P2.41 AN ARTIFICIAL NEURAL NETWORK TO FORECAST THUNDERSTORM LOCATION: A SEARCH FOR MORE RELEVANT LAND SURFACE INPUT DATA

Waylon Collins* and Philippe Tissot**
*NOAA/National Weather Service
**Texas A&M University – Corpus Christi

1. INTRODUCTION

A feed-forward, supervised, multi-layer perceptron Artificial Neural Network (ANN) was developed to test the hypothesis that an ANN can be developed to successfully forecast convective initiation (CI) with an accuracy of 400-km^2 (Collins and Tissot, 2007). The ANN domain is an area in South Texas composed of 286 20-km x 20-km box regions, 13 boxes north-south by 22 boxes east-west. A framework was established to train 286 separate ANNs (one for each box region) to predict thunderstorm occurrence within each box. The ANN inputs included both selected output from a deterministic grid point numerical weather prediction (NWP) model – with a horizontal grid spacing of 12km – and sub-grid scale data that contribute to CI. The logic underlying this strategy is that NWP model output provides a forecast of whether the larger scale mesoscale environment is conducive to CI while the sub-grid scale data determines the extent to which convection could be triggered at a particular location. By incorporating both grid scale NWP output and sub-grid scale data (thus not explicitly accounted for by the NWP) that contributes to CI, an improvement beyond NWP output alone is envisaged. This approach represents a paradigm shift away from the idea of increasing NWP model horizontal resolution to more accurately and explicitly forecast CI. The sub-grid scale inputs include land surface temperature (LST) gradients. Numerous studies (e.g. Avissar and Liu, 1996) demonstrate that surface heterogeneity (including vegetation and soil moisture variations) contribute to horizontal LST gradients which can trigger individual convective cells. This study represents an attempt to provide more relevant land surface data inputs to the ANN model, by identifying a parameter that clearly separates the heterogeneity pattern contributing to CI from the pattern not conducive to CI. In their prior work Collins and Tissot (2007) used statistical parameters *range* (RG), *standard deviation* (SD), and the *maximum finite difference* (MFD) values in orthogonal directions as proxies for LST gradients. However high values of these parameters can be associated to a broad range of atmospheric and land based processes. This work focuses on investigating the discriminating potential of selected land surface parameters in cases for which atmospheric conditions are favorable to the development of thunderstorms. Developing such parameters capturing specific land surface patterns and gradients should subsequently lead to more accurate thunderstorm ANN predictive models.

2. DATA AND METHODOLOGY

Several proxies for heterogeneity were calculated using three fundamentally different techniques. The first technique was an application of image processing. The second technique involved actual gradient calculations. The third technique was statistical in nature. The heterogeneity proxies were calculated within each of the 286 20-km x 20-km box regions.

The source data includes both the (1-km grid spaced) daily and 8-day averaged LST data from the NASA Moderate Resolution Imaging Spectro-radiometer (MODIS) instrument aboard the Terra and Aqua satellites. For additional information, view http://modis.gsfc.nasa.gov. Specifically, the output from the *LST_Day_1km* variable or *scientific data set* (SDS) from the daily and 8-day mean (the arithmetic average of valid daily values at each pixel) files with granule names MOD11A1 (from the Terra satellite) and MYD11A2 (from the Aqua satellite) respectively were used. The choice of 8-day averaged in addition to the daily files is based on the limitation of the MODIS data – the existence of clouds which prevents data collection. There is a greater likelihood that every pixel within each box region will contain valid 8-day averaged, rather than daily, data. However, it must be noted that daily values are more appropriate since the LST data will be compared to daily lightning data. The original MODIS data was available on an Integerized Sinusoidal (ISIN) projection and written to HDF-EOS files. The software package known as the *Modis Reprojection Tool* (http://edcdaac.usgs.gov/landdaac/tools/modis/info/MRT_Users_Manual.pdf) was used to reproject the data to a geographic projection grid and write the output to separate HDF-EOS files. The MATLAB® software was used to import the new HDF-EOS files and output a matrix containing the LST data for each 20-km x 20-km box.

