

LCC / Anvil Electrification Hazard - Preliminary Analysis

Dennis J. Boccippio NASA/MSFC SD-60

1. “PHILOSOPHICAL” OBSERVATIONS

Identification of an anvil electrification hazard (through analysis of in-situ and radar observations) is a “classic” binary classification problem. A training dataset determines hazard (presently defined as $E > 3$ kV/m) and a candidate set of inputs (radar observations) must be used to model a decision surface between no-hazard/hazard conditions.

Any such classification problem can be conveniently generalized as a (potentially) multivariate, categorical regression problem. The problem requires three main design steps: (1) identification of appropriate input parameters (subselection of available observations, or transformation of them), (2) selection of a regression model, (3) identification of a decision criterion upon the outputs of the regression model yielding a problem-appropriate tradeoff between Probability of Detection [POD] (loosely, safety) and False Alarm Ratio [FAR] (loosely, cost). From the briefing given on the problem at hand, it appears that significant effort has gone into steps (1) and (3), while by mutual consent step (2) (model selection) has been limited to simple, univariate threshold rules. This study seeks to “flesh out” step 2 a bit to determine whether more complicated models warrant consideration. It is acknowledged that powerful operational drivers nudge model selection towards single-input threshold rules; however, it appears that this model selection is not necessarily mandated.

1.1 SOME PERSONAL OPINIONS

On step (1): This is the most appropriate step for expert physical knowledge to come into play, especially given that we do not have an explicit process physics model relating radar inputs to electrification outputs (we are not, e.g., fitting coefficients in a population biology model where the dynamics are assumed to be known). Since we lack full understanding of process physics, and our input observations are not state variables themselves, the problem is fundamentally *empirical*, and the best we

can do is guide ourselves towards an optimal solution by appropriate selection/transformation of the inputs.

I’ll have a lot more to say on step (2) later.

On step (3): The current LCC review committee structure seems to hand decision-making authority on the POD/FAR tradeoff to the technical experts. Since this is essentially a safety/cost tradeoff, this is arguably a decision appropriate for management, rather than technical experts, to make (if the two overlap, all the better). This observation is especially relevant since the review team has already identified the appropriate “tool” to provide the necessary information on the POD/FAR tradeoff to management: the Receiver Operating Characteristic (ROC) curve, which parameterically displays POD vs FAR for all possible decision thresholds in a fitted model. This tool is (and should be) the *primary* means of assessment of model performance.

1.2 CLASSIFICATION PRIMER

At the risk of being simplistic and pedantic, a quick primer on classification may be useful to get us all “speaking the same language”. The following illustration has been helpful in illustrating classification problems to students.

Consider a room. There exists a master plan for this room to be filled with sand and rocks. A work-

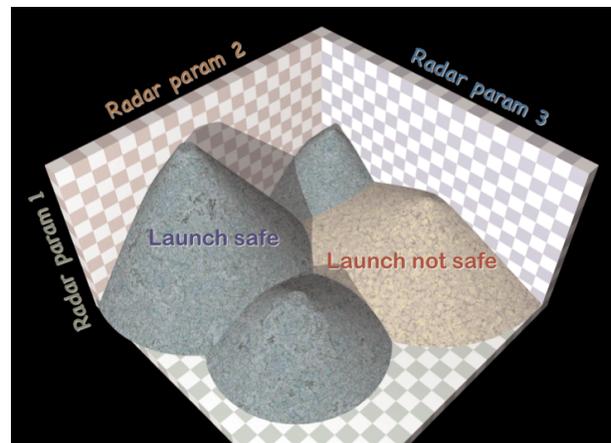
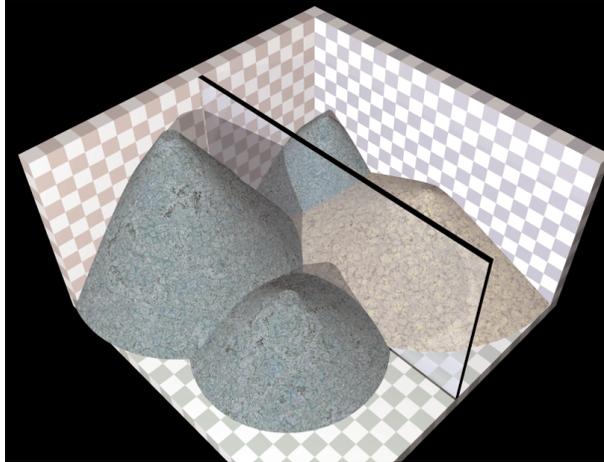
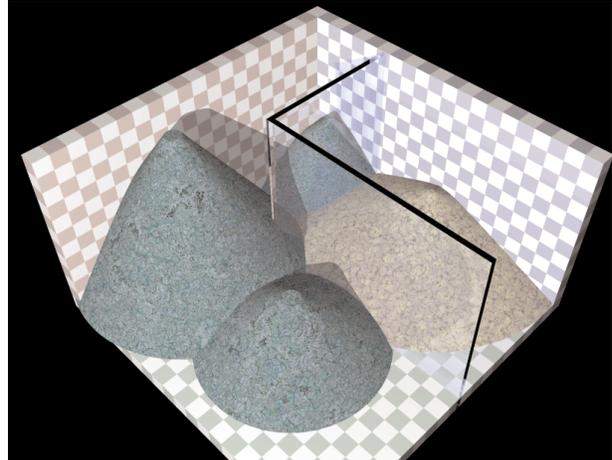


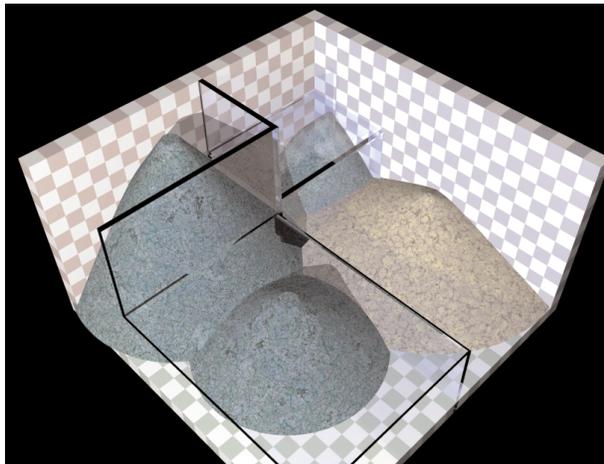
Figure 1 : Conceptual problem.



