AUTOMATED CLASSIFICATION OF STORM FLASHING / NON-FLASHING CONDITION FROM MICROPHYSICAL AND ENVIRONMENTAL OBSERVATIONS

D.J. Boccippio

NASA / Marshall Space Flight Center, SD-60, Marshall Space Flight Center AL 35812

ABSTRACT: Three years of paired TRMM radar, passive microwave (PM), lightning and NCEP reanalysis data are used to answer the question: “Can we classify a storm as flashing, or not, based only on microphysical and environmental observations?” Univariate, rule-based threshold models are tested, as well as multivariate linear (discriminant analysis) and multivariate nonlinear (neural network) logistic regressions. Simple reflectivity threshold rules achieve, over tropical land, 75% unbiased true skill (HSS), 79% probability of detection (POD) and 25% false alarm ratio (FAT), although their performance varies geographically. Multivariate neural network models achieve, over tropical land, 80% HSS, 82% POD and 19% FAT, exhibit much less geographic bias, are more robust over a range of decision thresholds, and are not overfitted. Automated classification from PM observations (DMSP/SSM-I, TRMM/TMI, Aqua/AMSR-E, NPOESS/CMI) is feasible and opens the possibility of a 20-yr “backfilled” global lightning climatology.

INTRODUCTION: Recent advances in ground-based and orbital lightning detection now provide some form of direct lightning observation over most of the globe [Christian et al., 2003]. However, these observations are often highly undersampled (e.g., 0.3% of all space-time from NASA’s Low-Earth-Orbiting optical detectors) or taken with very low detection efficiency (ground network observations at distant range). Needs (e.g., aviation) still exist for robust and automated classification of lightning hazard in remote regions, in absence of direct observations. Field programs have suggested that storm microphysical properties, e.g., radar reflectivity in the mixed phase region, might serve as proxies for lightning occurrence [e.g., Larsen and Stansbury, 1974; Petersen et al., 1996]. The Tropical Rainfall Measuring Mission (TRMM) satellite simultaneously measures radar reflectivity, PM brightness temperature and lightning flash rate, and provides a very large sample training/validation database for algorithm development, sufficient to both avoid model overfitting and generalize well beyond the local conditions of field programs.

Development of any classification model essentially involves specification of a function for the “decision surface”, in an n-parameter input space, between members of various categories; in this case, flashing and non-flashing storms. The simplest model is a scalar threshold placed on a single input (a univariate rule-based model). Univariate or multivariate rule-based models have the benefit of being easily understandable, but are limited in their ability to robustly parameterize the decision surface. Multivariate regression is a reasonable way to parameterize this surface, and indeed is the basis for traditional discriminant analysis (a multivariate, MSE, linear regression of the input parameters to define the surface). Classification neural networks are an example of multivariate, logistic nonlinear regressions, which allow both interactions between, and nonlinear transformations of, input variates, without a priori specification of model functional form. In this study the simplest (rule based) and most sophisticated (neural network) decision surface parameterizations are compared.

METHODOLOGY: Three years (12/97–11/00) of TRMM radar and PM observations from 35S-35N are used as input data, while TRMM lightning data are used as model truth. NCEP reanalysis [Kalnay et al., 1996] environmental variates (both surface and tropospheric) are also used as inputs; we allow the TRMM satellite to “virtually” sample the NCEP 4-D atmosphere and tag each storm observation with climatological (20-yr monthly mean) and concurrent (daily anomaly) environmental data. Storms are defined using the University of Utah “precipitation feature” approach [Nesbitt et al., 2000]; loosely, a storm is defined as contiguous raining TRMM radar pixel columns with adjacent anvil overhang, if any. Storms are only considered during “warm” seasons; the NCEP surface temperature at the storm time/location must be greater than 15C (arbitrarily chosen). Additionally, the 0C level must lie between 3-6 km altitude (thus excluding high mountain storms), and only cold-topped (15 dBZ tops < 0C) storms are considered. A total of 519,425 land storms and 1,004,033 ocean storms were considered. A storm is considered “flashing” if IC or CG lightning is detected within 10 km of any of its radar columns. Given the TRMM/LIS detection efficiency, “flashing” storms are thus those with flash rates > ~1 fl/min [Boccippio et al., 2002].