A literature review reveals that LST has not been used as a parameter to test the hypothesis that land surface heterogeneity contributes to CI. Instead, parameters such as vegetation and soil moisture were used. However, Collins and Tissot (2007) used LST because (1) the LST adjusts to soil moisture and vegetation changes and (2) NASA provides daily MODIS LST, yet does not provide daily MODIS soil moisture or vegetation. However, NASA does provide 1-km MODIS 16-day averaged *normalized difference vegetation index* (NDVI) data. NASA also provides daily soil moisture data from its Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) instrument. However, the resolution of the AMSR-E data is too coarse for this study. In this study, we used MODIS NDVI data, aqua granule MYD13A2. With respect to soil moisture, we utilized an empirical relationship to compute a soil moisture index (SMI) for a semiarid environment (Zeng et. al. 2004). The equation requires only NDVI and LST as inputs. In this study, the 16-day NDVI and 8-day LST data were used to compute the SMI.

*Corresponding author address: Waylon G. Collins, National Weather Service, 300 Pinson Drive, Corpus Christi, TX 78406; e-mail: Waylon.Collins@noaa.gov

Figure 1 depicts the time frame of the various data sources used in this study. The MODIS daily and 8-day averaged LST and 16-day averaged NDVI are derived from polar orbiting satellites. The approximate time frame for the MODIS Terra (Aqua) satellite to move across the ANN domain to collect data from the LST SDS *LST_Day_1km* is 1600-1745 (1845-2025) UTC. Output from the Eta NWP model (e.g. Rogers et. al. 1996), discussed later, are 9-hour forecasts valid at 2100 UTC. The proxy for CI is the presence of cloud-to-ground lightning data from the National Lightning Detection Network (NLDN) (e.g. Orville 1991). For each case all lightning strikes taking place between 19Z and 23Z inclusive are included. The overall period of the data collection is 749 consecutive days (1 June 2004 through 19 June 2006).

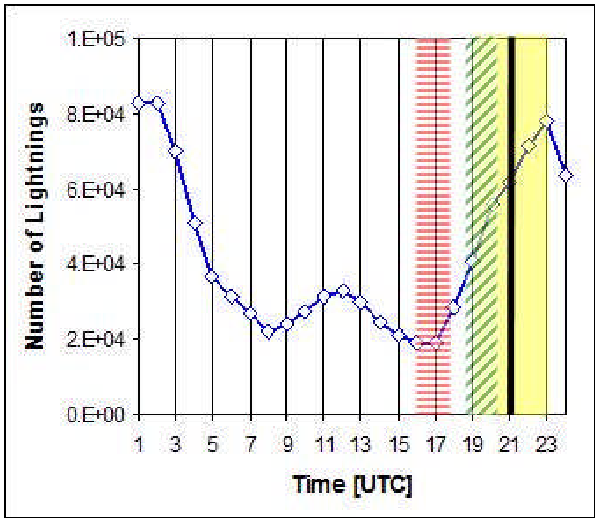

Figure 1: Source data time frames. Blue line: Lightning histogram for all 286 boxes for all 749 days. Yellow shading: Lightning used for this study. Red (Green) shading: Terra (Aqua) periods. Thick black line: Eta forecasts valid time. See text.

In order to extract the LST, SMI and NDVI patterns that truly correlate with lightning, a simple approach would involve a separation of the dataset into non-lightning and lightning cases. However, CI/lightning is hypothesized to develop in response to both favorable atmospheric conditions and favorable land surface patterns. This complicates the analysis since during non-lightning cases, it is likely that the entire range of possible LST, SMI and NDVI patterns would occur; in other words, for the non-lightning cases dataset, CI would not occur, either because the atmospheric conditions are not favorable, or because the land surface gradients were not sufficient. However, by restricting the dataset only to those cases whereby the atmosphere was conducive to CI, it is reasoned that variations in binary lightning output (lightning, no lightning) will be due primarily to horizontal gradients in LST, SMI and NDVI. The filtering of atmospheric conditions was based on the use of threshold values of specific parameters that correlate with CI generated by land surface heterogeneity. The thresholds, based on Eta output, are as follows: Convective Available Potential Energy (CAPE) >1000 Jkg^{-1}, Convective Inhibition (CIN) >-20 Jkg^{-1}, Precipitable Water (PW) >30 kgm^{-2}, both the 10-meter (m) and 850 millibar (mb) u and v components of the wind < 5 ms^{-1}, Lifted Index (LI) < -2, and vertical wind shear within the surface to 850mb layer < 0.005s^{-1}. These thresholds were used to separate "favorable" from "unfavorable" atmospheric conditions for CI. The choice of parameters and threshold values were based on the subjective evaluation of the lead author, the literature, and the time series data of these parameters and of lightning for the period 1 June 2004-19 June 2006 (not shown). See Collins and Tissot (2007) for more information. From the "favorable" dataset, histograms of the heterogeneity proxies for the lightning and non-lightning cases were developed to test whether certain land surface patterns are truly different between lightning and null cases. Note that the time resolution of the Eta output is daily while the LST (NDVI) output resolution is daily or 8-days (16-days). Thus, conditions can be favorable or unfavorable for CI during a given 8 (16) day period. Further, lightning can both occur and not occur during each 8 (16) day period.

The LST source data used has 1-km grid spacing. Thus, for each box region, a maximum of 400 data points were possible. Collins and Tissot (2007) used daily LST values, and computed the statistical parameters within each box region without regard to the number of data points. However, due to the predominance of clouds within the daily source data, output from box regions containing valid data points representing only a small percentage of the maximum, is less likely to be representative of the box region. Thus, gradients were calculated using the foregoing MODIS data with the restriction that missing data pixels for each box region comprise less than 10% of the maximum number of pixels. When applied to daily LST maps this restriction eliminated most lightning cases leading to the focus of the study on 8-day LST and 16-day NDVI data sets. For each box region, the following smoothing technique was implemented – Each of the data points was replaced with the mean of the eight (8) adjacent points and the point in question. The motivation behind the smoothing is to minimize gradients less than length scale 5-km. Results from Lynn et. al (2001) and Avissar and Schmidt (1998) suggest that when micro-α and meso-γ wind patterns develop in response to distinct regions of significantly different temperatures, the length scale of each region should exceed 5-km. No attempts were made to filter questionable data.

Heterogeneity Proxy #1: Image Processing Output

For this study, a proxy chosen for land surface heterogeneity within each box region was the Canny filter edge count output (EC) from the MATLAB® software package. The Canny algorithm (Canny 1986) is a widely used image processing technique to detect edges within an image. The Canny method is more likely than other edge detection algorithms to detect true weak edges, and less likely to be influenced by noise in the data set. The MATLAB® function BW=edge (I,'canny',thresh) was used to calculate the number of edge counts, where I=data matrix and 'canny' designates the Canny method. The parameter 'thresh'=[L,H], where L=low threshold and H=high threshold. For this study [L,H]=[0.40,0.99] and was held constant for the entire data set. See www.mathworks.com for additional information.

The processing of the data within each box region using the Canny method was based on the reasoning that an image processed edge would represent a boundary separating regions of significantly different LST, NDVI, and SMI values.

Heterogeneity Proxy #2: Traditional Gradient Calculation
Another proxy computed from the data was the *maximum gradient* (MG). The MG in each box region was determined by first calculating the gradient (finite difference divided by distance) between each data point and every other data point, resulting in a maximum of 400 gradient calculations. The maximum gradient was then chosen. It must be noted that although a horizontal thermal contrast is necessary in order to generate the mesoscale wind pattern conducive to CI, Segal et. al. (1998) has shown that the wind magnitudes may be invariant to the intensity of the gradient at the boundary. Nevertheless, we expect a minimum gradient threshold to emerge from the data set.