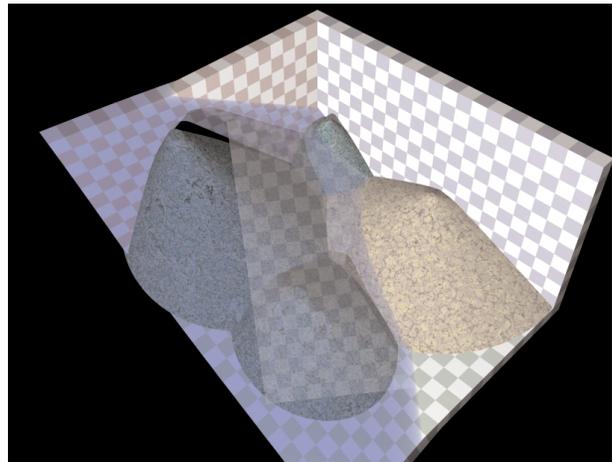
(2a) A univariate threshold rule.



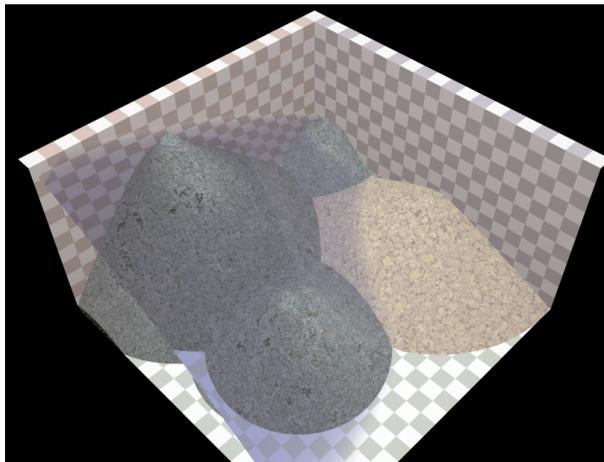
(2b) A bivariate threshold rule.



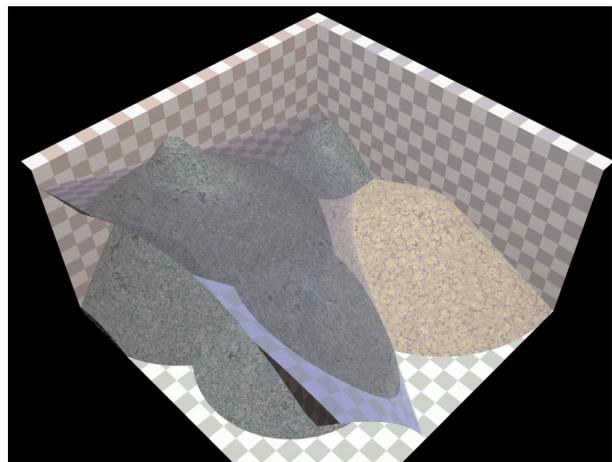
(2c) A multivariate threshold (logical) rule, sometimes known as a decision tree.



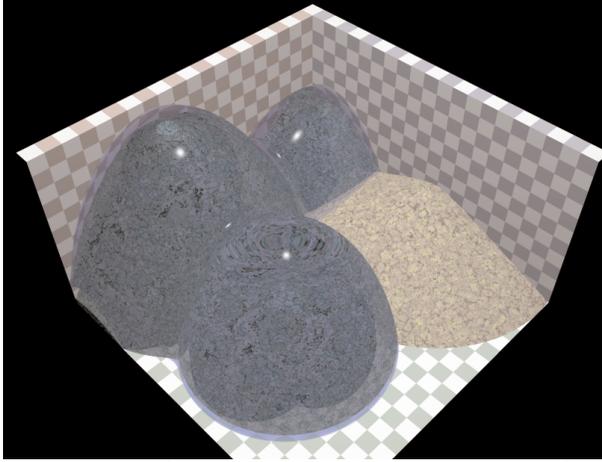
(2d) A linear multivariate regression with linear basis functions. Also known as “discriminant analysis”.



(2e) A linear multivariate regression with nonlinear basis functions. (e.g., $ax^2+by^2+cz^2$)



(2f) A nonlinear multivariate regression with “interaction” terms (e.g., $axy + bxz + cyz$)



(2g) A nonlinear multivariate regression with very “relaxed” basis functions (e.g., a neural network)

man is hired to lay out some sand and rocks according to the master plan. He does not have enough of either to fill the entire room, and then goes away. (He may also be a bit sloppy in his adherence to the plan). We are then given a sheet of plexiglass, and asked to recreate the master plan by creating a boundary between “sand areas” and “rock areas”.

To make this “concrete” (inadvertent pun): “sand area” = “anvil electrification hazard”; “rock area” = “anvil electrification safe”; the room axes correspond to three possible input (radar) observations; the plexiglass corresponds to a decision surface; how we use it corresponds to our model selection; where we slide it corresponds to POD/FAR trade-off. Figure 1 illustrates the analogy.

Figure 2a illustrates a simple univariate threshold rule, the model initially selected for this problem. Attention is restricted to one possible input, and the plexiglass is slid back and forth to determine and optimum “threshold” (decision surface). Figure 2b illustrates the case if we “cut” the plexiglass and arrange the panels orthogonally: a bivariate threshold rule. This process can be extended indefinitely, creating a “Lego world” (2c) in which the hazard areas are increasingly “boxed in”. When automated, this type of model is also known as a decision tree. It appears that the bulk of the weather-related LCC fall into “Lego world” model form. (As an aside, a Lego world model with slightly polished-off / rounded edges could be a reasonably approximate illustration of a fuzzy logic / expert system model).

Carried to its extreme, it is apparent that “Lego world” is simply a crude way of parameterizing what is actually a smooth, complicated and likely nonlinear boundary in the input parameter space. It is natural to wonder if smoother parameterizations exist. We could simply “tilt” the plexiglass (2d). This would be a linear multivariate regression with linear basis functions (the three radar inputs), and is also known as “discriminant analysis”. We could get fancy and bend or even flex the plexiglass, though still be limited by its material properties; this would be an example of linear multivariate regression with nonlinear basis functions (2e) or nonlinear multivariate regression (2f; if the plexiglass were sufficiently “wobbly”). An important point here is that the constraining material properties of the plexiglass (the mathematical form of the chosen regression basis functions) has nothing to do with the data distribution itself - again, by counterexample, we are not fitting coefficients to a population biology model. In our case, the choice of appropriate nonlinear basis functions or interaction terms would be, at best, obscure.

