There is significant interaction between:

- Precipitable water and relative humidity (500mb)
- Sea surface (water) temperature and relative humidity (500mb)
- Month and the pressure (sea level)

Thus the theoretical model that statistically characterize the behavior of the contributing variables along with significant interaction is given by

$$N = \begin{cases} \beta_0 + \beta_1 pw + \beta_2 w + \beta_3 \dot{w}. + \beta_4 P_w + \beta_5 \dot{P}_w + \beta_6 T_{range} + \beta_7 T_{850} \\ + \beta_8 T_w + \beta_9 rh_{1000} + \beta_{10} rh_{850} + \beta_{11} rh_{500} + \beta_{12} PNA + \beta_{13} sf \\ + \beta_{14} T_B + \beta_{15} p_4 + \beta_{16} r_5 + \beta_{17} r_6 \\ + \beta_{18} pw \times rh_{500} + \beta_{19} T_w \times rh_{500} + \beta_{20} m \times P_w + \varepsilon \end{cases}, \qquad (6)$$

where the coefficients $\beta_i's$ are the weights that drive the estimate of the contributing variables and ε is the random error.

Using the information available (real world data) we have structured the following statistical model to estimate the theoretical model given by equation (6) that will estimate the number of lightning strikes per month along with contributing variables and interaction, that is,

$$\hat{N} = \begin{cases} -190\times10^7 + 21913.9\, pw + 1491.85 w - 1315.12 \dot{w}. \\ - 5534.6 P_w - 0.934669 \dot{P}_w - 5815.26 T_{range} - 1.20\times10^4 T_{850} \\ - 4.91\times10^3 T_w - 4481.25 rh_{1000} - 2743.72 rh_{850} \\ - 10134.5 rh_{500} + 2.82\times10^3 PNA - 3194.69 sf \\ + 24949.8 T_B - 4075.7\, p_4 + 3628.99 r_5 + 2310.25 r_{13} \\ - 229.526\, pw \times rh_{500} + 634.607 T_w \times rh_{500} + 559.587 m \times P_w \end{cases} \qquad (7)$$

where $\hat{N}$ is the estimate of the number of lightning strikes.

This statistical model results in $R^2 = 98.4\%$, which is an improvement of the previous statistical model. All contributing entities except sea surface temperature are significant with all approximate results shown in Table 8.

Table 8: Least-squares regression for the number of lightning strikes in a month with respect to the ranked independent variables including interaction; also included are the associated *p*-values

Variable	Coefficient	SE of Coefficient	t-ratio	*p*-value
Constant	-1.90E+07	2400000.00	-7.93	< 0.0001
pw	21913.9	2226.00	9.84	< 0.0001
w	1491.85	156.10	9.56	< 0.0001
$\dot{P}_w$	-0.934669	0.02599	-36	< 0.0001
$\dot{w}$	-1315.12	165.80	-7.93	< 0.0001
T_{range}	-5815.26	909.30	-6.4	< 0.0001
rh_{1000}	-4481.25	825.90	-5.43	< 0.0001
rh_{850}	-2743.72	553.90	-4.95	< 0.0001
rh_{500}	-10134.5	2853.00	-3.55	0.0005
T_B	24949.8	1947.00	12.8	< 0.0001
r_5	3628.99	779.40	4.66	< 0.0001
p_4	-4075.7	1205.00	-3.38	0.0009
T_w	-4.91E+03	5893.00	-0.833	0.406
sf	-3194.69	1281.00	-2.49	0.0137
PNA	2.82E+03	1183.00	2.38	0.0185
r_{13}	2310.25	941.80	2.45	0.0153
T_{850}	-1.20E+04	3026.00	-3.96	0.0001
m	-569230	272400.00	-2.09	0.0383
$pw \times rh_{500}$	-229.526	56.17	-4.09	< 0.0001
$T_w \times rh_{500}$	634.607	183.50	3.46	0.0007
P_w	-5534.6	1751.00	-3.16	0.0019
$m \times P_w$	559.587	267.20	2.09	0.0379

The result in the above table statistically justifies the overall significant development of the proposed model based on actual data.