Heterogeneity Proxy #3: Bimodality Coefficient
The bimodality coefficient (BC) was calculated from an empirical relationship that relates BC to the statistical parameters kurtosis and skewness (SAS Institute Inc. 1999). A BC value exceeding 0.555 suggests a bimodal or multimodal distribution. Again, results from Lynn et. al (2001) and Avissar and Schmidt (1998) suggest that when micro-α and meso-γ wind patterns develop in response to distinct regions of significantly different temperatures, the length scale of each region should exceed 5-km. It is reasoned that a perfectly bimodal distribution of LST, SMI, or NDVI within the 20-km x 20-km box regions in this study would represent two distinct regions with a length scale of at least 10-km.

3. RESULTS

Figures 2-7 depict the histograms of the SMI and 8-day LST (NDVI not shown) EC, MG, and BC for lightning (L) and non-lightning (NL) cases, from a maximum of 214,214 box cases (286 boxes/day x 749 days). Owing to cloud cover near CI, gradient proxy output based on the daily LST data was miniscule and thus insufficient to draw meaningful conclusions. Thus, results from the 8-day LST and 16-day NDVI data were used. Unfortunately for each proxy, the difference between the histograms of L versus NL cases was not significant. These results suggest the following possibilities: **(1)** land surface heterogeneity was not a significant contributor to CI for the specific storms that developed over the 286 box domain. Our dataset was 749 consecutive days. However, it's possible that the contribution of land surface heterogeneity is seasonally dependent. We did not examine that possibility; **(2)** the MODIS 8-day (16-day) LST (NDVI) source data was not appropriate when assessing the relationship between MODIS data and daily lightning. We are interested only in binary lightning output with a time resolution of 1 day or less. Thus, a comparison between the 8-day and 16-day MODIS output and lightning with a corresponding time resolution was not conducted. We need to acquire/calculate gradient proxies on a time resolution of ≤ 1 day; **(3)** most thunderstorms within a box region originated from another box region. A thunderstorm within a box at a given time did not necessarily form in that box, or entirely from processes within such box. A storm could form within a box then simply move into another box. In addition to surface heterogeneity, gust fronts from neighboring storms, sea breezes, synoptic fronts, and other processes can also trigger convection.

4. CONCLUSIONS

There exists a broad consensus that surface heterogeneity can generate mesoscale wind patterns that can trigger CI. However, to the extent that land surface heterogeneity contributed to CI during the period of this study, the results herein suggest that the specific land surface heterogeneity patterns that contribute to CI were not identified. The inability to discover a relationship between convection and land surface heterogeneity could simply reflect the lack of MODIS data with a higher time resolution. Further, the contribution of land surface heterogeneity to CI could be seasonally dependent. Lastly, land surface heterogeneity may have represented a small fraction of the triggering mechanisms responsible for CI in this study. The search for a surface pattern that contributes to CI is ongoing. The literature is replete with studies which suggest that high resolution soil moisture gradients contribute to CI. Based on the reasoning that the use of higher time resolution data is warranted, we plan to calculate (on the order of 1-km) soil moisture at higher time resolutions. A technique to consider is the one proposed by Jiang and Cotton (2004) who used an ANN to estimate soil moisture. Further, data mining techniques such as clustering and singular value decomposition (SVD) (e.g. Bretherton et. al. 1992) will be examined. SVD will allow for the extraction of the LST, soil moisture and NDVI horizontal patterns that correlate with lightning variations. Both clustering and SVD focus on patterns rather than gradients. Until surface patterns or gradients that discriminate between lightning and non-lightning cases can be identified, the use of land surface data cannot be relied on to improve the performance of the ANN to predict CI.