A final and novel solution would be to get around the annoying “material properties” of the plexiglass by melting it (2g). This would be a nonlinear multivariate regression with somewhat “arbitrary” or “very flexible” basis functions. It is essentially what a classification neural network does, and it would be best to ignore any emotionally loaded reactions against neural networks by simply thinking of them as nonlinear regressions. The caution (as with any nonlinear regression) is that we would not want to melt the plexiglass “too much”, or it would overfit the training data. Well-established bootstrapping techniques exist to avoid this.

The illustration is intended to drive home several basic points: (1) The decomposition of the problem into definition of inputs (room axes), model selection (what to do with the plexiglass) and decision threshold selection (where to slide the plexiglass). (2) The possible importance of considering multivariate models, especially when the data suggest their importance, and when the inputs are readily available. (3) The intuitive appeal of rule based models rapidly vanishes if we consider multiple inputs. Drawing a “wall” or “box” in one or two dimensions is relatively easy. Drawing “boxes” in many dimensions is not, and not necessarily intu-

itively superior to other models. [End, for now, rant against decision tree models].

In this study, models of the form (2a), (2d) and (2g) will be considered.

1.3 CAVEAT

Those with experience in statistics and classification will already recognize that I'm using somewhat "fuzzy" language to describe these concepts, and will observe that I'm glossing over some formal details in the analysis below. This is a "communication design tradeoff" to get the primary points across (and a byproduct of generating a rapid turnaround analysis at a late stage in the game). I'm aware of the formal deficiencies in my "fuzzy" approach, but from experience with similar classification problems, do not believe that the primary results would significantly change given a more exhaustive treatment.

2. DATASET OBSERVATIONS

The dataset contains ~2600 observations at *pixel* level, collected over a much smaller number of *cases*. After rejection of bad data points (pixels where a significant number of input parameters are missing), this yields about 2600 WSR and 2000 NEXRAD observations. Of these, 15-16% (WSR) and 12-13% (NEX) pass $E > 4$ kV/m / $E > 3$ kV/m hazard criteria (the alternate 4 kV/m E hazard criterion will be discussed below). The problem is thus not necessarily "rare event", but neither is it a "balanced" classification problem (comparable numbers of events and non-events). The dataset size and rarity of hazards place fundamental limits on the complexity of nonlinear models that could be applied.

Several key parameters have significant numbers of missing observations. This tends to occur most often with the 0 dBZ-thresholded values, and ACIntSum is a good example. For the preliminary analysis, these data were "repaired" by filling them in with the non-thresholded values (if available). In the case of ACIntSum, if even the non-thresholded value was missing, it was replaced by the TotSum, which is supposed to be highly analogous and spans the same dynamic range.

The input parameters are acknowledged to be covarying and in some cases (by design) highly

collinear. (This is relevant in the interpretation of weights in some of the multivariate linear models, below).

Preliminary tests (using ROC curves as diagnostics) suggested that 0-thresholded inputs consistently outperformed non-thresholded inputs, and the multivariate models considered below all use 0-thresholded inputs.

Preliminary tests also revealed a tendency for many models' ROC curves to "fall apart" near $POD \sim 1$. I.e., the "order of performance", gauged by the overall model robustness for $POD < 1$, got very confused near $POD \sim 1$. This has two immediate possible implications: (1) the hazard criterion ($E > 3$ kV/m) may be too lenient, and the models have a difficult time being "flexible" enough to accommodate marginal cases, or (2) a small number of truly noisy or poor input or E observations cause the models to "go through unnecessary contortions" (at the expense of FAR) to accommodate this noise. [(1) and (2) may also be related]. As a result:

(a) In addition to FAR @ $POD = 1$ (nominally the most "conservative" metric), FAR @ $POD = 0.995$ and FAR @ $POD = 0.990$ are reported for each model. I will argue that since these curves are constructed from *pixel* data, we could even look at FAR @ $POD =$ 'much lower' as a model discriminator, since the underlying *problem* is detection of a hazardous anvil, rather than every hazardous pixel, and even marginal or noisy pixels are likely to have more "useful" pixels nearby. (As an aside; with ~ 85 or so transects in the 1-yr dataset, POD/FAR statistics on *feature* hazard detection could easily be run). For now, I will err on the side of caution and report only FAR @ $POD = 0.990$, although the full ROC curves are presented for the most promising models.

(b) Hazard has been alternatively defined (again erring on the side of caution) as $E > 4$ kV/m. In general, this yielded superior performance over an $E > 3$ kV/m criterion. In the preliminary writeup, results are only reported for the 4 kV/m hazard criterion. It is for the committee to decide whether that redefinition is acceptable.

Finally, reviewing the initial writeups, it appears that *all* models trained to date were trained on the entire input dataset - no data were reserved for independ-

ent validation. For intercomparability I will continue that approach, with the understanding that all models discussed are likely overfitted and the POD/FAR results thus overly optimistic. A “final” solution should really reserve about 1/3 of the data for independent assessment. The current approach should be reasonable for coarse intercomparison *between* models, although that is an *assumption*.

3. APPROACH

Three *types* of models are considered here: (1) univariate threshold rules, (2) multivariate linear models, (3) multivariate nonlinear models. All have been tested using a common framework, configuring the models as neural networks. [A multivariate categorical linear regression is simply a neural network with no hidden layers, and a univariate threshold rule can be shown to map to a univariate linear regression, or neural network with one input and no hidden layers]. The results should be completely equivalent to (1) conventional rules, (2) discriminant analysis, (3) neural networks. Each model yields as an output (prediction) of the probability P (from 0-1) that a set of input observations yields a hazardous E-field. The ROC curves are created by examining POD vs FAR as the *decision threshold* for a hazard/no-hazard call based on the output (predicted) probability is varied from 0-1. Each curve represents a model; each point on the curve represents a different decision threshold. (Note that the outputs of a simple univariate threshold rule can easily be transformed to probabilities as well based upon the training sample, without using my neural network implementation framework. Having models yield probabilities as outputs is, in general, a cognitively useful thing to do...)

Three types of multivariate cases were run. A “kitchen sink” neural network (all 30 inputs) was trained to illustrate the “limits of predictability” in the input data. Obviously, computation of 30 inputs is operationally prohibitive; the “kitchen sink” network is simply a (likely overfitted) best case scenario or “tall pole”.