5. STATISTICAL MODEL VALIDATION

The following statistical criteria were used to identify and attest to the quality of the developed statistical models: the *p*-values determining significance of each contributing term in conjunction with the R^2 and R^2_{adj} statistics, the F statistics, and Mallow's $C(p)$ statistics. The uniform response of the statistical tests attests to the high quality of the proposed model.

Here, we shall give a brief discussion of the statistical test used to validate the quality of the proposed model.

5.1 THE COEFFICIENT OF DETERMINATION R^2 AND THE ADJUSTED COEFFICIENT OF DETERMINATION R^2_{adj}

This is the last statistic used in the F-test. We have already discussed the significance of the R^2 and R^2_{adj} statistics previously. Using the first of these statistics, we can rank the contributing variables with respect to the maximum increase in R^2. Where in the non-interactive model, precipitable water explained 61% of the variation in the subject response and in the interactive model, the interaction between the precipitable water and the relative humidity (500mb) explain 60.59% of the variation in the number of lightning strikes. Second, explaining an additional 13.61% of the variation is the water pressure (anomaly). All other contributing variables explain less than 10% of the explanatory power; however with all 21 contributing variables, given in Table 9, the developed model explains 98.41% of the variation in the subject response.

5.2 THE F - STATISTIC

Consider the linear model, = + , where is an $n \times m$ matrix, $\{x_{ij} \mid i = 1,2,..,n \text{ and } j = 1,2,...,m\}$ are known constants and full rank; that is, $m < n$. The vector is a vector of unknown parameters $\beta_0, \beta_1, ..., \beta_m$ and $= [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]$ is a vector of non-observable independent normal random variables (RVs) with common variance σ^2 and mean $E(\ ) = 0$. The general linear regression test for testing the null hypothesis H_0 : $= 0$, where is an $m \times n$ matrix of full rank $m \le n$, is to reject the null hypothesis at the confidence level α if $F \ge F_\alpha$. The significance level is

$\alpha = P\{F \geq F_\alpha \mid H_0\}$ and F is given by equation (8) and $\frac{n-m}{m}F$ follows the F probability distribution with m and $n-m$ degrees of freedom.

Using the F-test, at the 0.10 significance level, the 21 contributing variables listed in Table 9 are found to be significantly contributing. At the 0.05 level the variables solar flux, sea surface pressure and the interaction between the month and the sea surface pressure are not significantly contributing. Therefore, we will use the following criteria to determine if these variables should be removed.

Table 9: Summary for Forward Selection including Mallow's $C(p)$ statistics

Rank	**Variable**	**Partial R^2**	R^2	$C(p)$	F	**Pr>F**
1	$pw \times rh_{500}$	60.59	60.59	3627.57	265.93	< 0.0001
2	$\dot{P}_w$	13.61	74.2	2318.1	90.7	< 0.0001
3	w	6.79	80.99	1665.27	61.11	< 0.0001
4	$\dot{w}$	8.77	89.76	822.406	145.46	< 0.0001
5	rh_{500}	2.55	92.31	578.423	56.07	< 0.0001
6	T_B	1.58	93.89	428.4	43.33	< 0.0001
7	$T_w \times rh_{500}$	1.68	95.57	268.963	63	< 0.0001
8	T_{range}	0.6	96.17	212.804	26.11	< 0.0001
9	r_5	0.46	96.63	170.474	22.47	< 0.0001
10	PDO	0.26	96.89	146.971	13.94	0.0003
11	rh_{1000}	0.2	97.09	129.976	11.02	0.0011
12	pw	0.51	97.6	82.8983	34.29	< 0.0001
13	rh_{850}	0.23	97.83	62.4231	17.28	< 0.0001
14	T_{850}	0.24	98.07	41.2136	19.94	< 0.0001
15	PNA	0.07	98.14	36.7854	5.69	0.0183
16	p_4	0.03	98.17	35.5194	2.92	0.0893
17	r_{13}	0.07	98.24	30.2948	6.7	0.0105
18	sf	0.04	98.28	28.2024	3.86	0.0511
19	m	0.05	98.33	25.8096	4.23	0.0413
20	P_w	0.04	98.37	23.8977	3.84	0.0519
21	$m \times P_w$	0.04	98.41	22	3.9	0.0502

5.3 MALLOW'S $C(p)$ STATISTIC

Using forward selection with a 0.10 level of significance to enter the model and a 0.10 level of significance to remain in the model, then we have as shown in Table 9, $C(p) \to 22$ where $p = 21$; and this is a strong indication of the high quality of the developed model.