Radar-based inputs include the temperature of (15, 20, … 40 dBZ) echo tops (with nonoccurrence set to the surface temperature), maximum reflectivity at (0, -5, … -40C) expressed as Ze^{-0.5}, total area of the storm colder than ~10C and greater than (20, 25 … 40 dBZ), total volume of the storm meeting the same criteria, and total ice mass in the storm, using a simple Z-M relationship [Petersen and Rutledge, 2001]. The percentages of the storm classified by the TRMM 2A23 algorithm as convective, stratiform and non-raining were also input. PM-based
inputs include the minimum 85 and 37 GHz brightness temperatures, the total storm area at 85 GHz colder than 275, 250, 225 and 200K, and at 37 GHz colder than 280, 270 and 260K. NCEP inputs include surface temperature, pressure, sensible and latent heat fluxes, and Bowen ratio, lower troposphere lifted index and lapse rate (850-500 mb), column precipitable water, relative humidity at the surface, 850, 500 and 300 mb, vertical pressure velocity at 850, 500 and 300 mb, and tropopause pressure.

Radar-based threshold rule models were tested for the maximum reflectivity at each temperature level; the model consisting of the temperature level (input) and reflectivity decision threshold yielding the highest unbiased true skill was selected as the “best” univariate model. Neural networks including radar-only, radar+NCEP, PM-only, PM+NCEP, and radar+PM+NCEP inputs were trained. The latter (most sophisticated) network included 87 inputs and 15 hidden nodes, and used a cross-entropy rather than MSE error minimization criterion (appropriate for dichotomous flashing/non-flashing outputs [Marzban and Witt, 2001]). Network structure and training hyperparameters were not optimized; the networks shown are simply reasonably well-performing models, not the best possible. Overfitting was prevented by splitting the storm cases into 2/3 and 1/3 subsets. The 2/3 partition was used for training and validation; it was repeatedly split in half, with half used for model parameter estimation and half used to assess the onset of overfitting. After 25 such splits, an optimal convergence criterion was determined and the entire 2/3 partition used for parameter estimation. The remaining 1/3 partition was reserved and used to test the fitted models; <1% differences in skill scores were found between the training/validation and testing data, confirming that the nonlinear models were not overfitted.

Performance was assessed using conventional contingency-table based statistics [Marzban, 1998]: the Heidke Skill Score (loosely, “true” skill after correcting for correct diagnoses due to chance), the Gilbert Score (loosely, “threat” skill in real or forecast lightning situations, after correcting for correct diagnoses due to chance), the Probability of Detection (POD) and the False Alarm Ratio (FAT). Skill is also diagnosed in a non-scalar fashion by constructing Receiver Operating Characteristic (ROC) plots; i.e., parametric plots of POD vs FAT as all possible decision thresholds on the model outputs are considered. ROC plots help show model robustness (some end-users, e.g., may be constrained to use a model which limits the maximum allowable FAT, rather than maximizing HSS). The total area under ROC curves provides an additional scalar measure of model skill and robustness.

RESULTS: The formally “optimal” maximum-reflectivity based univariate threshold rule (yielding the highest HSS) is found to be: maximum dBZ @ -20C > 29 dBZ. This yields unbiased true skill (HSS) of 0.80, unbiased threat skill (GS) of 0.60, POD of 0.79 and FAT of 0.25, over land within the experiment domain. Practically, however, maximum reflectivity at any level from -14 to -21C, with thresholds from 33 dBZ to 28 dBZ over this range, yield virtually the same skill scores. This is remarkably similar to the “rules-of-thumb” gleaned from individual field programs (30-35 dBZ occurrence in the mixed phase region) which typically consider much smaller sample sizes, although optimal performance is strongly dependent upon isolating the best input and threshold in the mixed phase region. Fig. 1 shows the skill score performance as thresholds are varied. Over oceans, the “optimal” rule is: maximum dBZ @ -15C > 35.5 dBZ (effectively a “stronger” criterion than at the equivalent temperature level over land); this yields HSS/GS/POD/FAT of 0.59 / 0.42 / 0.59 / 0.40, significantly lower skill.