Recognizing that many of the 30 inputs are essentially redundant, 9-input nonlinear and linear models were tested using the 0-thresholded variants of: ColSum1, Thick11, Top11, Base11, Avg11, Frac11, SumAvg11, ACIntSum11 and TotSum11. This is a more reasonable (though still likely operationally

infeasible) “tall pole” based on my guess at “most physically distinct, interesting or relevant” inputs.

To test multivariate models that *could* be operationally feasible, 3-parameter linear models were tested for every combination of 3 of these 9 inputs (84 tests). A 3-parameter discriminant analysis should be a viable model to implement. For comparison, a nonlinear neural network was trained using the 3 inputs from the *best* 3-parameter linear model.

Finally, univariate linear models (which reduce to univariate threshold rules) were tested for the 9 unthresholded and 9 thresholded parameters described above. This set thus includes some of the tests already run by the committee, with the difference that I am using $E > 4$ kV/m as the hazard criterion.

4. FINDINGS

A preliminary finding is that the WSR and NEX results *must* be considered separately. Optimal predictors differ significantly between the two input sources, and the committee should resign itself to selection of two different, radar-appropriate models. The results are thus presented separately below. If desired, I could attempt to train an optimal model using the inputs from both radars as redundant data-points, and we could explore what the performance “hit” would be in seeking a “unified” model. I suspect it will be significant. Note that this is different than consideration of a multivariate model using inputs from *both* sensors, which would almost certainly improve performance at the expense of operational complexity.

4.1 WSR RADAR INPUTS

Figure 3 presents the ROC curves (zoomed in to $POD = [0.85...1.00]$) for: ‘interesting’ unthresholded univariate rules (Avg*, ACIntSum*), ‘interesting’ univariate 0-thresholded rules (ACIntSum, Avg), two of the best 3-parameter linear models (ACIntSum, Frac, Base; ACIntSum, Frac, Top), a 3-parameter neural network (ACIntSum, Frac, Base), 9-parameter linear and nonlinear models, and the “kitchen sink” 30-parameter nonlinear model. The inset numbers correspond to FAR @ $POD = 0.990$, shown by the dashed line. Note that the plot illustrates the point made earlier: the overall sequence of model “robustness” at $POD < \sim 0.98$

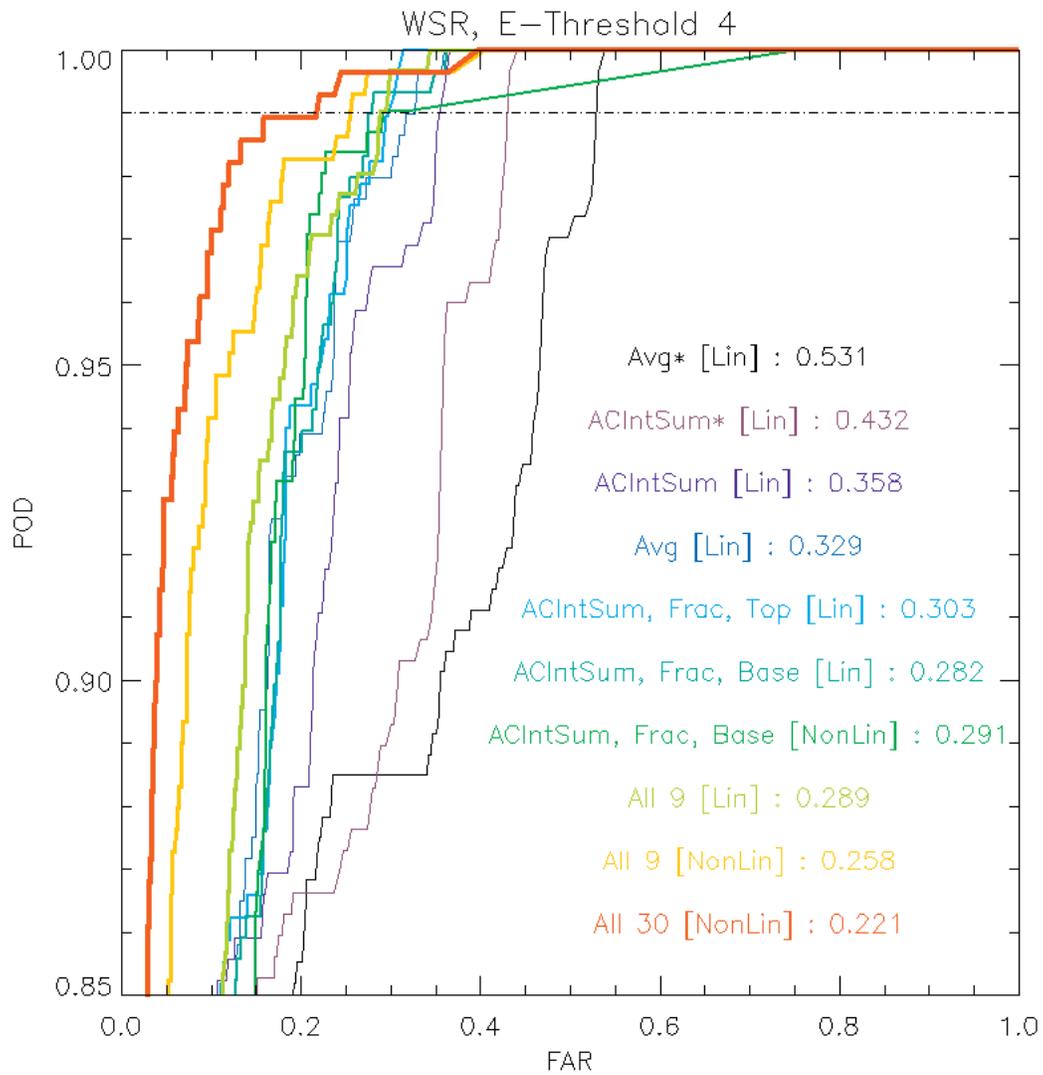


Figure 3: ROC curves (POD vs FAR as a function of decision threshold) for the “best” models trained on WSR data with an $E > 4$ kV/m hazard threshold. Inset numbers show FAR @ POD = 0.990 (dashed line). The full ensemble of model runs is shown in Table 1.

ANVIL LCC PRELIMINARY REPORT

Experiment	VSR_4KV			MeanFAR	Color Scale									
	FAR @ POD=1.000	FAR @ POD=0.995	FAR @ POD=0.990		0-1	1-2	2-5	5-10	10-20	20-40	40-60	60-80	80-90	90-100
					CoSum1	Thick	Top	Base	Avg	Frac	SumAvg	ACInt	TotSum	
All30NN	0.397	0.244	0.221	0.287										
All9NN	0.404	0.274	0.258	0.312										
All9Lin	0.349	0.299	0.269	0.312										
70NN	0.750	0.750	0.291	0.597										
70	0.362	0.350	0.282	0.331										
81	0.336	0.315	0.297	0.316										
60	0.313	0.307	0.303	0.308										
47	0.365	0.337	0.307	0.337										
31	0.359	0.333	0.309	0.334										
79	0.377	0.357	0.313	0.349										
80	0.367	0.325	0.314	0.335										
37	0.391	0.391	0.314	0.366										
48	0.381	0.363	0.318	0.354										
19	0.332	0.327	0.322	0.327										
50	0.335	0.333	0.324	0.331										
36	0.387	0.356	0.324	0.356										
66	0.346	0.343	0.324	0.337										
68	0.350	0.337	0.325	0.337										
32	0.346	0.331	0.327	0.335										
67	0.359	0.344	0.329	0.344										
Avg	0.359	0.329	0.329	0.339										
49	0.366	0.329	0.329	0.342										
4	0.451	0.334	0.330	0.372										
40	0.357	0.342	0.340	0.346										
43	0.348	0.346	0.331	0.342										
12	0.373	0.356	0.331	0.353										
64	0.390	0.387	0.332	0.370										
22	0.434	0.354	0.333	0.374										
82	0.343	0.339	0.333	0.338										
76	0.355	0.341	0.334	0.343										
23	0.338	0.335	0.335	0.336										
83	0.349	0.349	0.335	0.344										
61	0.367	0.345	0.336	0.350										
10	0.413	0.385	0.337	0.378										
75	0.358	0.344	0.338	0.346										
53	0.359	0.356	0.339	0.351										
20	0.363	0.344	0.340	0.349										
17	0.358	0.341	0.341	0.347										
15	0.374	0.360	0.342	0.359										
45	0.386	0.366	0.342	0.364										
18	0.366	0.359	0.344	0.356										
39	0.365	0.347	0.345	0.352										
78	0.360	0.346	0.346	0.351										
65	0.368	0.351	0.347	0.355										
5	0.392	0.364	0.347	0.368										
71	0.370	0.355	0.348	0.358										
38	0.373	0.355	0.348	0.359										
77	0.379	0.354	0.349	0.361										
1	0.357	0.353	0.350	0.353										
35	0.375	0.355	0.352	0.361										
72	0.369	0.359	0.357	0.361										
ACInt	0.365	0.362	0.358	0.364										
63	0.379	0.367	0.361	0.369										
58	0.378	0.362	0.362	0.367										
TotSum	0.368	0.368	0.363	0.366										
54	0.381	0.377	0.363	0.373										
16	0.404	0.375	0.363	0.381										
52	0.367	0.367	0.364	0.366										
62	0.377	0.377	0.366	0.373										
33	0.376	0.372	0.366	0.372										
26	0.371	0.371	0.369	0.371										
Thick*	0.455	0.400	0.369	0.408										
55	0.385	0.375	0.369	0.376										
56	0.385	0.374	0.370	0.376										
57	0.387	0.379	0.371	0.379										
74	0.383	0.379	0.373	0.378										
42	0.403	0.391	0.378	0.391										
69	0.389	0.389	0.378	0.385										
6	0.385	0.381	0.378	0.382										
13	0.396	0.391	0.379	0.389										
84	0.395	0.391	0.381	0.389										
73	0.397	0.397	0.386	0.393										
51	0.398	0.395	0.391	0.394										
14	0.395	0.395	0.393	0.395										
2	0.520	0.442	0.394	0.452										
7	0.412	0.404	0.395	0.403										
34	0.479	0.404	0.396	0.426										
24	0.408	0.408	0.397	0.404										
8	0.436	0.417	0.399	0.417										
27	0.412	0.405	0.399	0.406										
25	0.425	0.409	0.402	0.412										
ColSum1	0.429	0.421	0.404	0.418										
41	0.443	0.411	0.404	0.419										
9	0.413	0.413	0.406	0.411										
11	0.413	0.407	0.407	0.409										
29	0.478	0.410	0.407	0.432										
28	0.426	0.414	0.409	0.415										
Frac*	0.496	0.425	0.417	0.446										
21	0.448	0.448	0.448	0.448										
59	0.470	0.450	0.450	0.457										
3	0.478	0.478	0.460	0.472										
46	0.623	0.623	0.469	0.571										
Thick	0.480	0.473	0.473	0.475										
SumAvg	0.508	0.488	0.478	0.491										
Top*	0.685	0.604	0.525	0.605										
Avg*	0.538	0.531	0.531	0.533										
Col*	0.648	0.618	0.549	0.605										
SumAvg*	0.741	0.692	0.650	0.694										
Base*	0.959	0.948	0.839	0.915										
Base	0.951	0.946	0.849	0.915										
Top	0.885	0.875	0.875	0.878										

Table 1

“falls apart” near $POD \sim 1$, suggesting that very good models can be found but only at the “expense” of handling likely marginal or noisy data cases poorly. I argue that this is only an “expense” in a world where attention is inappropriately focused too close to $POD=1$ (such a world may overemphasize poor data or ambiguity in the hazard definition itself).

Observe that at $POD \sim 0.99$, the models span a range of about 0.35 False Alarm Ratio. The problem of optimal model selection is thus a relevant one, and the gains of considering even “slightly” multivariate models should be weighed appropriately.

The entire ensemble of model runs is shown in Table 1. For each model, FAR @ $POD=1.000, 0.995$, and 0.990 is reported (the 3-parameter and 1-parameter models are sorted by FAR @ $POD=0.990$). The mean of these three FARs is also shown as an additional metric. The “relevances” of the inputs for each model are “painted” in the next 9 columns. Some caveats: these relevances are of limited use in the nonlinear models, and should be handled with caution even in the linear models, as collinearity among input parameters may be present. Nonetheless, the “big picture” is highly instructive.