Moreover, when considering the one variable t - test with the null hypothesis; that is, that the mean residual is zero. Therefore, the F statistic indicates that all contributing variables are significant at the 0.10 level and $R^2 = 98.41\%$, that is 98.41% of the variation in the subject response (number of lightning strikes) is explained by the least-squares regression model. Hence, all 21 variables will remain in the proposed model. Therefore, all criteria uniformly support the high quality of the statistical model.

6. USEFULNESS OF THE STATISTICAL MODEL

Lightning affects us in several ways; one lightning casualty occurred for every 86,000 strikes (over the United States) and one death occurred for every 345,000 flashes (NOAA). Second, lightning causes power outages. It would be useful for the energy supply company to have a statistical model which would estimate lightning storms in order for them to better serve customers and which minimize expense, not just to be able to estimate the number of lightning based on the surrounding environmental data but to appropriately allocate resources – when should additional workers be scheduled to make repairs to the system by estimating potential occurrence and in general, to be able to develop strategies for the safety of our citizens, among others. The developed model can be used effectively to address these issues.

7. CONCLUSION

When the electrostatic energy within storm conditions is unbalanced and ephemeral discharges of static electricity, this discharge is seen as light or lightning. The phenomenon is a common occurrence in the State of Florida. Basic descriptive statistics indicate that the mean number of lightning strikes is approximately 93,961 strikes in a given month. To verify these estimations' accuracy, parametric analysis is used to show

that the number of lightning strikes per month is not Gaussian distributed due to the extreme skewness in the data.

The Weibull probability distribution best characterize the behavior of the subject response (number of lightning strikes). Both the two-parameter and three-parameter Weibull probability distribution gives very good fitness results for the subject data. Employing the law of parsimony, as well as the fact the true minimum number of lightning strikes is zero, the two-parameter Weibull was used when estimating the return period for various numbers of lightning strikes. Further analysis utilizing the technique of bootstrapping to generate 500 count shows that the true number of lightning strikes per month at the 95% confidence level is between 81,583 and 102,080 strikes in a given month.

Second, non-linear modeling of the number of lightning strikes per month with respect to the amount of **perceptible water**, **wind shear**, **anomalies** in **sea level pressure**, various **temperatures**, **relative humidity** at different levels in the atmosphere and several other significantly contributing variables, which explains approximately 98.4% of the variation in the subject response. The contributable variables to the response were identified along with the significant interaction and ranked in accordance to the percent of contribution. Using the developed model we can estimate the **number of lightning strikes** per month given the environment along with their interaction with $R^2 = 98.4\%$ and $R^2_{adj} = 98.2\%$. Significant interactions include the following interactions: perceptible water and relative humidity at the 500 level, relative humidity at the 500 level and sea surface (water) temperature, the sea surface pressure and month of year.

Finally, the quality of the proposed statistical model was uniformly verified by using the following tests: R^2 in conjunction with R^2_{adj} and p-value, the F statistic and Mallow's $C(p)$ statistic. Although this model was developed using real data from the State of Florida, it can similarly be developed using data from other regions where lightning strikes are of significant importance to the public.

8. REFERENCE

1. Fieux, J. F., Paxton, C. H., Stano, G. T., and DiMarco, J. P., 2005. Monthly Lightning Trends over Florida 1989-2004, NOAA/NWS

2. Hodanish, S., Sharp, D., Collins, W., Paxton, C. H., and Orville, R. E., 1996. A 10-yr Monthly Lightning Climatology of Florida: 1986-95, American Meteorological Society

3. Lucretius, 58 B.C., "The Nature of Things", Book VI. W. & W. Norton 152-159

4. Sheridan, S. C., Griffiths, J. F., and Orville, R. E., 1997, "Warm Season Cloud-to-Ground Lightning-Precipitation Relationship in the South-Central United States", American Meteorological Society