Gus Sinnis
10/25/2000
P-23/LANL
Milagro Memo

A Maximum Likelihood Approach to Background Rejection

Introduction: In this memo I discuss a generalization of the "direct integration" method of background estimation. This new method allows for a source search analysis that is based purely on the shapes of the ON and OFF source X_2 distributions (# PMTs >2 PE / MAXPE), and on the expected shape of a signal. The approach uses a binned maximum likelihood estimator and is easily generalized to take account for any D.C. excess. This memo is more of a progress report on the application of the method. Several issues have arisen during this investigation and further work on understanding our detector is needed before this method can be put to optimal use. First I will review the salient points from my previous memo on gamma/hadron separation and show some results on its application to the Crab. I will then explain the generalization of the direct integration method and derive the maximum likelihood approach to gamma/hadron separation. Along the way I will point out the problems that we must still resolve, which relate to the PE calibrations and the Monte Carlo simulation of Milagro. An appendix delves a little deeper into calibration issues directly relevant to gamma/hadron separation.

Gamma Hadron Separation:

The parameter used to distinguish gamma showers from hadronic showers is denoted as X_2, where:

$$X_2 = \frac{N_\mu(\geq 2PE)}{\max_i(PE_i)},$$

and the maximum is found over the PMTs in the muon layer. For a complete discussion of this parameter see my previous memo. In that memo I showed that if one wishes to simply make a cut on the data, the Monte Carlo predicts that an optimal cut would reject all showers with X_2 less than 2.5. However, if one uses the data to determine the proton distributions (and Monte Carlo gammas), the optimal cut is at very large values of X_2. This is due to an incorrect prediction of the shape of the X_2 distribution by the Monte Carlo. In particular the data has more events with X_2 values larger than 2.5 than the Monte Carlo predicts. For completeness I show these distributions in Figure 1. It turns out that both the numerator and the denominator are incorrectly predicted by the Monte Carlo. I believe that the error in the numerator is due to the values used for the absorption and scattering lengths of light in the pond. The simulations used here are from Version 22 of the Monte Carlo, which has the Milagrito water. The Version 23 Monte Carlo has insufficient numbers of triggered events to be used for this analysis (~300 as of 10/11/00). The PEMAX distribution (denominator) is incorrect because of the water and also because of the TOT-to-PE calibrations. This is discussed in detail in the next section.

Andy has applied the above parameter (cut on 2.5) on the data from the Crab and observed a Q factor of 1.7 (resulting in a significance of 3.5σ), as predicted in my previous memo. I have applied the new generalized analysis method to determine the significance as a function of the X_2 cut value. The data was processed with a NFIT > 20 and >80 cut applied. The results are shown in Figure 2. Note that I obtain a smaller

significance than Andy at the value of X_2=2.5, (3.1σ for NFIT>20 and 3.3σ for NFIT>80) but the same value for Q (1.7). The maximum value I obtain for the significance is 3.3σ a cut on X_2 at 3.2 (for NFIT>20) and 3.5σ at X_2=2.6 (for NFIT>80).

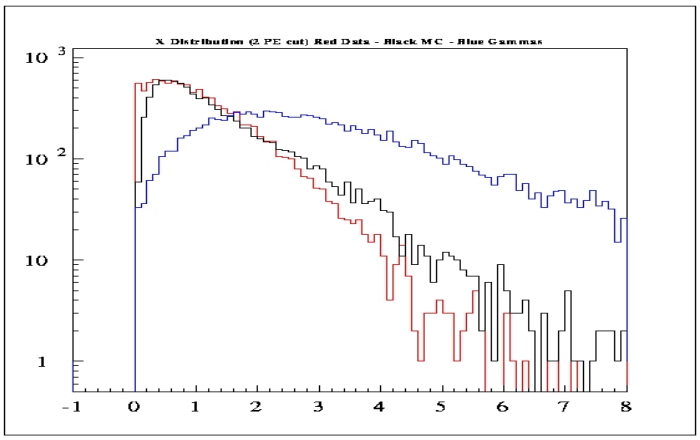

Figure 1 X_2 distributions for data (red) Monte Carlo protons (black) and Monte Carlo gamma showers (blue).

Note that the X_2 distributions do depend upon the value of NFIT and the optimization in my previous memo was for NFIT>20. (With the current uncertainties in the Monte Carlo there is little point in re-optimizing the value of the X_2 cut for each NFIT value.) The difference between this analysis and Andy's has been traced to a slight difference in the definition of the bin centered on the Crab and to a slightly different data set (I did not use sub-runs if the event rate changed by more than 7% from the previous sub-run).

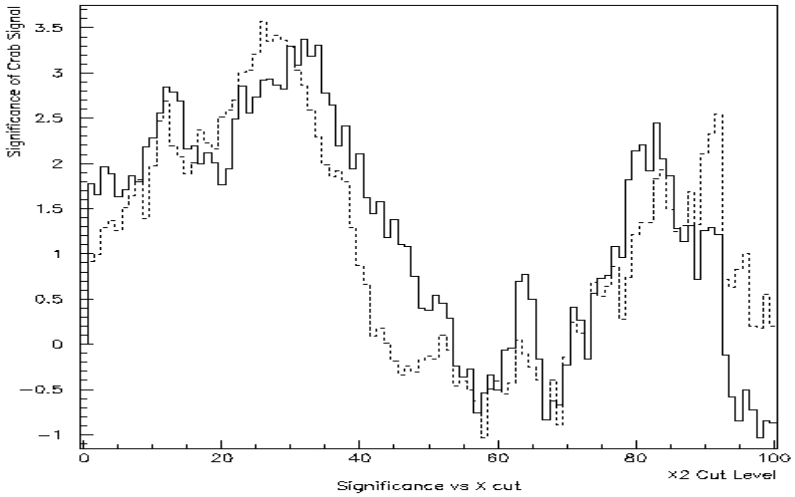

Figure 2 Significance of Signal from Crab vs. value of cut level in X_2 parameter. Solid line is for NFIT>20 and dashed line for NFIT>80.

Calibrations and PEMAX:

Figure 3 shows the PEMAX distribution predicted by Version 22 of the Monte Carlo (protons). Note the 2-peaked structure. Previous examination of the data failed to observe this 2-peaked structure. The presence of this structure in the data is strongly dependent upon the calibrations used to process the data. Since my previous memo there has been an improvement in the calibrations that allows us to see this structure. In Figure 4 I show the PEMAX distribution using the old "spectrum"-based TOT-to-PE conversions. One can see a shoulder on the distributions but the first peak is missing. Superimposed on this plot is the PEMAX distribution using the occupancy method in its first incarnation. In this calibration if the HI TOT was not within it's useful range the value of the LO-TOT was extrapolated to arbitrarily high PE values. In the latest version of TOTPE_OCC4, the PE value from the LO-TOT is truncated at the minimum value that HI-TOT can give. While this approach gives strange looking PE distributions for each PMT (see Figure 5), it makes a marked improvement in the PEMAX distribution, Figure 6. A way of improving upon this approach is to calibrate the HI-TOT to lower values; currently HI-TOT starts to be used between 200 and 300 counts, though above 100 counts it contains reliable data.