For the WSR radar, the best “reasonable” (i.e., 3-parameter) models are essentially ACIntSum (thresholded) + a quality metric (Frac) + some nearly irrelevant 3rd input ... i.e., they are essentially 2-parameter models. This is fascinating in that it recollects what I believe was a complaint about the integrated parameters; that they were sensitive to scan gaps, which is precisely what Frac is reporting. “Below” these models lie a number of models with Avg as the primary input. Avg alone (thresholded) is the 17th best “feasible” model, which itself outperforms ACIntsum alone (much further down the list). If I understand correctly, significant discussion occurred about the scan gap issue. These results illustrate that while “reject the input” (univariate world) is certainly an option, “correct for it” (multivariate world) is a much better option.

As an aside, note that ACIntSum (univariate) indeed outperforms Avg when unthresholded, but the reverse is true (and both are better) when using thresholded inputs. In the context of “really good” multivariate models shown here, the unthresholded univariate models are ‘disastrous’. This should not

be surprising for two reasons: First, we can legitimately *physically* question the microphysical relevance of anvil layers with < 0 dBZ reflectivity. Second, and more importantly, the minimum detectable reflectivity of the WSR radar is highly range-dependent. Unthresholded inputs thus do not *mean* the same thing across the sampling domain. This can add unnecessary “effective” noise to the dataset. It also means that with limited training data, our models may easily contain implicit bias based on the circumstantial radar-relative locations of the anvils studied... not a good thing. In this case, more was *not* necessarily better.

Finally, note that the performance gains of significantly more multivariate *linear* models (“All9Lin”) are minimal for WSR data. The performance gains of significantly more multivariate *nonlinear* models are significant, but it would require significant effort to (a) verify that these models are not overfitted and (b) implement a neural network module in the operational system.

4.2 NEXRAD RADAR INPUTS

Before discussing Figure 4, a “big picture” comparison of Tables 2 and 1 is in order. Clearly, the order of “optimal” predictors differs significantly between NEXRAD and WSR inputs. For 3-parameter models, the “leaders of the pack” involve either Top (!), SumAvg or Avg. Also, while SumAvg (thresholded) alone is near the “top of the pack”, the best 3-parameter linear model (Avg, Top, Base) yields a .07 FAR gain over SumAvg alone; non-trivial (a 20% relative gain).

Figure 4 also reveals an interesting feature of the univariate models. First, thresholding appears to make little difference; not surprising given the low sensitivity of the NEXRAD radar itself. Second, while the Avg parameter yields good performance for $POD > \sim 0.98$, it is by no means the most robust model for $POD < \sim 0.98$; rather, ACIntSum appears superior (SumAvg actually does best). Recalling the earlier observations about overall model behavior near $POD \sim 1$, this should caution us that Avg may “only” be working there because it happens to accommodate marginal or anomalous data well (and indeed may do this at a cost of lower overall robustness). However, when “corrected” (for whatever reason) using information about cloud base and top, this deficiency is removed. Interestingly, in this

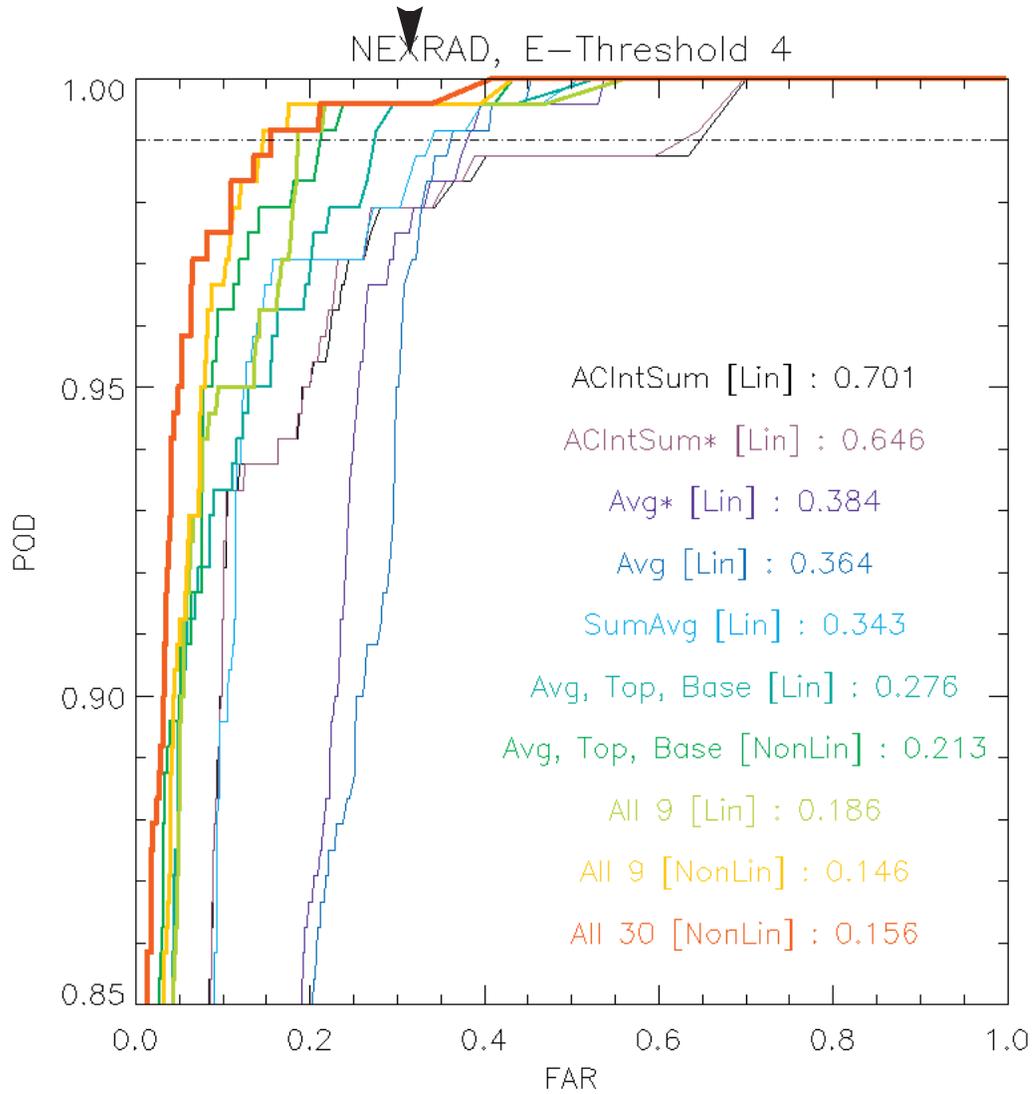


Figure 4: ROC curves (POD vs FAR as a function of decision threshold) for the “best” models trained on NEXRAD data with an $E > 4$ kV/m hazard threshold. Inset numbers show FAR @ POD = 0.990 (dashed line). The full ensemble of model runs is shown in Table 1.