5. REFERENCES

Avissar, R., and Y. Liu. 1996: Three-dimensional numerical study of shallow convective clouds and precipitation induced by land surface forcing. *J. Geophys.Res.* **101**, 7499-7518.

---- and T. Schmidt, 1998: An evaluation of the scale at which ground-surface heat flux patchiness affects the convective boundary layer using large-eddy simulation. *J. Atmos. Sci.,* **55,** 2666–2689.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An Intercomparison of Methods for Finding Coupled Patterns in Climate Data. *J. Climate*. **5**, 541-560.

Canny, J., 1986: A Computational Approach To Edge Detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **8**, 679-714.

Collins, W., and P. Tissot, 2007: Use of an Artificial Neural Network to Forecast Thunderstorm Location. Preprints, 5th Conference of the Artificial Intelligence Applications to the Environmental Sciences. San Antonio, TX, Amer. Meteor. Soc.

Jiang, H. and W. Cotton, 2004: Soil moisture estimation using an artificial neural network: a feasibility study. Can. J. Remote Sensing, Vol. 30, No. 5, pp. 827-839.

Lynn, B. H., W. Tao, and F. Abramopoulos, 2001: A Parameterization for the Triggering of Landscape-Generated Moist Convection. Part I: Analysis of High-resolution Model Results. *J. Atmos. Sci.,* **58,** 575–592.

Orville, R. E., 1991: Lightning ground flash density in the contiguous United States—1989. *Mon. Wea. Rev.,* **119**, 573-577.

Rogers, E., T. L. Black, D. G. Deaven, and G. J. DiMego, 1996: Changes to the operational "early" Eta analysis/ forecast system at the National Centers for Environmental Prediction.*Wea Forecasting,* **11,** 391–413.

SAS Institute Inc., SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999.

Segal, M., R. Avissar, M. C. McCumber, and R. A. Pielke, 1998: Evaluation of Vegetation Effects on the Generation and Modification of Mesoscale Circulations. *J. Atmos. Sci.,* **45**, 2268-2292.

Zeng, Y., F. Zhaodong, and N. Xiang, 2004: Assessment of soil moisture using Landsat ETM+ temperature/ vegetation index in semiarid environment. *International Geoscience and Remote Sensing Symposium (IGARSS)*, Vol. 6, pp. 4306-4309.

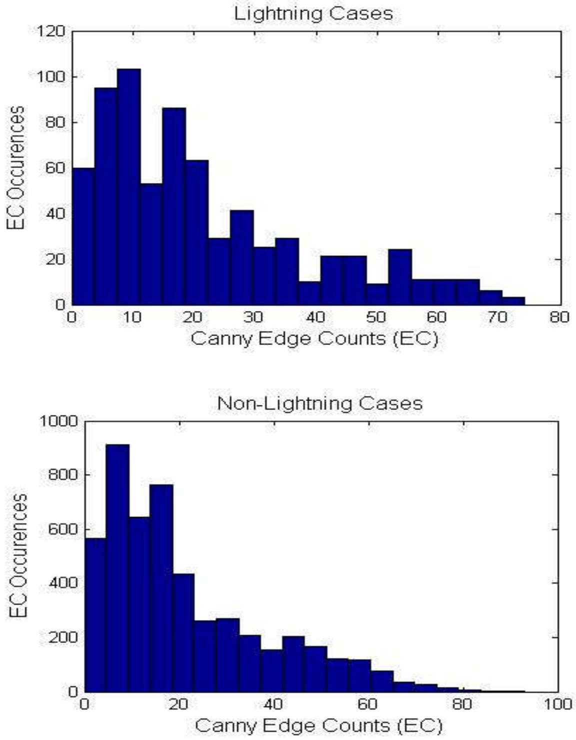

Figure 2: Histograms of **SMI** Canny Edge Counts (EC) from "favorable" data set, with (top) and without (bottom) lightning.