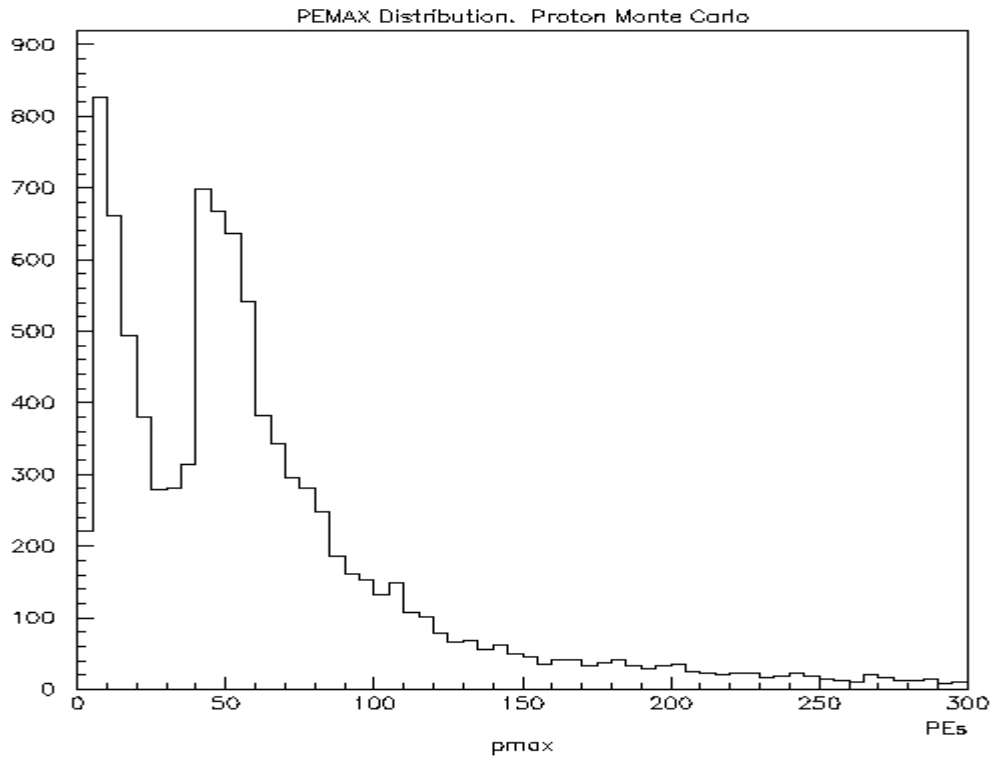

Figure 3 PEMAX distribution Monte Carlo (Version 22) protons.

First some comments on the PEMAX distribution. The second peak is due to local light from muons in the bottom layer. The first peak is due to showers without muons and thus light that travels from the top layer to the bottom layer. Thus, the position of the second peak is sensitive to the quantum efficiency of the PMTs and the position of the first peak is sensitive to both the quantum efficiency of the PMTs and the

transparency of the water. Clearer water should yield larger PEs in the bottom layer. In fact, Version 22 of the Monte Carlo used Milagrito water (10 meter scattering and 10 meter absorption). This is consistent with the position of the first peak in the data.

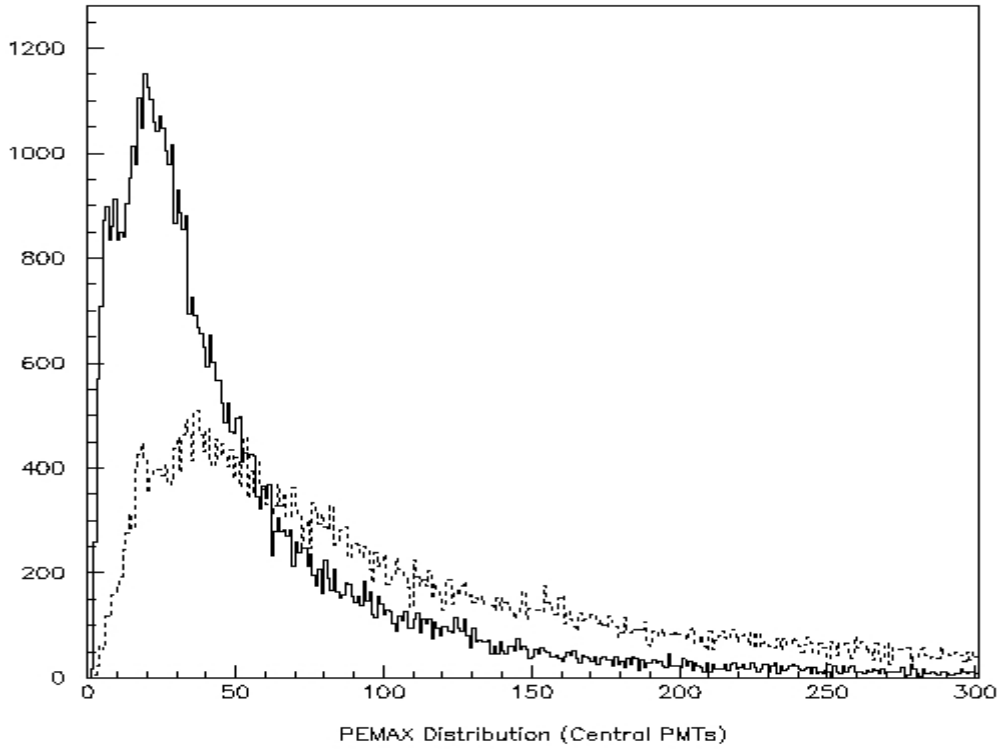

Figure 4 PEMAX distribution. Solid line uses "spectrum"-based TOT-to-PE calibration and the dashed line uses the occupancy TOT-to-PE calibration (without the truncation of LO-TOT).

I have tried to improve upon these results by extrapolating HI-TOT to *smaller* values. I used a linear extrapolation provided by Roman and Lazar. The problem is knowing where a linear extrapolation breaks down. By playing around with the start of HI-TOT one can play around with the relative amplitudes of the two peaks, and by linearly extrapolating HI-TOT down to 100 counts, one can make the first peak disappear completely. Without being able to trust the Monte Carlo on a quantitative basis one cannot use this distribution to determine which is the best PE calibration. Further work is needed in understanding the PE calibration in the region near high threshold.