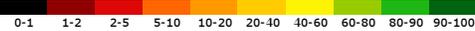
IIE_X_4KV													
Experiment	FAR @ POD=1.000	FAR @ POD=0.995	FAR @ POD=0.990	MeanFAR	CoSum1	Thick	Top	Base	Avg	Frac	SumAvg	ACInt	TotSum
All30III	0.406	0.212	0.156	0.258									
All9III	0.433	0.176	0.146	0.252									
All9Lin	0.561	0.216	0.186	0.321									
50III	0.433	0.237	0.213	0.295									
50	0.526	0.296	0.276	0.366									
52	0.559	0.314	0.303	0.392									
25	0.372	0.372	0.335	0.359									
SumAvg	0.509	0.397	0.343	0.416									
8	0.440	0.387	0.344	0.390									
41	0.454	0.356	0.356	0.389									
57	0.602	0.478	0.364	0.482									
Avg	0.454	0.408	0.364	0.409									
58	0.611	0.478	0.365	0.485									
55	0.500	0.389	0.366	0.418									
7	0.476	0.388	0.369	0.411									
59	0.542	0.431	0.370	0.448									
10	0.554	0.400	0.372	0.442									
6	0.481	0.481	0.373	0.445									
32	0.558	0.441	0.376	0.458									
64	0.514	0.428	0.378	0.440									
24	0.431	0.431	0.380	0.414									
31	0.516	0.443	0.380	0.447									
35	0.559	0.425	0.381	0.455									
54	0.584	0.418	0.382	0.461									
61	1.000	0.509	0.382	0.630									
Avg*	0.538	0.395	0.384	0.439									
9	0.506	0.398	0.385	0.430									
11	0.586	0.415	0.387	0.463									
30	0.558	0.458	0.388	0.468									
47	0.494	0.431	0.390	0.438									
33	0.551	0.434	0.390	0.458									
39	0.591	0.484	0.394	0.490									
84	0.487	0.467	0.395	0.485									
36	0.529	0.529	0.398	0.450									
Top*	0.614	0.519	0.399	0.511									
63	0.553	0.427	0.401	0.460									
60	1.000	0.463	0.402	0.622									
48	0.469	0.434	0.403	0.436									
37	0.546	0.454	0.411	0.470									
62	0.586	0.457	0.412	0.485									
66	0.753	0.495	0.414	0.554									
16	0.508	0.440	0.419	0.456									
14	0.524	0.463	0.419	0.469									
SumAvg*	0.499	0.468	0.422	0.463									
42	0.664	0.423	0.423	0.503									
12	0.597	0.436	0.436	0.490									
73	0.517	0.517	0.439	0.491									
69	0.489	0.440	0.440	0.456									
56	0.671	0.474	0.442	0.529									
53	0.551	0.446	0.446	0.481									
72	0.507	0.507	0.452	0.488									
28	0.498	0.478	0.459	0.478									
27	0.498	0.478	0.459	0.478									
18	0.551	0.466	0.466	0.494									
15	0.574	0.471	0.471	0.506									
75	0.661	0.661	0.491	0.604									
Top	0.678	0.650	0.503	0.610									
74	0.569	0.507	0.507	0.527									
3	0.510	0.510	0.510	0.510									
70	0.556	0.556	0.511	0.541									
67	0.592	0.513	0.513	0.539									
81	0.653	0.514	0.514	0.560									
71	0.517	0.517	0.517	0.517									
82	0.649	0.517	0.517	0.561									
51	0.575	0.575	0.518	0.556									
13	0.672	0.524	0.524	0.573									
44	0.678	0.537	0.537	0.584									
65	0.541	0.541	0.541	0.541									
29	0.709	0.709	0.594	0.671									
21	0.688	0.599	0.599	0.629									
Col1*	0.781	0.659	0.604	0.681									
Thick	0.756	0.756	0.616	0.709									
78	0.780	0.780	0.630	0.730									
TotSum*	0.678	0.678	0.639	0.665									
68	0.844	0.844	0.644	0.777									
ACInt*	0.698	0.698	0.646	0.681									
38	1.000	1.000	0.657	0.886									
26	0.777	0.657	0.657	0.697									
Col1	0.886	0.658	0.658	0.734									
Thick*	0.865	0.791	0.661	0.772									
79	0.664	0.664	0.664	0.664									
TotSum	0.748	0.748	0.672	0.723									
49	1.000	0.700	0.700	0.800									
4	1.000	0.700	0.700	0.800									
ACInt	0.701	0.701	0.701	0.701									
5	1.000	0.704	0.704	0.803									
2	1.000	0.716	0.716	0.810									
1	1.000	0.719	0.719	0.812									
34	1.000	0.728	0.728	0.909									
46	0.872	0.872	0.765	0.836									
22	0.901	0.809	0.809	0.840									
83	0.810	0.810	0.810	0.810									
23	0.906	0.819	0.819	0.848									
Frac	1.000	1.000	0.824	0.941									
17	1.000	0.836	0.836	0.891									
77	0.884	0.884	0.884	0.884									
40	1.000	1.000	0.887	0.962									
Base	0.983	0.979	0.968	0.977									
43	1.000	1.000	0.926	0.975									
20	1.000	1.000	0.931	0.977									
80	1.000	1.000	0.931	0.977									
Base*	0.987	0.983	0.970	0.980									
19	1.000	1.000	1.000	1.000									
45	1.000	1.000	1.000	1.000									
76	1.000	1.000	1.000	1.000									

Table 2

case, “correction” of ACIntSum in a 3-parameter model does not yield optimum performance.

Unlike with WSR data, a 9-parameter linear model yields a significant gain over 3-parameter models, reducing FAR @ POD = 0.990 to 0.186 (a 0.09 or 33% gain over 3-parameter models, and a 0.16 or 46% gain over univariate models). NEXRAD inputs, although poor quality individually, appear to have enough information content *collectively* to yield good performance.