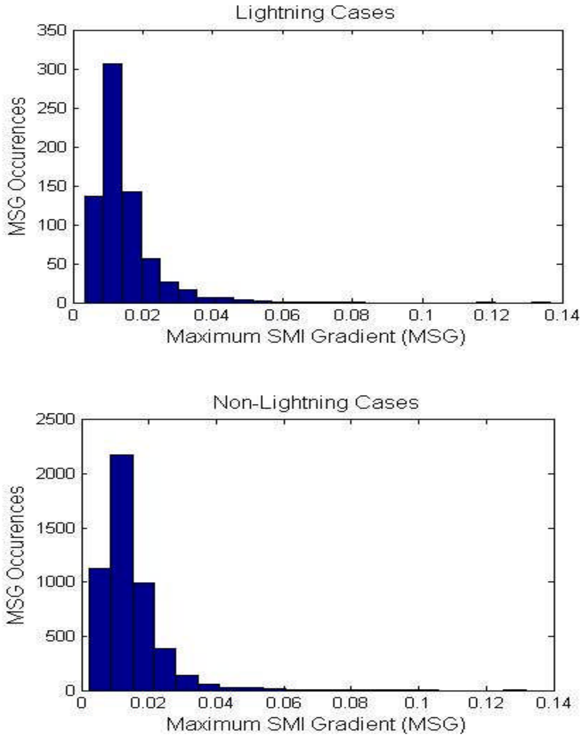

Figure 3: Histograms of **SMI** Maximum Gradient (MG) from "favorable" data set, with (top) and without (bottom) lightning.

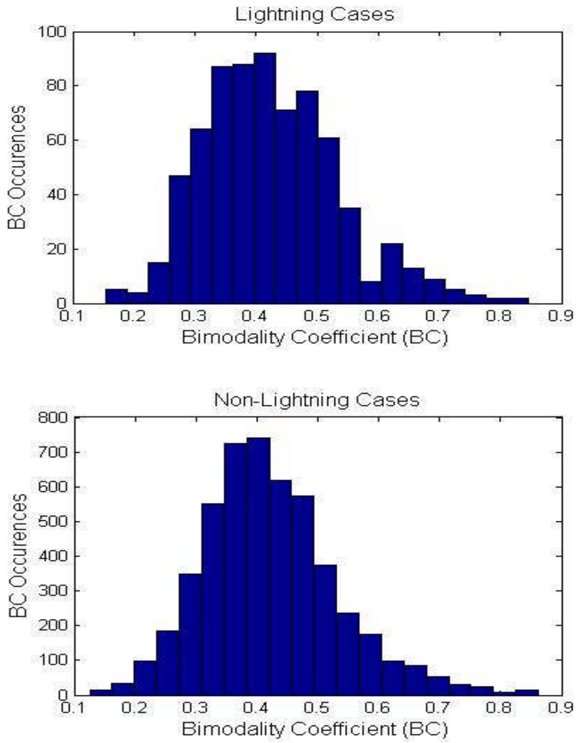

Figure 4: Histograms of **SMI** Bimodality Coefficient (BC) from "favorable" data set, with (top) and without (bottom) lightning.