Extension of the Direct Integration Method:

This technique estimates the background level at a given point in the celestial sky, by integrating the observed rate over the shape of our "efficiency" in local coordinates. The detector efficiency map is determined from the data and is simply the probability that a given event came from any point in the sky. The efficiency map is made in the local coordinates of hour angle (HA) and declination (DEC), with the same binning as the celestial maps of observed (OBS) and expected (EXP) events in right ascension (RA) and DEC. By keeping track of how many events arrived during each sidereal interval

(NSIDER) one can both estimate the background and determine the excess with a single pass through the data. To eliminate systematic errors due to bin misalignments the array

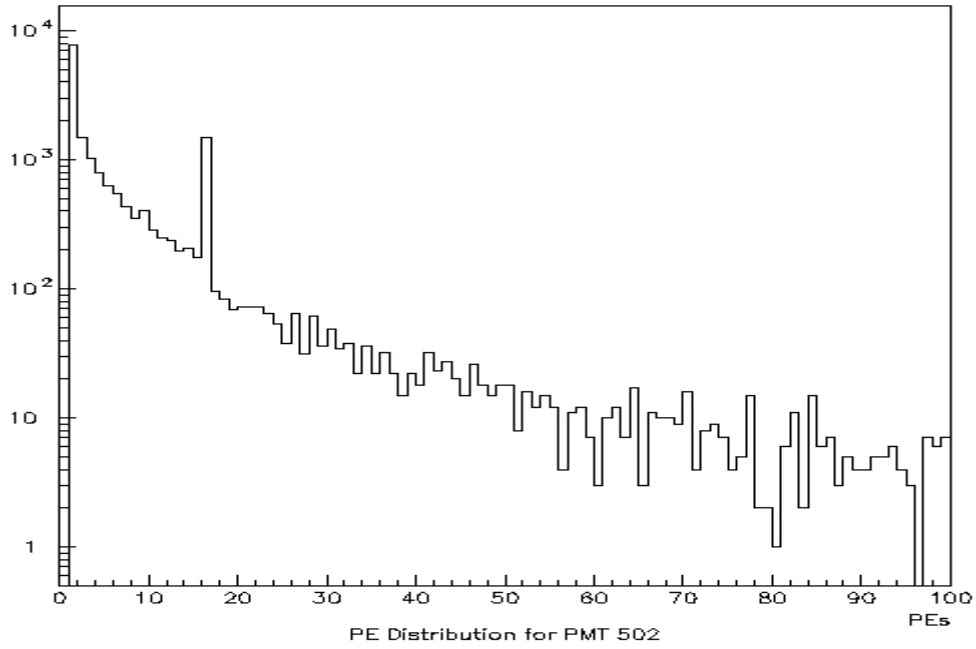

Figure 5 PE distribution from a single PMT using the occupancy TOT-to-PE calibration, with a truncation set on the PEs from LO-TOT.

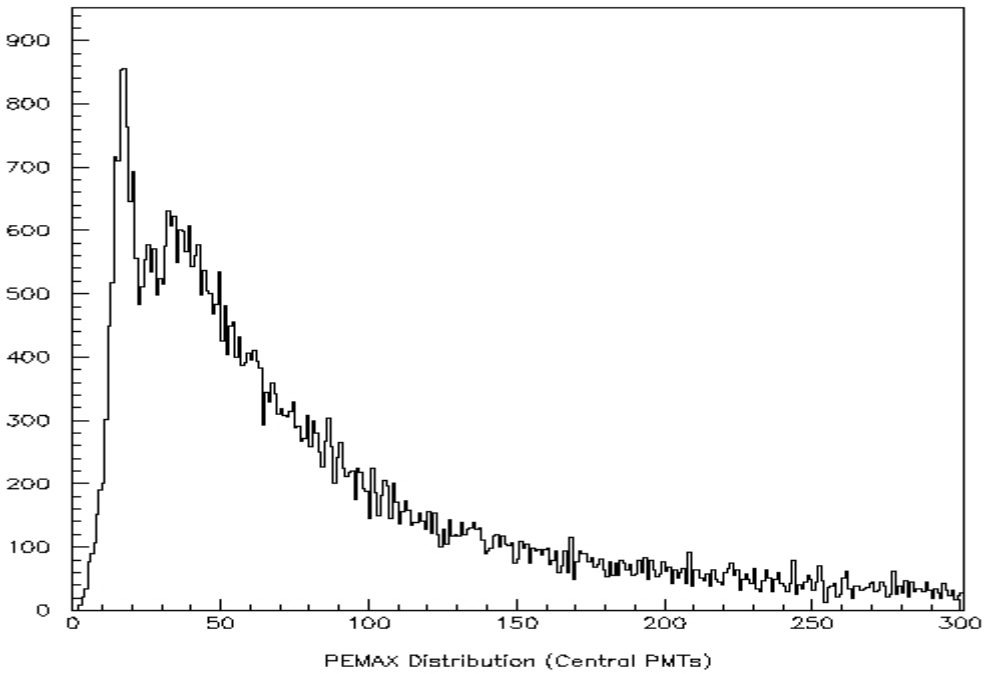

Figure 6 PEMAX distribution using occupancy TOT-to--PE calibration with a trunctation of the LO-TOT.

NSIDER is binned with the same width as the EFF, OBS, and EXP maps (0.1 degrees in HA or RA corresponds to 24 sidereal seconds). (See Appendix B for a coding example.) The extension performs the following integration:

$$EXP(RA, DEC, X) = \iint_{t,\Omega} R(t)\mathrm{E}(HA, DEC, X)\delta((HA - t) - RA)d\Omega dt$$

While this example I will give has to do with a parameter that is used to differentiate gamma-ray showers from proton showers, it may be used for any parameter (i.e. NFIT). The basis of the method is to expand the dimension of the EFF, OBS, and EXP arrays to include any/all parameters of interest. For example, EFF[HA][DEC] would become EFF[HA][DEC][X]. Note that the same assumption applies to this third parameter as to the other two: The *shape* of the detector response function is independent of rate for each 2-hour interval. We assume that the shape of the X distribution may depend upon local coordinates HA, DEC, but that for the 2-hour interval over which we accumulate background data the shape of the X distribution is constant for each HA, DEC pair.

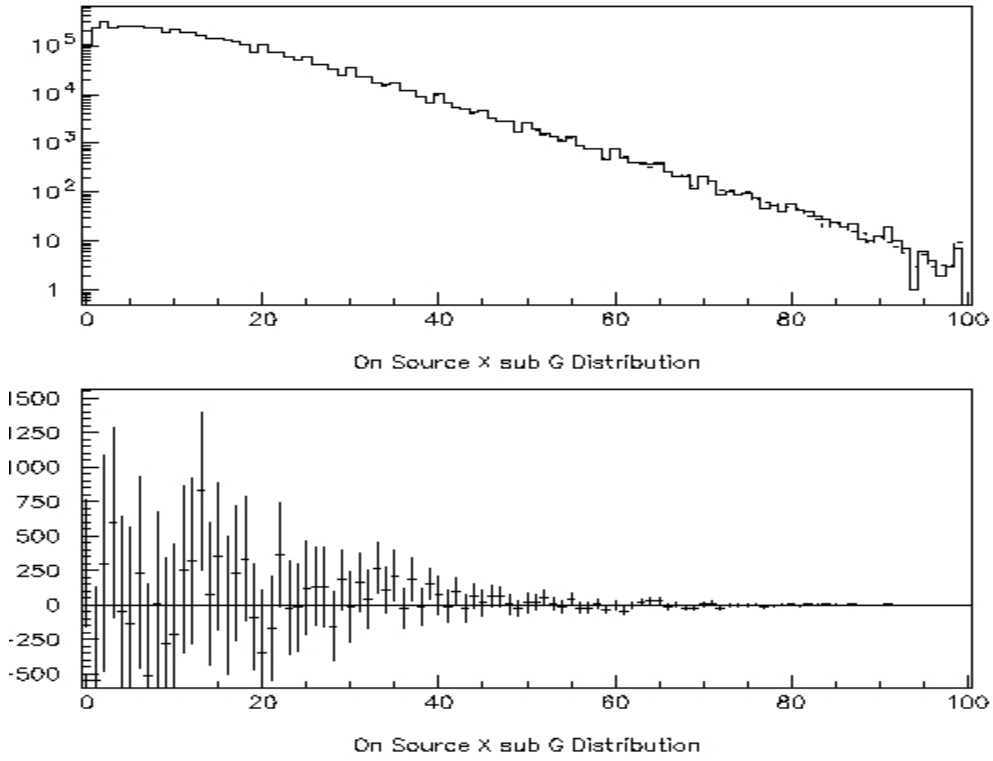

Figure 7 ON (solid) and OFF (dashed) source X distributions in Crab bin. The values given on the X-axis are bin number (X times 10). NFIT>20 2.1 degree wide square bin. Bottom panel is the difference ON - OFF.

Maximum Likelihood Method of Source Analysis:

Using,

$$X = \frac{N_{muon}(> 2PE)}{PEMAX(muon)}$$

with 100 bins in X spanning [0-10] (overflows going into the last bin) I analyzed the Crab data re-reconstructed at Maryland (runs 1121-2451). I then summed the arrays EXP[RA][DEC][X] and OBS[RA][DEC][X] over RA and DEC (around the Crab: 2.1 degrees bin width for NFIT>20 and 1.7 degree bin width for NFIT>80), to obtain an EXP[X] and OBS[X] distribution. These are shown in Figure 7 (NFIT>20) and Figure 8 (NFIT>80) superimposed upon each other with the difference (OBS-EXP) shown in the lower panels of each figure.

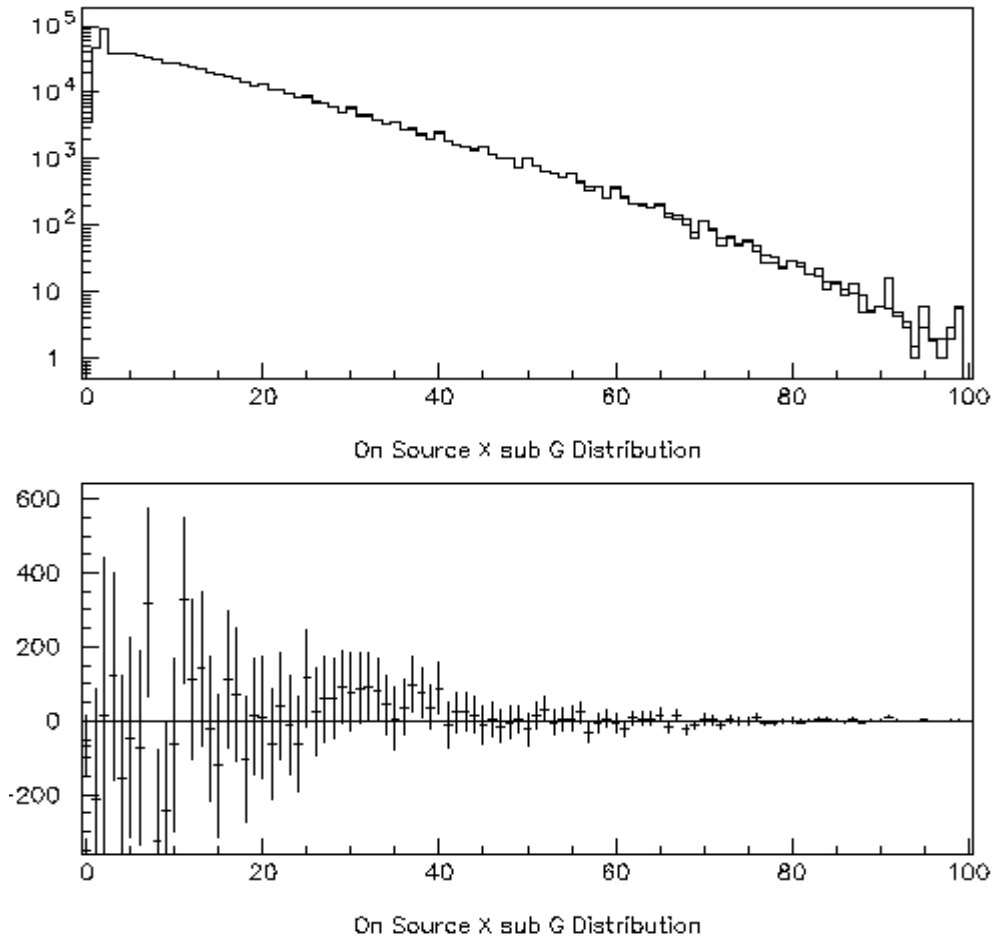

Figure 8 Same as Figure 7 with NFIT>80 and 1.7 degree wide bin.

In Figure 9 I show the expected X distribution for gamma showers with NFIT>20 (solid line) and NFIT>80 (dashed line). The problem is how to use the ON source, OFF source, and Monte Carlo gamma distributions to search for an excess of "gamma-like" events from the Crab. I derive a method that uses only the shapes of the distributions, with no account taken for the overall D.C. excess observed. This technique relies solely on the gamma-like nature of the images from the direction of the Crab in Milagro.

I use a binned likelihood ratio approach, where the goal is to determine the likelihood that the ON source distribution is made up of a contribution from the OFF source distribution plus a contribution from the gamma distribution (as determined by the

Monte Carlo). This is a binned likelihood approach because the product is over the 100 bins in the X distributions. The likelihood function is defined as:

$$L(N_S, N_{OBS} - N_S) = \prod_{i=1}^{100} P(ON_i \,|\, N_S \Gamma_i + (N_{OBS} - N_S) \times OFF_i)$$

Where the ON_i is the number of ON source events in the i^{th} bin of the X distribution, Γ_i is the fraction of Monte Carlo gamma-ray events in the i^{th} bin of the X distribution, and OFF_i is the fraction of OFF source events in the i^{th} bin of the X distribution. $P(\mathbf{a} \mid \mathbf{b})$ is simply the Poisson probability of observing **a** events when one expects **b** events. I have applied the constraint $N_B + N_S = N_{OBS}$. One then maximizes the above quantity over N_S and finds the ratio:

$$\lambda = \frac{L(0, N_{OBS})}{\max(N_S)[L(N_S, N_{OBS} - N_S)]}.$$

Then, $-2\ln(\lambda)$ is distributed as a χ^2(1 dof), so the square root of $[-2\ln(\lambda)]$ is the significance of the likelihood ratio. In Figure 10 I show $-2\ln(\lambda)$ as a function of N_S for the NFIT>20 and NFIT>80 data sets. In neither of these cases is the significance of the likelihood ratio very impressive (though it is better than the observed D.C. excess for each case). A better result is obtained by making a hard cut on the data.