5. PRELIMINARY CONCLUSIONS

- For WSR data, unthresholded inputs yield overall poorer performance than thresholded inputs (for very plausible reasons). For NEXRAD data, thresholding appears much less relevant (for very plausible reasons). Attention should be given to fixing the “significant missing data” problem in important WSR thresholded inputs.
- While not shown, an $E > 4$ kV/m hazard criterion yields higher performance than an $E > 3$ kV/m criterion. The committee should consider whether such a hazard criterion is acceptable.
- The fact that many models’ performance “falls apart” for $POD \gg 0.98$ suggests the presence of a small subset of marginal cases or noisy cases in the training dataset. This subset should be fairly easy to isolate and examine in greater depth to determine how important it is. Regardless, strong arguments (pixel detection vs feature detection) can be made to guide model selection by FAR at POD slightly below a strict 1.0 requirement.
- “Simple”, 3-parameter linear models (discriminant analysis) outperforms univariate threshold rules with a 0.05-0.07 FAR reduction at $POD=0.990$ (a 14%-19% relative gain). Given the costs of overwarning, these gains are nontrivial.
- For WSR data, the complaints about ACIntSum sensitivity to scan gaps can be mitigated by “correcting” for sampling in a 2- or 3-parameter model with Frac, yielding a final model superior to a univariate rule.
- For univariate models, 0-thresholded Avg performs best for WSR data (ACIntSum does

well, but really needs a quality “correction” to shine). 0-thresholded SumAvg works best with NEXRAD data. 0-thresholded Avg *appears* to do well for NEXRAD data, but only for very high POD, and is thus suspect. Over a broader performance range, SumAvg and ACIntSum fare much better (despite the latter appearing to do very badly at $POD=1$).

- The benefits of multiparameter linear models are nontrivial, especially for NEXRAD data. Since discriminant analysis is relatively simple to code, and since many of the inputs are already routinely available, serious consideration should be given to multivariate linear models in general.
- The benefits of multiparameter nonlinear models (neural networks) may be significant (pending proper testing against overfitting).
- Given the prevalence of multiparameter rule-based / decision tree models in *other* weather-related LCC, it may be worth investing some effort into coding “generalized” discriminant analysis or neural network functionality into whatever operational decision support system is in place at KSC.
- As Monte has pointed out: Regardless of model selection, these are all *really good* ROC curves. For the vast majority of cases, models can be constructed from radar observations alone which rarely predict significant E-fields when none are indeed present. (The issue of whether instantaneous, in-situ ice mass is *causative* for those fields is of course separate from whether or not it can be a good *predictor*, implicitly capturing the time history of the anvil’s development and decay).

APPENDIX: SOME OBSERVATIONS ON RULE-BASED MODELS (OR, “WHILE I HAVE YOUR EAR, I’LL GO OUT ON A LIMB...”)

Almost all of the weather-related LCC currently live in “Lego world” in the sand-and-rocks allegory; i.e., multivariate, logical threshold rule-based decision trees. Recall that this is simply one way to discretely parameterize a smooth/continuous decision surface, which happens to project well onto a FORTRAN-ish operational implementation (either manual or algorithmic). As a useful exercise, run through the weather LCC with a yellow highlighter

and mark off all quantities which count as “inputs” (these could be spatial variates, temporal variates, quality variates, or quantitative direct observations). It is immediately obvious that this is a “Lego world” in a room with far more than 3 dimensions, and we should have very little confidence that the Lego boundary is anything like an *optimal* parameterization of the true decision surface in the various hypercubes. Finally, recall that this is a high-impact issue; we’re balancing two of the most important drivers possible, safety and cost, so “optimization” is not just a technical nicety here. As a final note, I seem to recall that significant energy in the past has gone into decisions about altering the decision *thresholds* (e.g., sliding a wall in Lego world by a little bit), while very little attention has gone into the much more basic issue of how the decision *surface* itself has been parameterized (‘Lego world is the only world we shall consider’). This expenditure of energy strikes me as ... disproportionate ... within the context of overall “optimization”.

Having said that, entrenched reasons exist for rule-based models. Here are a few:

- “*It’s what we’ve always done.*” Fair enough, leveraging prior investment is important. “If it ain’t broke, don’t fix it.” The problem here is that we don’t *know* if it’s broke; we haven’t explored models *other* than Lego world. As has been pointed out, it’s largely “what we’ve always done” because up til now, we haven’t actually *had* training datasets to work with, only expert opinion (for which rule-based models are perfectly legitimate tools). This is thus both a data collection and experiment design issue.
- “*It’s easy.*” Time and money constraints are legitimate. However, I again note that this is a very high impact problem. Also, in a Lego world with many, many dimensions, it is most definitely *not* easy. It’s easy to read, is all (and arguably not even that). This is a human resources issue; a bigger toolbox increases the likelihood of better solutions.
- “*It’s procedural.*” Follow a checklist. This argument seems to have legs, and not knowing how actual launch operations proceed, may be quite valid. On the other hand, I don’t see how “Thou shalt not launch if X and Y (in the case

of Z) or A or B (in the case of C) is any more or less procedural than “Thou shalt not launch if the hazard probability is greater than 10% and the likelihood of error is less than 5%.” A solution can be procedural without exposing the inner details of the model to those in charge of implementation. This is a cultural, technical and cognitive issue.

- “*It’s safe.*” (The biggest box in Lego world yields the safest conclusions.) This is simply an end-run around the optimization problem, in which technical experts decree that cost is *never* an issue. If the multivariate inputs can support equivalent safety (POD) with lower cost (FAR) using a model other than Lego world, we have failed in the optimization task. This is a management (or failure to manage) issue.
- “*It’s easy to fix.*” This one gets used a lot ... we “understand” how rule-based models work. If something goes wrong, move a wall of the box. The problem is that this is just a round-about way of saying that the model was the wrong design choice from the start. If we had made the “best” fit to available data, nothing could “go wrong” *except* insufficient data *or* the management decision itself in balancing detection vs acceptable false alarms, in selection of a decision threshold. And again, in a Lego world with many dimensions, I’ll further argue that we “understand” the rule based models far less than we think we do.

It should be pretty obvious that I’m not a huge fan of decision tree models, especially in cases where we already *have* available a significant number of input variates and good-sized training datasets to work with. If or when additional LCC components come up for review, I’ll suggest that it may be constructive to keep these issues in mind.