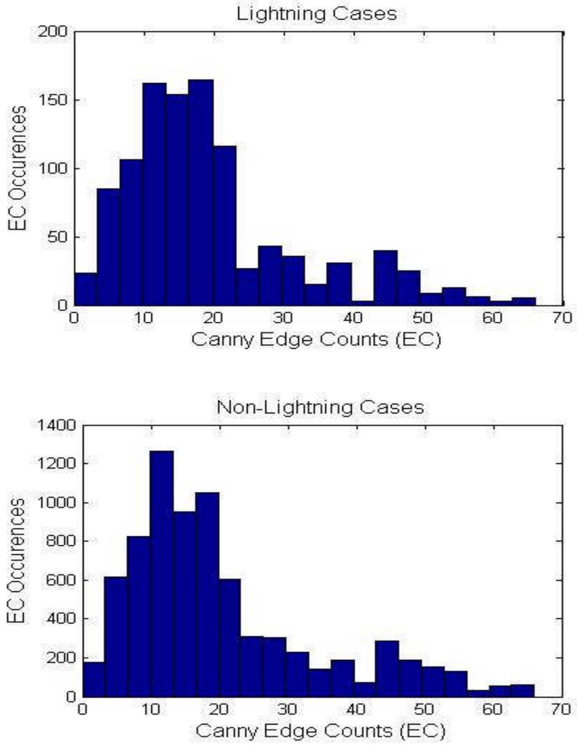

Figure 5: Histograms of **8-day LST** Canny Edge Counts (EC) from "favorable" data set, with (top) and without (bottom) lightning.

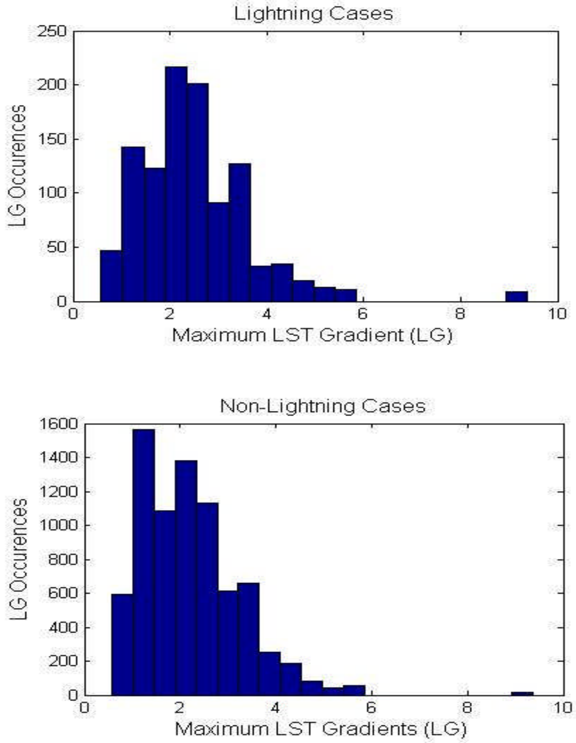

Figure 6: Histograms of **8-day LST** Maximum Gradient (MG) from "favorable" data set, with (top) and without (bottom) lightning.

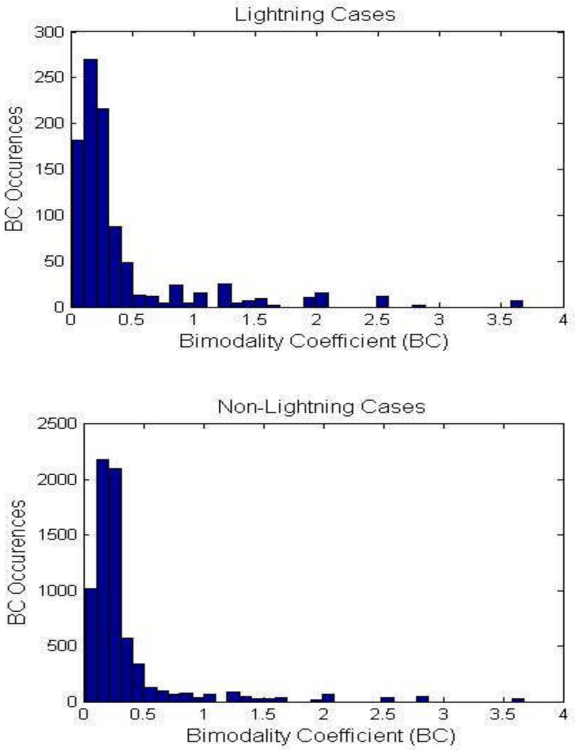

Figure 7: Histograms of **8-day LST** Bimodality Coefficient (BC) from "favorable" data set, with (top) and without (bottom) lightning.