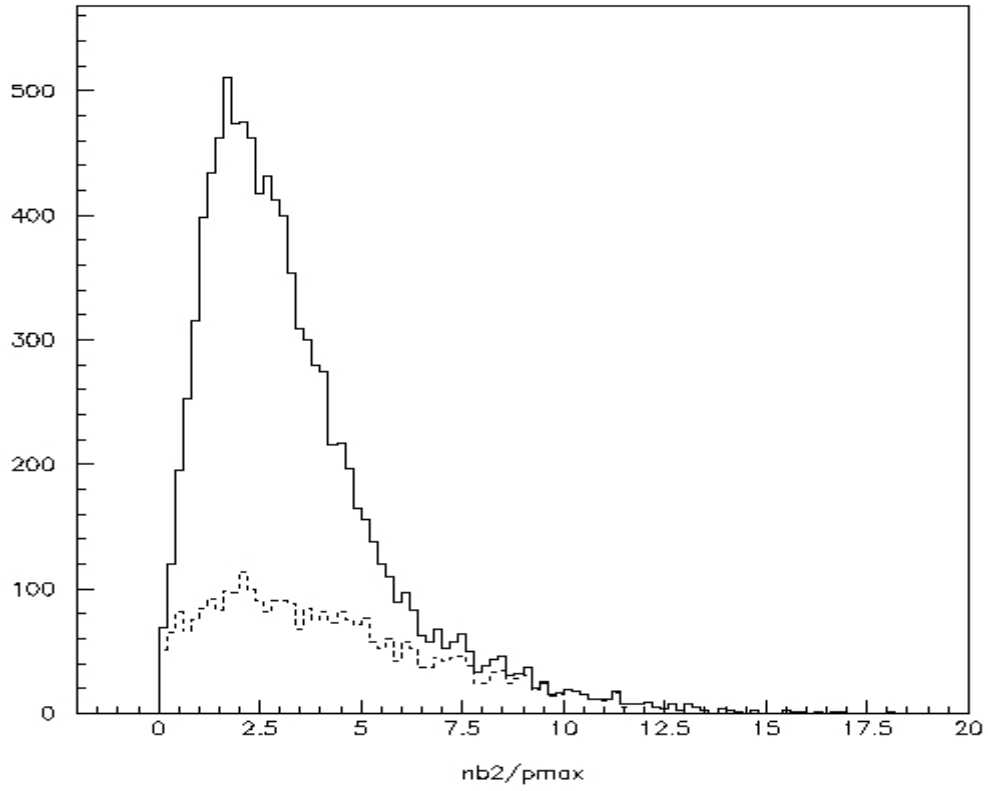

Figure 9 X distribution from Monte Carlo gamma rays. Version 22 of Monte Carlo with 50 PMT trigger requirement. Solid line is for NFIT>20 and dashed line for NFIT>80.

The table below shows the results of the analyses. It is interesting that though the likelihood approach maximizes at small number of excess events, the significance of

these events is much greater than the simply Gaussian excess. This would seem to indicate that not only is the shape important, but that the simulation does a reasonable job of predicting the shape for the X distribution for gamma ray showers.

Table 1 Result of Crab analysis.

Analysis	NExp	NObs	D.C. Excess	Sigma	Likelihood: NSignal	Likelihood: Significance
NFIT>20	4,525,489	4,529,268	3378	1.8σ	648	2.4σ
NFIT>80	730,560	731,345	784	0.9σ	596	2.25σ

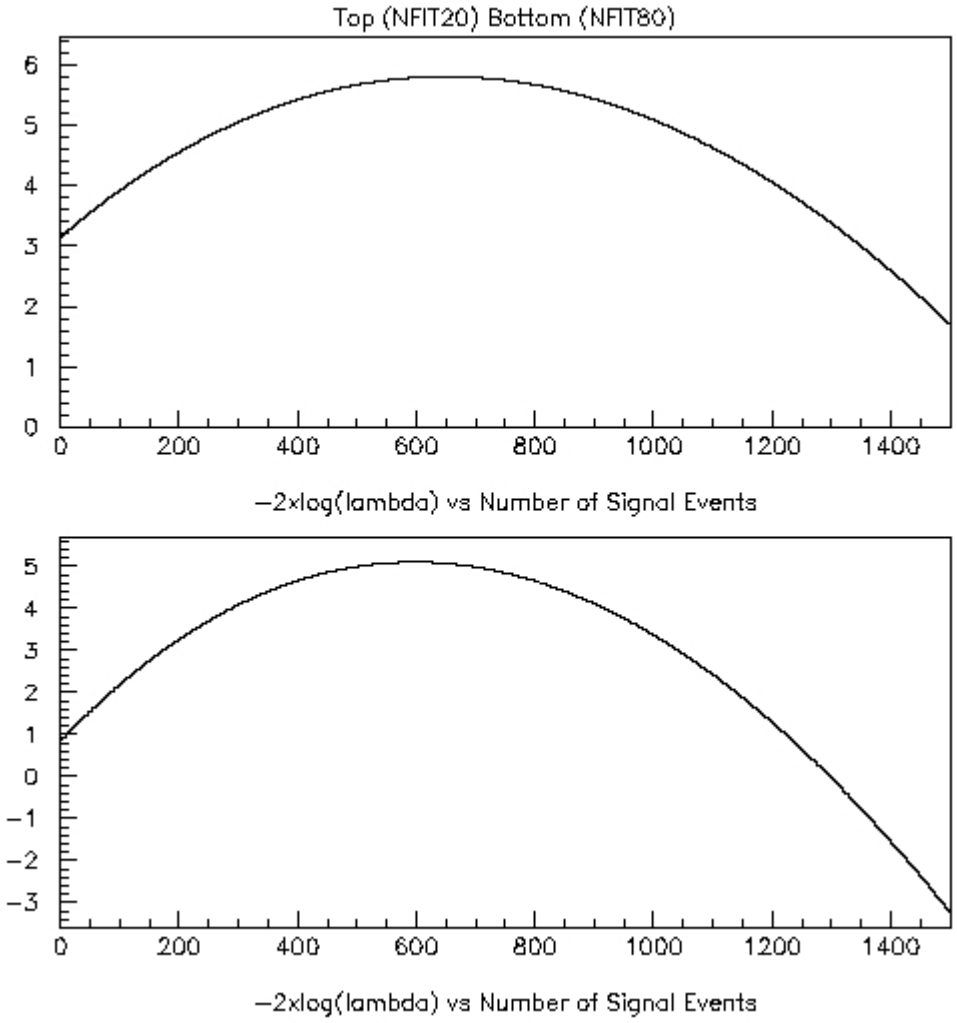

Figure 10 -2ln(λ) as a function of the number of signal events. Top plot is for NFIT>20 and bottom for NFIT>80. The significance is the square root of the y-axis.

What Went Wrong?: Examination of Figure 9 shows that the Monte Carlo predicts that there should be significant overflow in the X distribution (i.e. many gamma ray events with X>10.). In fact examination of Figure 8 shows that this is not the case. Secondly,

the number of signal events that maximized the likelihood function is smaller than the observed excess, especially for the NFIT>20 analysis. Recall that for proton showers the Monte Carlo over predicted, by nearly an order of magnitude, the number of events at large X values. I will hypothesize that the Monte Carlo has a similar problem in its prediction of gamma ray events at large X values. To investigate this hypothesis, I set the number of Monte Carlo events in the i^{th} bin of the X distribution to zero for all *i*'s greater than some value X_{max}, and investigate the behavior of the likelihood function as a function of X_{max}.

Figure 11 shows the significance of the results for all values of X_{max} between 3 and 10 (a value of X_{max}=10 corresponds to no truncation of the X distribution). Figure 12 shows the number of signal-like events (that maximizes the likelihood function) as a function of X_{max}.

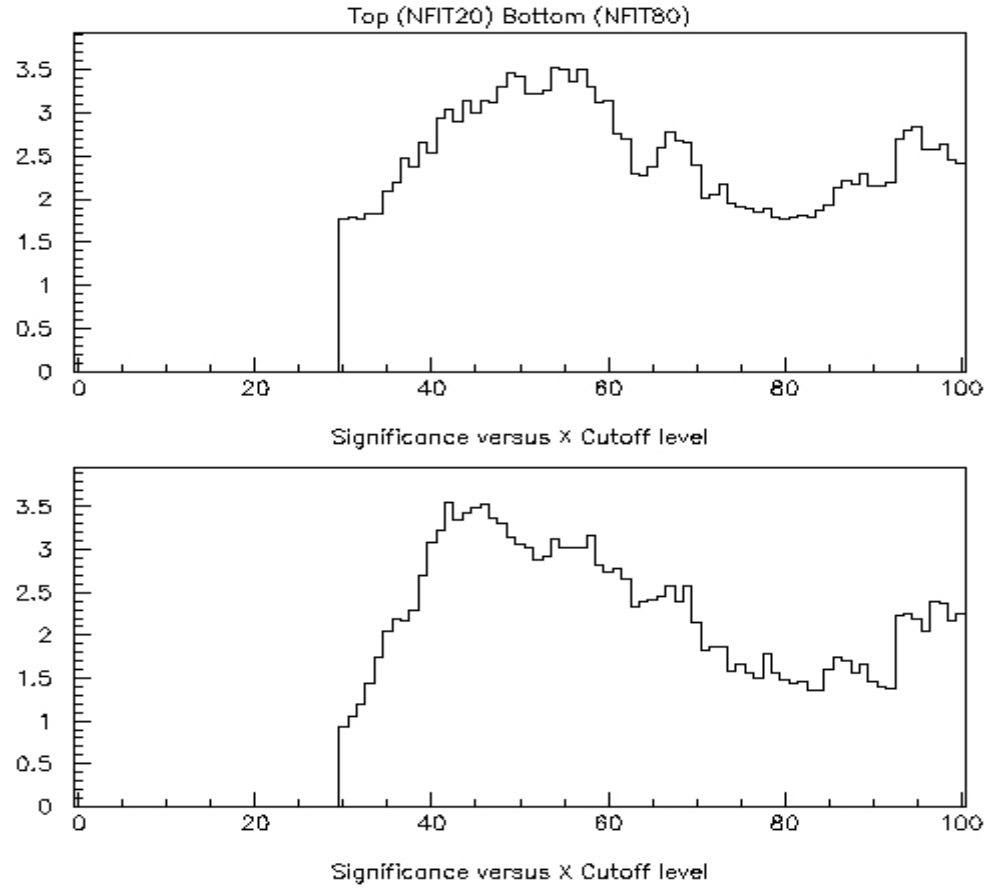

Figure 11 Significance of likelihood ratio as a function of X_{max}. Top plot is for NFIT>20 and bottom plot for NFIT>80.

Both analyses (NFIT>20 and NFIT>80) maximize at roughly 3.5σ and have broad areas of X_{max} for which they are above 3σ. In both cases the number of signal events deduced is reasonable. While the above significances have been tuned, we have reason to suspect that the Monte Carlo is incorrect in precisely the manner investigated. The above discussion is meant to provide a feeling for how good we could do in principle if the Monte Carlo matched the data. Clearly we must work on improving the simulations to

match the observed proton distributions. Hopefully this will also correct the gamma distributions.

Conclusions: I have developed an extension to the "direct integration" method of background estimation that allows for a binned likelihood analysis of the Milagro data. I have applied this technique to the Crab data using a gamma/hadron discrimination parameter previously discussed. The likelihood method gives results that are better than a straightforward D.C. excess analysis, but not as good as they could be. I believe that an improvement in the Monte Carlo simulation of both proton and gamma showers should enable us to detect the Crab nebula solely from the shape of the X distributions.

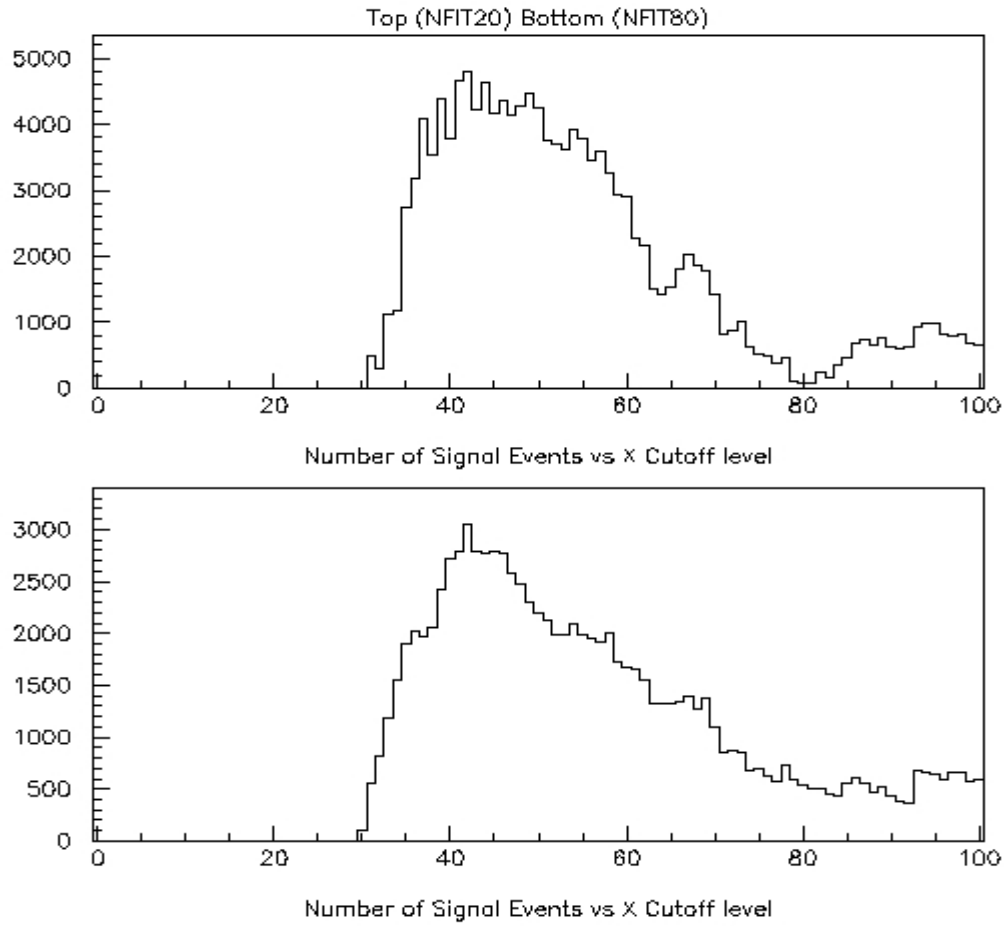

Figure 12 Number of signal events (that maximizes the likelihood function) as a function of X_{max}. Top plot is for NFIT>20 and bottom plot for NFIT>80.

Appendix A: Discussion of the quantum efficiency of the PMTs:

The current TOT-to-PE calibration method tells us how many PEs where detected at each PMT. For gamma/hadron separation we would like to know how many incident photons struck each PMT. The difference between the two numbers is the relative efficiency of the PMTs (quantum efficiency, collection efficiency, and baffle efficiency). In principle one could use through-going muons to give an absolute calibration for each PMT. Here

I demonstrate a straightforward method using the measured average PE level in each PMT to adjust for the relative differences in the PMT efficiencies.

To demonstrate why this is a problem I examine the uniformity of the frequency distribution of the PMTs selected to have the maximum pulse height. (i.e. for each PMT the of times that PMT was selected as having the highest pulse height.) The Monte Carlo predicts that there are three classes of PMTs: center PMTs, edge PMTs, and corner PMTs. The corner PMTs are selected the most frequently, than the edge PMTs, and finally the central PMTs. Within each class the frequency distribution should be "relatively" uniform (see below). In Figure 13 I show the frequency distribution for the central PMTs. It is highly non-uniform. The χ^2 to a straight line is 3109 for 193 degrees of freedom. The lower panel shows the same distribution after the corrections for relative efficiency (described below) have been applied to the data.

The relative efficiency of the PMTs can be determined by measuring the average number of PEs detected by each PMT. Assuming that all PMTs in the same class should have identical PE distributions (and therefore identical <PE>) and if $<PE_{Class}>$ is the average PE value for all events and all PMTs in the class, then:

$$Q_i = \left\langle PE_{Class} \right\rangle \Big/ \left\langle PE_i \right\rangle \quad \text{and} \quad PE_i^{true} = Q_i PE_i \,.$$

Where the latter formula is the correction applied to the current TOT-to-PE calibration on a PMT-by-PMT basis. Figure 14 is a histogram of the Q_i for all PMTs in the muon layer. Figure 15 shows the PEMAX frequency distribution after applying the correction for the relative PMT efficiencies. For comparison the same distribution derived from Monte Carlo protons is shown in the lower panel of the Figure. The χ^2 is now 1070 for 193 dofs. While this is a marked improvement, it is still not as uniform as the Monte Carlo predicts it should be. The remaining discrepancy is due to 17 PMTs that lie far from the mean. After removal of these outliers there is good agreement with the Monte Carlo.

Table 2 compares the uniformity of the PEMAX frequency distribution in the data with the Monte Carlo. After the efficiency correction has been applied to the data and the 17 outliers removed, the uniformity is consistent with the Monte Carlo predictions. The outliers may be PMTs that are poorly extrapolated to very high pulse heights (beyond 100 PE). Up to 100 PE there appears to be no difference in the PE distributions of these outlier PMTs and the remaining PMTs. The remaining observed (and predicted) non-uniformity is most likely due to the same effect that causes the outer PMTs to be selected as "hottest" the most frequently, most of our shower cores lie outside of the pond. I believe that the difference between the data and Monte Carlo in regards to the relative frequency of selecting a central versus edge PMT has to do with the arrangement of dead PMTs in Milagro. The Monte Carlo has no dead PMTs.

Table 2 Uniformity of PEMAX frequency distribution. There are three rows that describe the data: the first without the correction, the second after the correction, and the third after the correction and after removal of 17 "outlier" PMTs. An outlier is defined as any PMT that is more that 4 standard deviations from the mean.

Data Source	χ^2 to Flat	Central PMT Freq.	Edge PMT Freq.	Corner PMT Freq.
Data uncorrected	3109 (16/dof)	0.0036	0.0050	0.0071
Data corrected	1070 (5.5/dof)	0.0036	0.0050	0.0071
Data corrected outliers removed	627 (3.5/dof)	n/a	n/a	n/a
Monte Carlo Protons	684 (3.3/dof)	0.0028	0.0063	0.0084

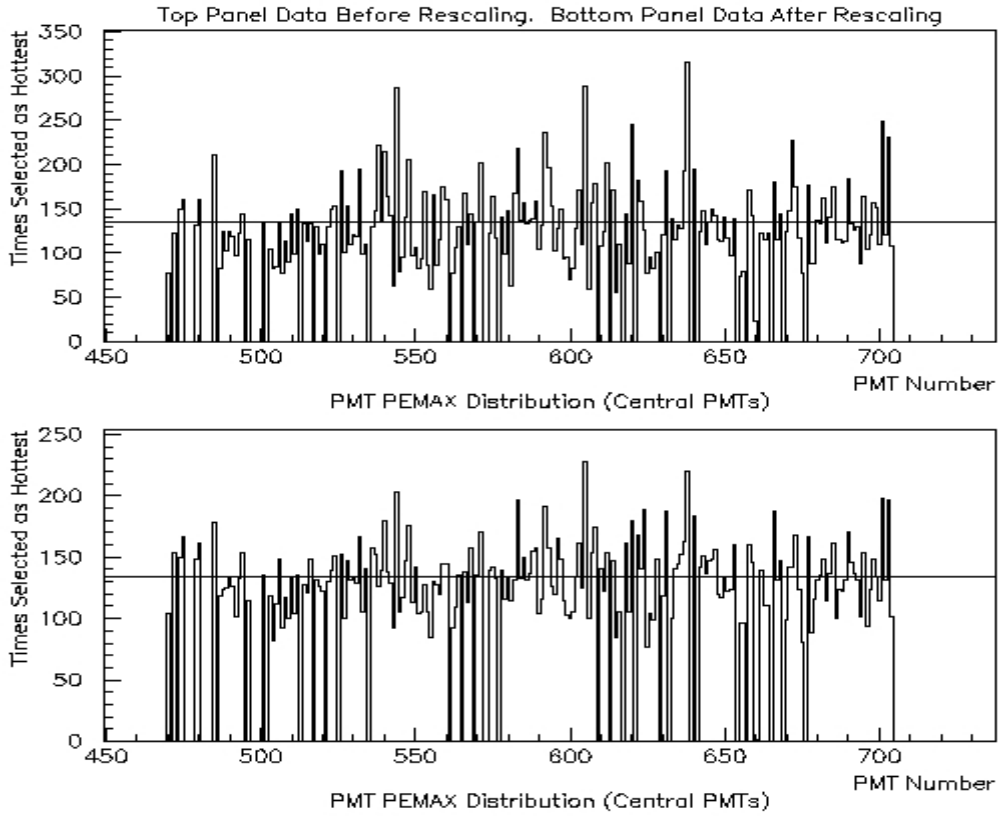

Figure 13 Frequency Distribution of PEMAX. Top plot is before rescaling of PMTs bottom plot after rescaling. Lines are drawn at the average.

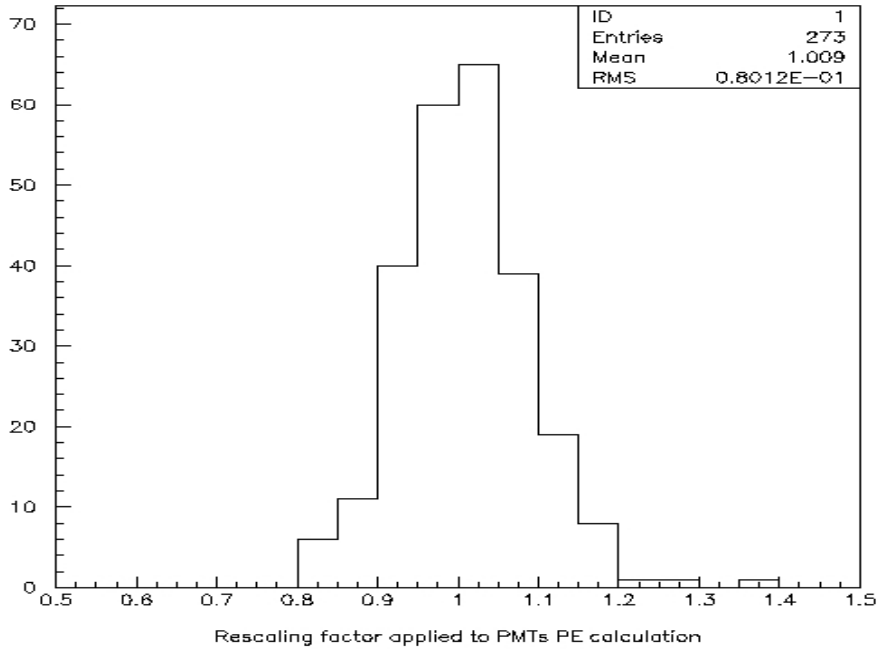

Figure 14 Rescaling factors applied to PMTs to correct for relative efficiencies.

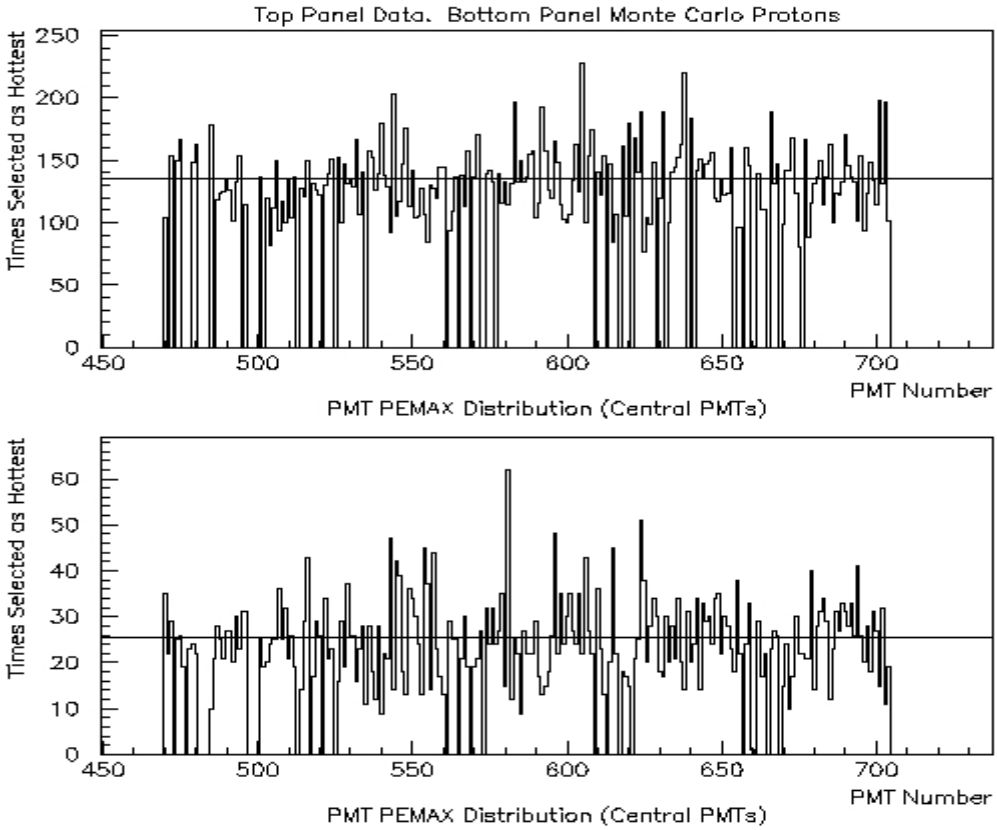

Figure 15 Frequency distribution of PEMAX. Top panel is data after rescaling and bottom panel is Monte Carlo protons. Lines are drawn at the average.

Appendix B: Direct Integration Code Fragment:

```
While MilagroEvent2( &eventData, &dataCon, &coords){
        upDateSKY(RA, DEC, OBS);
        sidereal_time = sider(julianDate, time);
        HA = sidereal_time - RA;
        upDateSKY(HA,DEC,EFF);
        is = IS(sidereal_time);
        NSIDER[is]++;
        If (time - tZero >= 2 HOURS){
                upDateBKG(NSIDER, EFF, EXP);
                tZero = time;
                bzero(EFF);
        }
}
void upDateSKY(float alpha, float dec, float array[][DECBINS]){
        ir  = IR(alpha);
        id = ID(dec);
        array[ir][id]++;
        return;
}
void upDateBKG(int NSIDER[], float EFF[][DECBINS], float EXP[][DECBINS]){
   for(is = 0; is < RABINS; is++){
        evts = NSIDER[is];
        for(ir=0; ir<RABINS; ir++){
                ih = (is - ir);
                for(id=0; id<DECBINS; id++){
                        EXP[ir][id]  += ((float) evts*EFF[ih][id]);
                }
        }
  }
  return;
}
```

In the code fragment IR, ID, and IS are macros that define/determine the binning used. Every 2-hours the EFF map is convoluted with the array NSIDER to determine the expected background (EXP) (this is the purpose of the upDateBKG function). The resulting EXP and OBS arrays may be examined over any time interval of interest to search for an excess.