Training of a neural network including all radar and NCEP inputs improved skill scores, over land, to 0.80 / 0.66 / 0.80 / 0.18, effectively lowering the false alarm incidence by 7% (and hence improving the threat skill) without compromising the probability of detection. Over ocean, this approach
CONCLUSIONS: A very large sample (1.5 million) storm dataset confirms that microphysical parameters can be used to estimate thunderstorm occurrence / non-occurrence, and confirm the importance of high reflectivity in the mixed phase region. Simple, single-input threshold rules yield fairly high skill but may not generalize well (they exhibit geographic performance bias). Multivariate nonlinear (neural network) regression models can yielded scores of 0.64 / 0.47 / 0.64 / 0.36, an improvement in all skill scores of 4-5%. The area under the ROC curves of land and ocean NN models was 6 and 8% greater than for rule-based models, demonstrating an improvement in overall robustness.

Passive-microwave based storm classification is perhaps of greater interest than radar-based classification, as LEO satellite-based PM sensors have been standard orbital instruments for several decades, and will continue to be in the foreseeable future. Fig. 2 shows the ROC curves for a variety of univariate and multivariate linear and nonlinear classification schemes, using PM or PM+NCEP inputs. A neural network with both PM and NCEP inputs provides 6-9% gains over a single PM input (min 85 GHz FCT), and a significant portion of these gains are attributable to the ability of environmental parameters to statistically “assist” in the interpretation of PM brightness temperatures. Importantly, single-input PM models, while yielding reasonable global skill, are geographically biased over land, with too many false alarms over “maritime”-like regions such as the Amazon and Maritime Continent, and too many misses over arid regions. Inclusion of the NCEP inputs in a NN significantly mitigates this geographic bias (not shown). Skill scores for the (highest HSS) NN model over land are 0.75 / 0.60 / 0.77 / 0.24, 3-6% worse than with radar-based inputs – but still yielding a useful model. Over oceans, the NN yields up to 13% gains over a single-input model, although the best skill scores are only 0.55 / 0.38 / 0.52 / 0.41. ROC area gains for land and ocean NNs were 8 and 12% for land and ocean, respectively. The land results also demonstrate that the NN is capable of retaining 20-25% higher POD if a low FAT (e.g., 5-10%) is required. Thunderstorm classification using PM inputs (which convolve responses from both precipitation and cloud ice) may be “confused” by the dominance of cloud ice in deep ocean storms; alternatively, the large number of ocean storms likely flashing at rates below the LIS detection threshold of 1 fl/min may compromise the “truth sample” used to train the network.

Fig. 2: ROC curves [parametric plots of POD vs FAT as the decision threshold on model output (probability of lightning) is varied from 0-1], for various passive-microwave based model types.

Fig. 3 shows results from a “kitchen sink” NN accepting all radar, PM and NCEP inputs. While an artificial case (such a set of inputs is, practically, only available from the TRMM satellite itself), it is instructive of the “best” model skill we might find, barring further optimization of the NN hyperparameters. The NN outputs a probability that a storm is flashing, and Fig 3a shows the distributions of storms truly flashing and nonflashing at each output probability. The NN correctly assigns very high (very low) probabilities to a large number of truly flashing (nonflashing) storms, indicating that a vast majority of cases are relatively easy to diagnose. A small number of truly flashing storms are suspiciously classified with near-0% probability; these are likely storms for which the 10-km proximity “matching” criterion incorrectly assigned lightning to the radar-defined storm. Fig. 3b shows the skill scores vs decision threshold (probability of lightning) for this NN. The model is very robust, as shown by maintenance of high skill scores over a wide range of end-user applied decision thresholds. The ROC curve for this model is shown in Fig 3c.
improve diagnosis skill by 5-10%, generalize better than rule-based models, and are more robust over a wide range of possible end-user decision criteria. Inclusion of environmental variates from the NCEP reanalyses helps models statistically “interpret” lumped ice measures such as reflectivity and brightness temperature. Successful development of models using passive microwave inputs suggests that long-term “backfilling” (and future diagnosis) of global lightning occurrence / non-occurrence from LEO satellite observations is feasible.

ACKNOWLEDGEMENTS: The author thanks Caren Marzban for helpful discussions regarding NN interpretation, the University of Nevada at Reno for development of the NevProp NN package used in this study, and Steven Nesbitt for development of the reduced TRMM PR/TMI dataset used as inputs.

REFERENCES:


