γ-h Separation and MARS

Milagro Note
Curtis Lansdell
Oct. 7, 2004
Updated Oct. 26, 2004

This note presents some results from using Multivariate Adaptive Regression Splines (MARS) to determine good γ-h event separation variables. Details of how MARS is used in the Milagro software framework are also shown. Traditionally, X2 cuts have given Q-factors of roughly 1.6. With MARS, Q-factors of greater than 2 are obtainable.

What is MARS?

MARS[1] is a technique which predicts the values of an outcome variable given a set of independent predictor variables. In terms of γ-h separation in Milagro, MARS was implemented by Frank Samuelson to give an indicator of how γ-like an event appears. MARS requires γ showers and some kind of background, e.g., proton showers. Within the Milagro code framework, MARS takes various input parameters, such as x2, nb2, mxPE, etc., and returns

$$MARSvalue = \ln[P(\gamma) / P(h)] \tag{1}$$

In the simple case of protons only for background events, the probability of being a hadron event is

$$P(h) = P(p) = 1 - P(\gamma) \tag{2}$$

For any event, a value from Equation (1) less than 0 means the event is more proton-like, while a value greater than 0 is more γ-like.

Q-factors are calculated for particular ranges of MARS values from histograms created via Milinda and ROOT.

[1] Friedman, J., "Multivariate Adaptive Regression Splines", Annals of Statistics **19** (1991)

Milinda, ROOT, and MARS

The Milinda framework is used in the first step of generating MARS values. A ROOT TTree object is filled with variables calculated or read directly from data files. This is done for raw data and for Monte Carlo proton and gamma data. The TTrees are written to separate files, and then pruned down to the relevant parameters for use in MARS. These parameters include the variables of interest, code numbers identifying real data, proton, and gamma events, and a weight associated with each event. For the results presented here, the weighting was based only on the MC core position since simulated events were thrown flat in radius. This means that all real data events had a weight of 1.

Histograms of the MARS values are filled for the different data types. Q-factors for the different MARS regions are calculated by looping over all of the histogram bins and gradually filling Q-factor and efficiency histograms, as in this example:

```
Float_t pFitnorm = hPrMarsDistFit->Integral(0,101);
Float_t gFitnorm = hGammaMarsDistBin->Integral(0,101);
for (int i=0, i<101, i++) {
  Float_t pFint = hPrMarsDistFit->Integral(i,101)/pFitnorm;
  Float_t gFint = hGammaMarsDistBin->Integral(i,101)/gFitnorm;

  hPrFracFit->SetBinContent(i,pFint);
  hGammaFracFit->SetBinContent(i,gFint);

  if (pFint > 0.00001) {
    hQfacMCFit->SetBinContent(i,gFint/sqrt(pFint));   // Q-factor
  }
}
```

The preceding example made use of 100 bin histograms. By integrating from 0 to 101 in bin number, it explicitly considers underflow and overflow bins in its calculation. The MARS distributions including the Q-factors are then complete.

Results

While various combinations of variables have been examined using the method outlined above, presented here are a couple of combinations of current interest. Both combinations use the same event cuts on the same data samples: nTop>55, nFit>80, triggered events (VMEword≠0), and on MC γs, $\Delta(\theta,\phi)<1.2$. As a point of reference, the all-sky[2] paper used all triggered events, x2>2.5, and nFit>20. In this note, 206,172 real data, 51,991 MC proton, and 59,842 MC γ events compose the samples. MC proton events have energies from 50 GeV to 100 TeV, while MC γ events range from 100 GeV to 100 TeV. The MC events were thrown out to 1 km with a -2.7 and -2.4 energy spectrum slopes for protons and γs, respectively. The real data events come from the sub-run 5268_0384, taken on Dec. 13, 2003 at 22:05 UT.

Tony Shoup's all-layer fitting routines are used here. For both angle and core fitting, the air shower, muon, and outrigger layers are used.

First are the results from a 5 parameter fit. These particular parameters were chosen based on visual differences in proton and γ distributions.

MARS input: 5 parameters

The parameters are x2, mxPE, sumPEbot, nb2, and nb8.

x2 = nb2/mxPE
mxPE = maximum number of PEs in a muon layer tube
sumPEbot = sum of the PEs in the muon layer
nb2 = number of muon layer tubes with at least 2 PEs
nb8 = number of muon layer tubes with at least 8 PEs

[2] Atkins, R. *et al.*, ApJ **608**, 680 (2004)

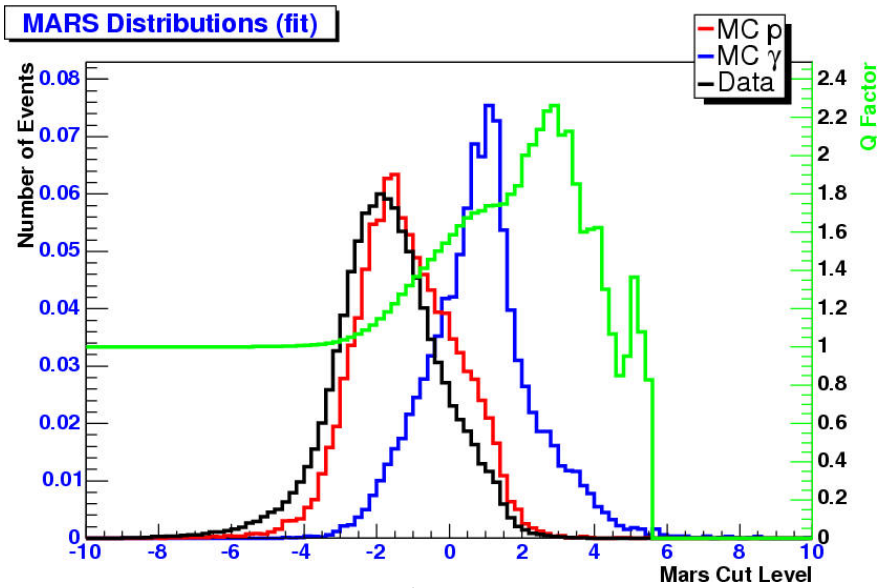

Figure 1

5 parameters, on+off pond. These are normalized MARS distributions, including the Q-factors, when protons (red) are used as the background with γs (blue) as the signal. Real data should primarily be composed of proton events, so the discrepancy between real data (black) and protons needs to be examined. These histograms are for all events (cores can be on or off the pond).

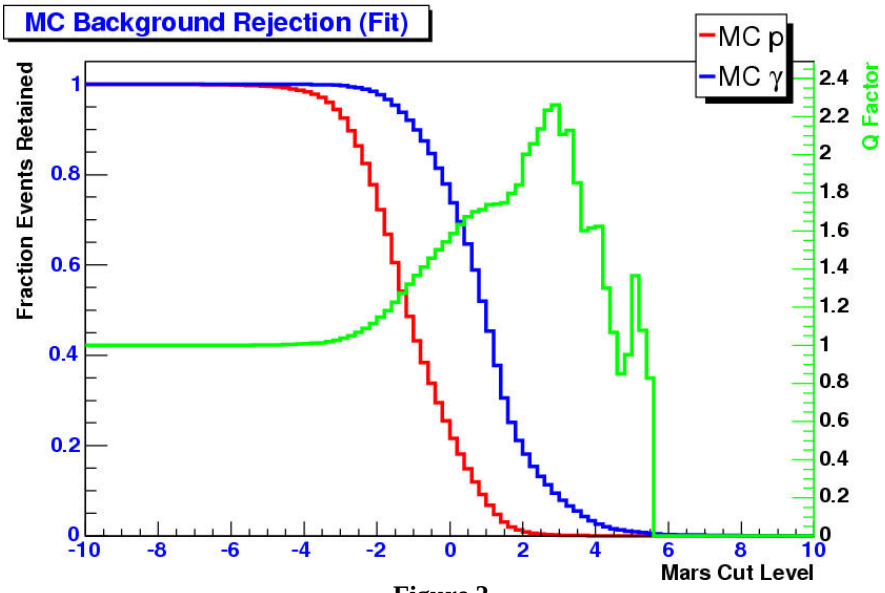

Figure 2

5 parameters, on+off pond. Background and signal efficiencies for various Q-factors.

Figure 1 shows normalized distributions for all events, where the event cores can be on or off the pond. Figure 2 shows the efficiencies for different Q-factors. The following two figures are for on-pond and off-pond events separately.

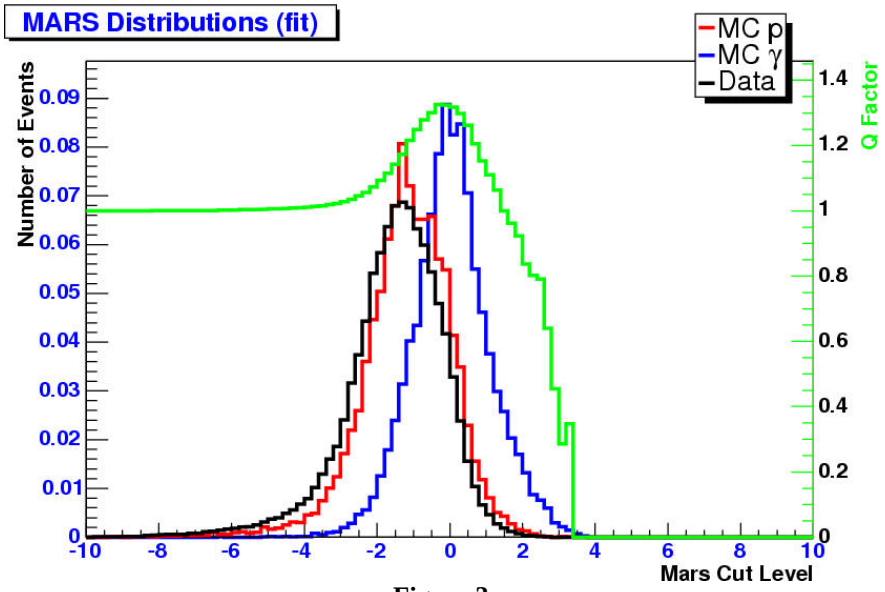

Figure 3

5 parameters, on-pond. These are MARS distributions for events with cores falling on the pond. There is better agreement between real data and MC protons, though the γ events appear to be a little too far over into the proton-like regime.

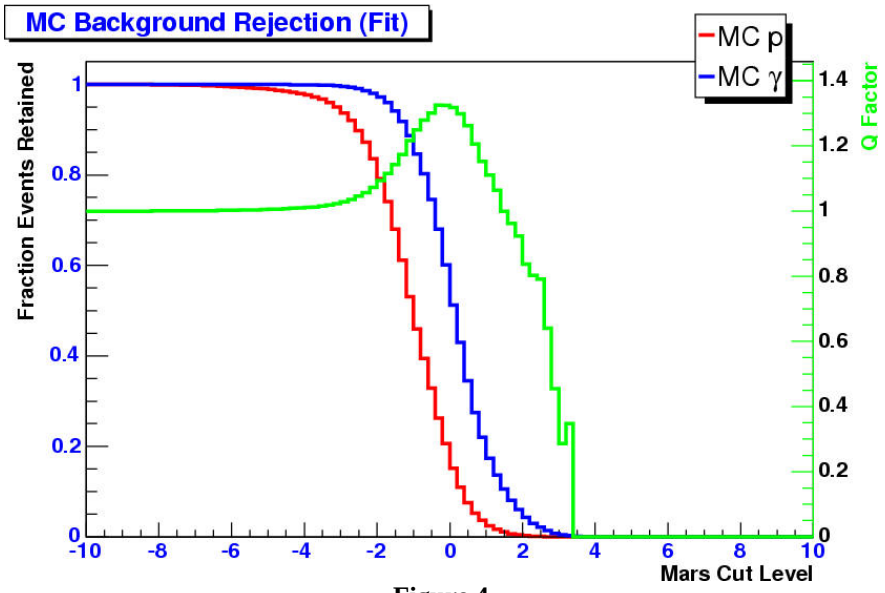

Figure 4

5 parameters, on-pond. Background and signal efficiencies for various Q-factors.

Figure 3 shows distributions for on-pond events while Figure 5 below is for off-pond events.

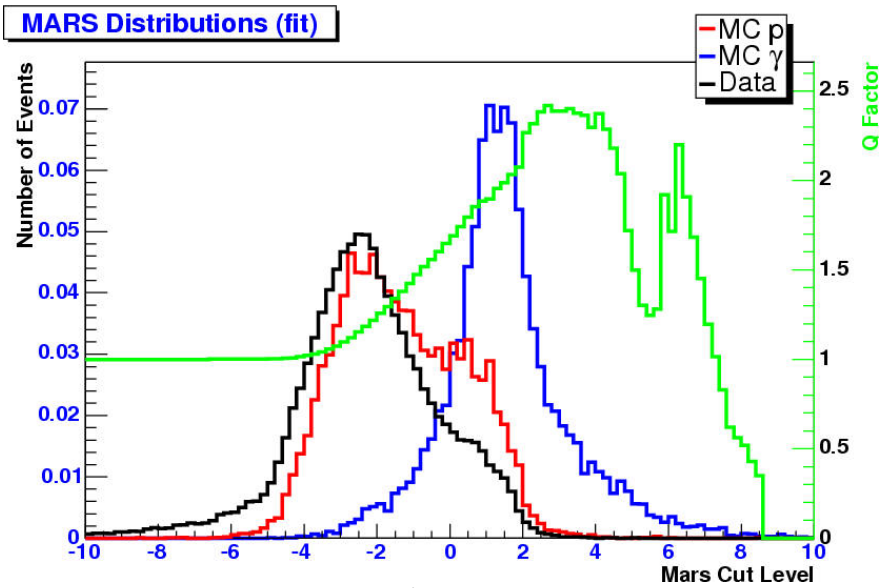

Figure 5

5 parameters, off-pond. These are MARS distributions for events with cores falling outside of the pond. The MC protons do not agree well with real data.

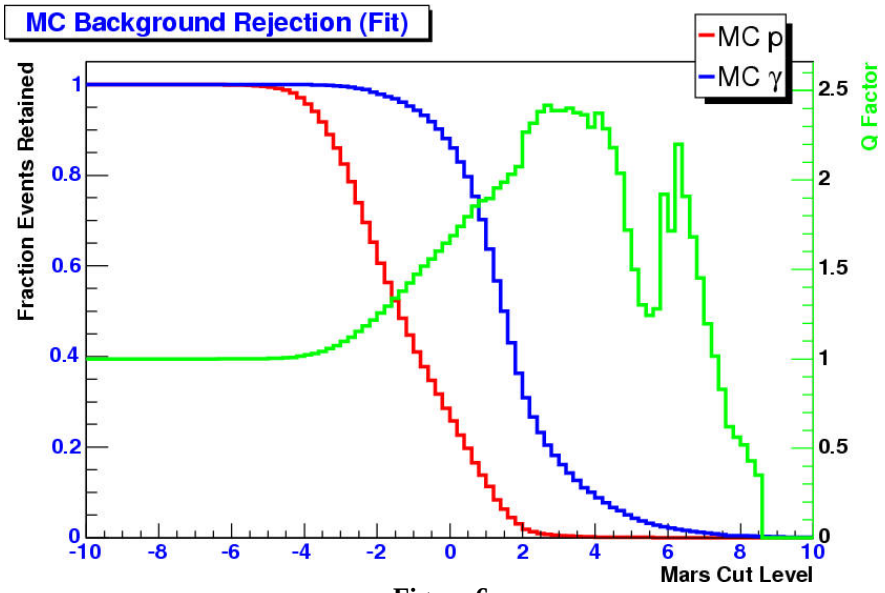

Figure 6

5 parameters, off-pond. Background and signal efficiencies for various Q-factors.

One immediate problem that was seen from making these histograms was that it was difficult to match Monte Carlo proton data with real data. MC-data matching is another topic which is being pursued currently. Nevertheless, Q-factors of more than 2 can still be seen without much difficulty in the above histograms. In fact, the real data curves seem to lie below the proton curves in the γ-like region of the histograms, meaning that even higher Q-factors can be found if one uses real data as the background instead of MC proton events.

Next, we look at the results from a 2 parameter MARS model. Gus Sinnis first implemented these parameters as part of a γ-h parameter study in July, 2004.

MARS input: 2 parameters

The parameters are x2 and nOut/nTop – 0.05(nb2/cxPE).

nOut = number of hit outrigger tubes

nTop = number of hit air shower layer tubes

cxPE = maximum number of PEs in a muon layer tube where a region 10m around the core is excluded from consideration; cxPE is now in Milinda as part of GHStat.h and so can be used by anyone

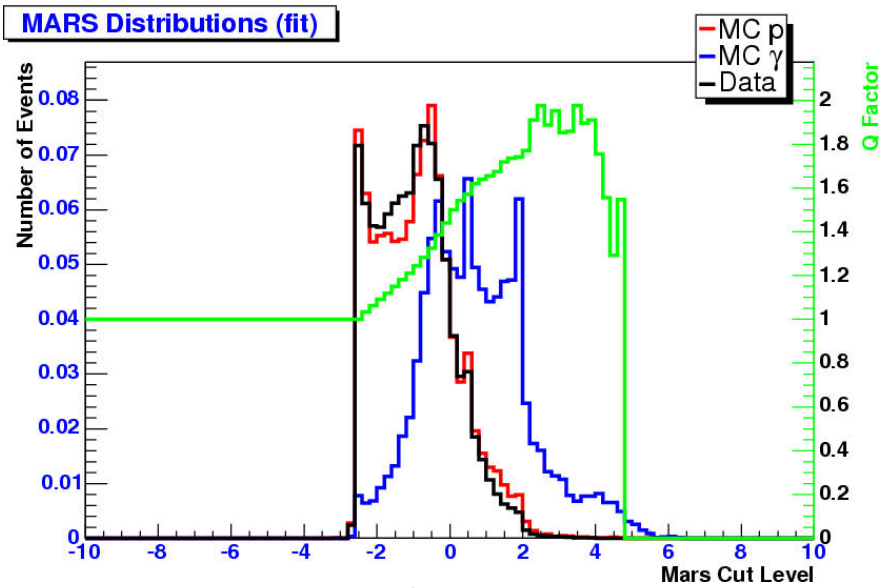

Figure 7

2 parameters, on+off pond. MARS distributions for the 2 parameter case with cores either on or off the pond. MC-data agreement is better than in the case of 5 parameters, though the peak Q-factor is slightly lower.

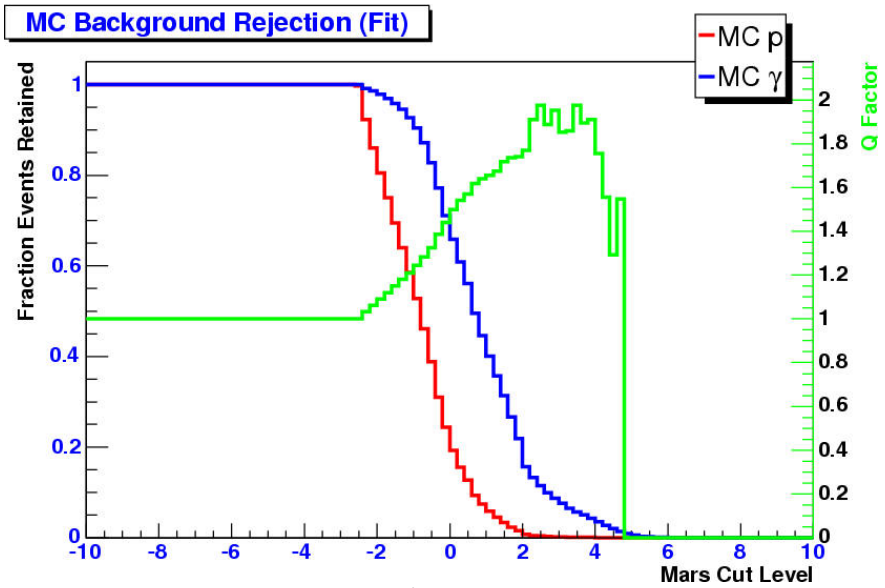

Figure 8

2 parameters, on+off pond. Background and signal efficiencies for various Q-factors.

Figure 7 shows MARS distributions for this 2 parameter case. Events with cores falling on and off the pond are included. The agreement between MC data and real data is much better here than in the 5 parameter case, though the peak Q-factor is slightly less.

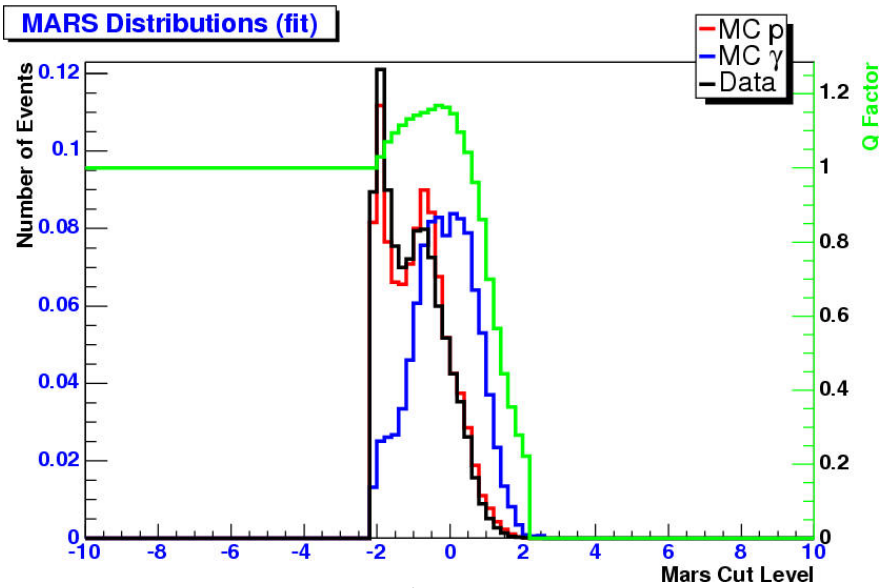

Figure 9

2 parameters, on-pond. MARS distributions for cores falling on the pond with the 2 parameter case. The double-peak is more pronounced for this type of event. As in the 5 parameter case with cores on the pond, the γ-distribution is centered about a MARS value of zero.

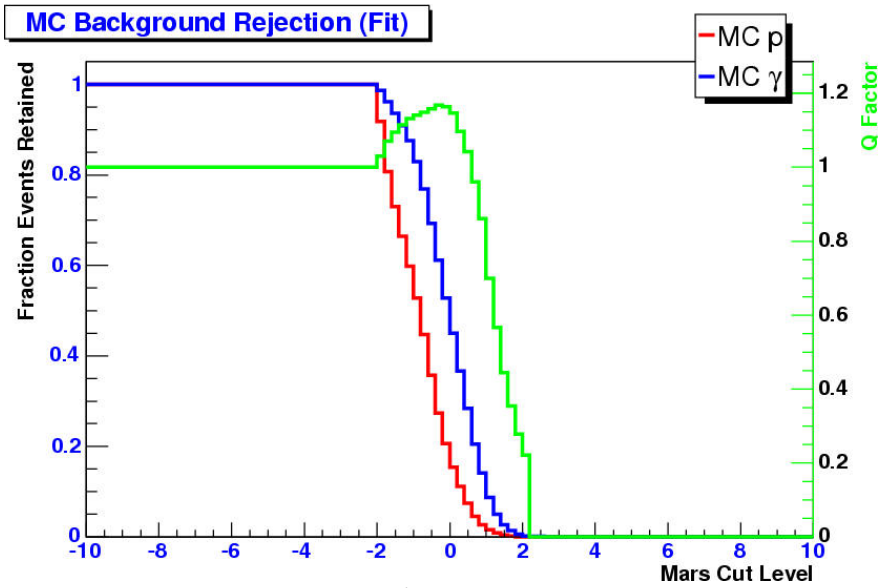

Figure 10

2 parameters, on-pond. Background and signal efficiencies for various Q-factors.

Figure 9 shows the distributions found when looking at cores falling on the pond only. The double-peaks are still seen for on-pond events.

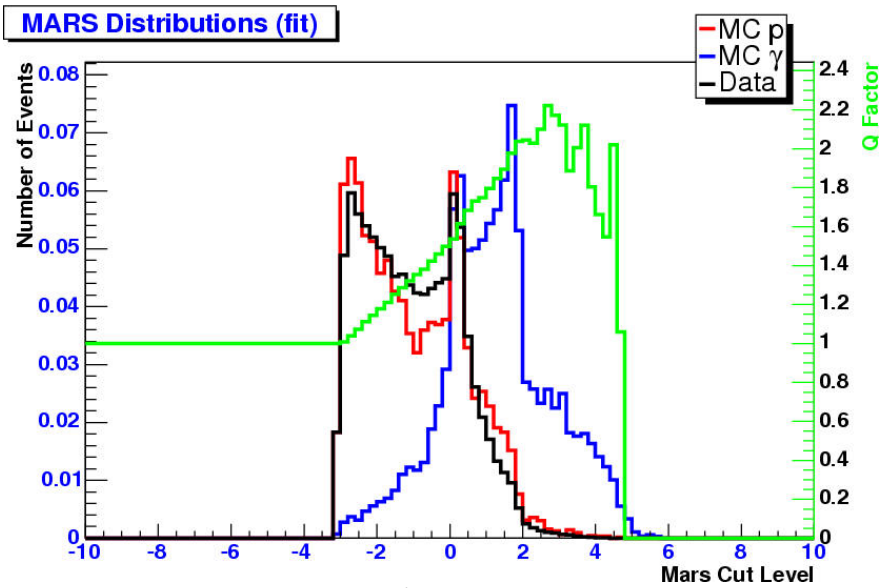

Figure 11

2 parameters, off-pond. MARS distributions for off the pond events. The double-peaks are still very visible, though they are wider and further apart than in the on-pond case.

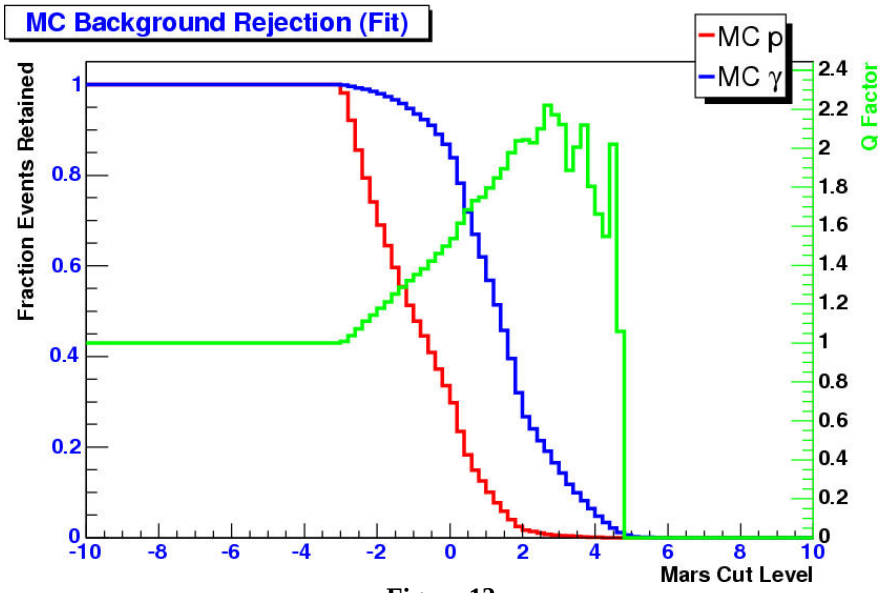

Figure 12

2 parameters, off-pond. Background and signal efficiencies for various Q-factors.

Figure 11 shows the off pond events for the 2 parameter case. The double-peaks are again quite visible. The cause of the double peaking is still being looked into.

The Q-factors in both 2 and 5 parameter cases reach values of 2 or more in the on+off-pond and off-pond distributions. However, the on-pond distributions give Q-factors that are quite small, roughly about 1.2.

Due to differences between proton and real data distributions, using real data as the background is also considered. The preceding results used MC proton events in the

MARS modeling. The following histograms are found by using real data, from sub-run 5269_0384, in the MARS modeling step as opposed to MC protons.

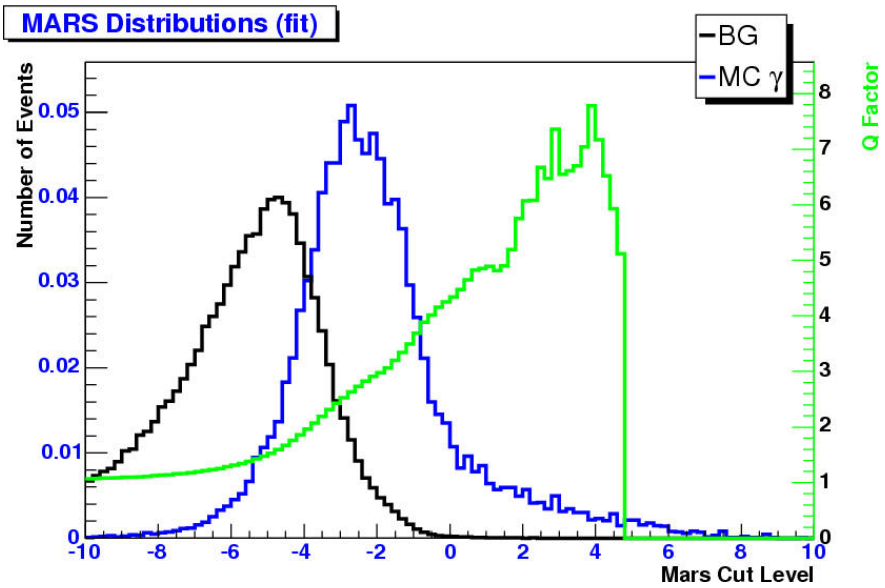

Figure 13

5 parameters, on+off pond, data as background. MARS distributions using data in the modeling step.

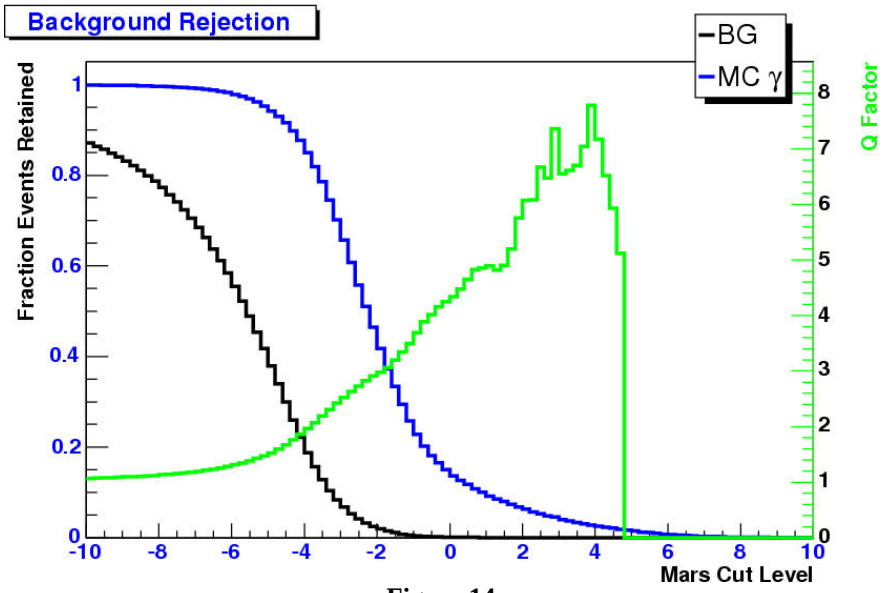

Figure 14

5 parameters, on+off pond, data as background. Background and signal efficiencies for various Q-factors.

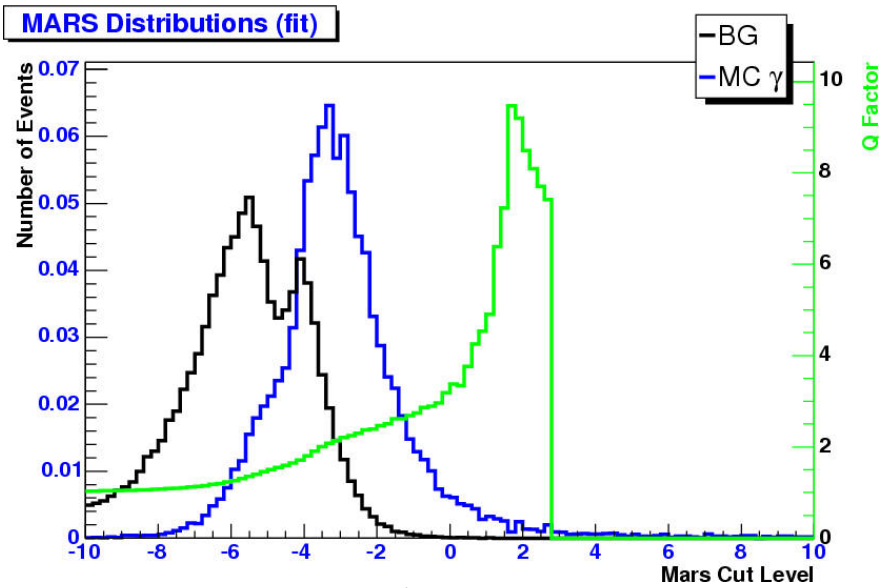

Figure 15

5 parameters, on-pond, data as background. MARS distributions using data in the MARS modeling as the background.

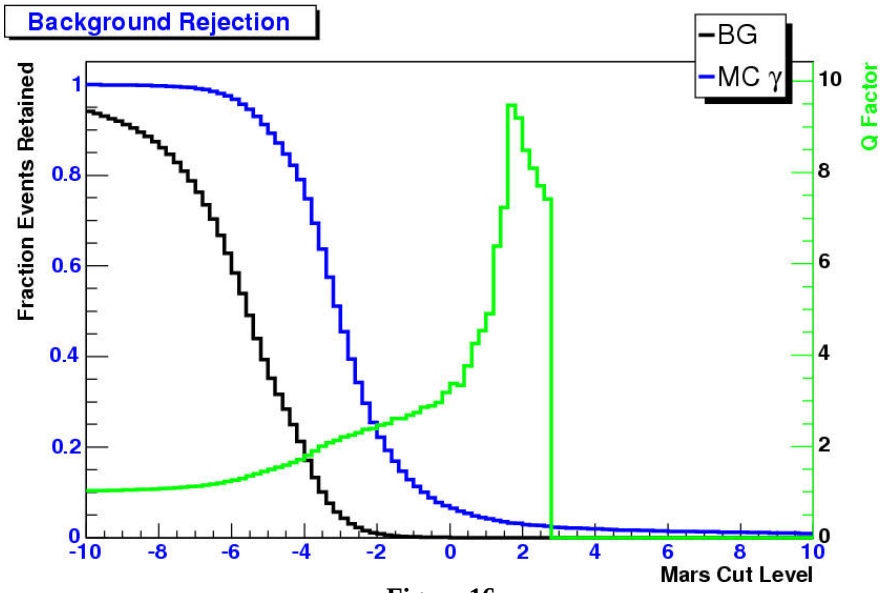

Figure 16

5 parameters, on-pond, data as background, efficiencies. Background and signal efficiencies for various Q-factors.

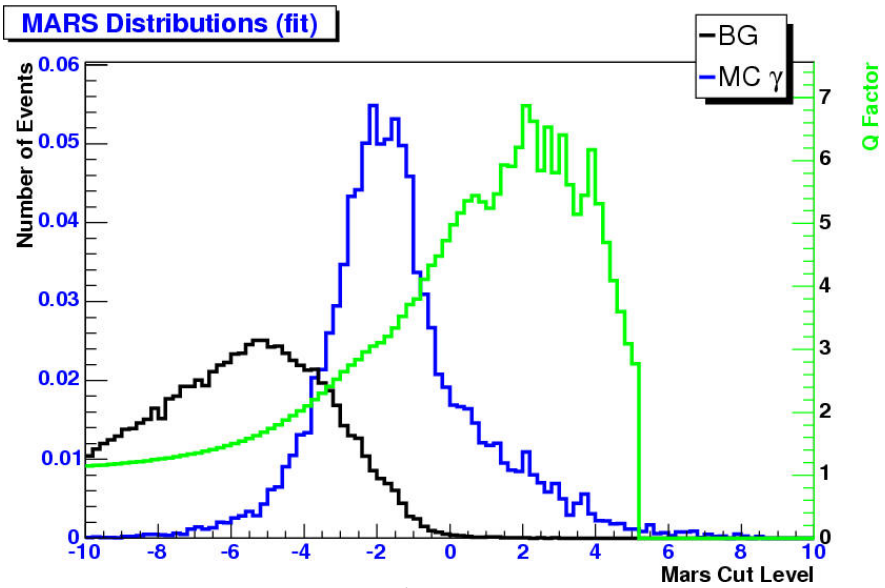

Figure 17

5 parameters, off-pond, data as background. MARS distributions for off-pond data using data in the MARS model.

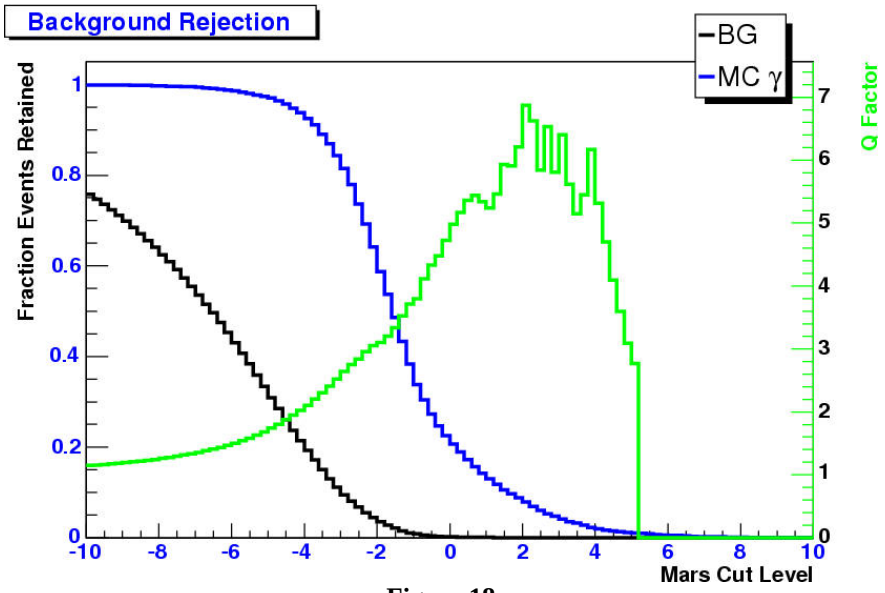

Figure 18

5 parameters, off-pond, data as background. Background and signal efficiencies for various Q-factors.

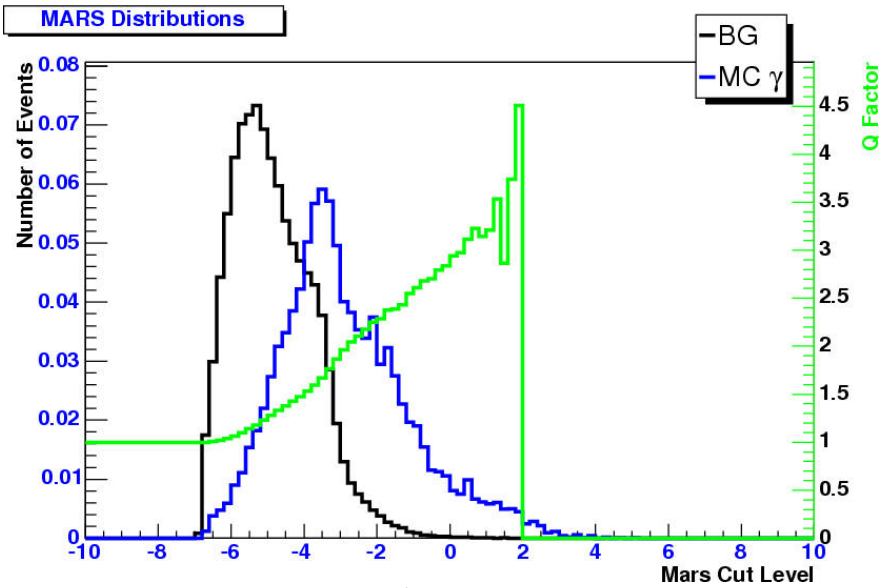

Figure 19

2 parameters, on+off pond, data as background. MARS distributions for on+off pond events using data in the MARS model.

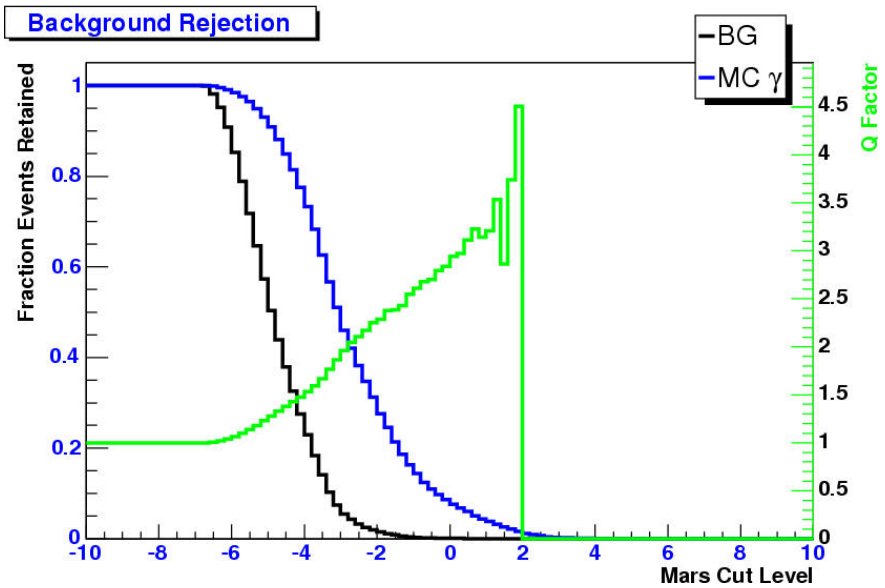

Figure 20

2 parameters, on+off pond, data as background. Background and signal efficiencies for various Q-factors.

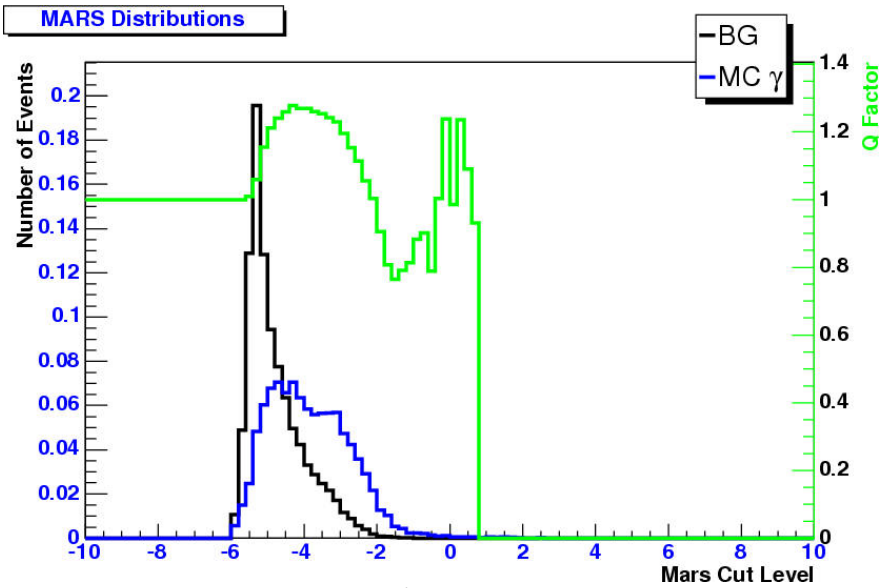

Figure 21

2 parameters, on-pond, data as background. MARS distributions for on-pond events using data in the MARS model.

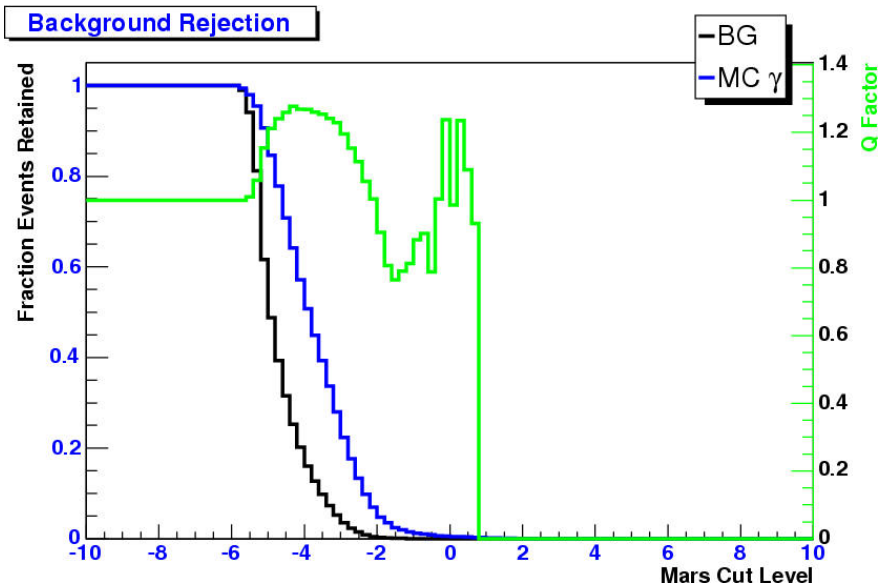

Figure 22

2 parameters, on-pond, data as background. Background and signal efficiencies for various Q-factors.

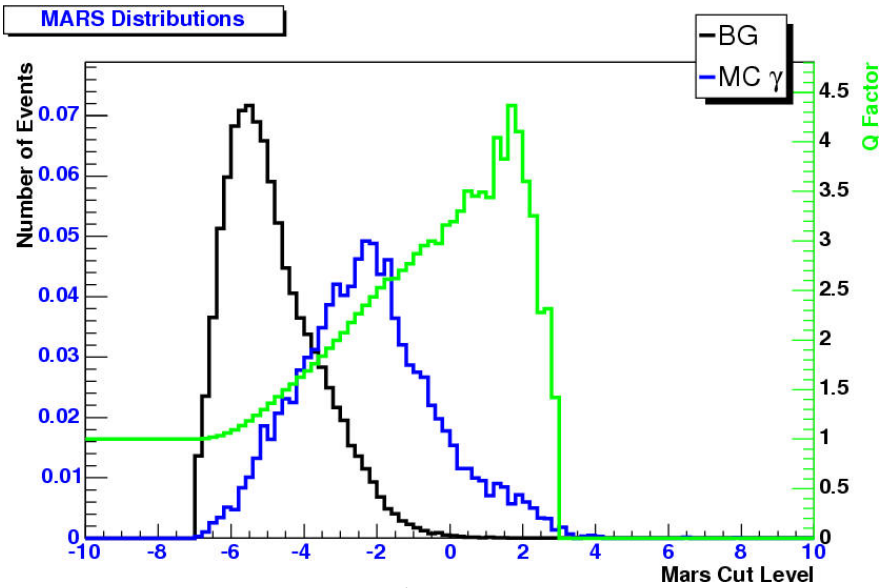

Figure 23

2 parameters, off-pond, data as background. MARS distributions for off-pond events using real data in the MARS model.

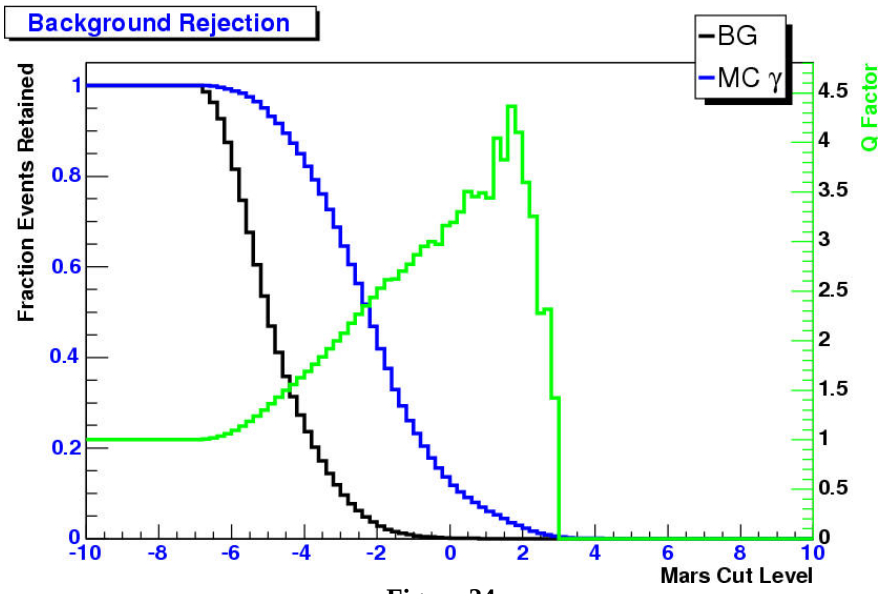

Figure 24

2 parameters, off-pond, data as background, efficiencies. Background and signal efficiencies for various Q-factors.

One striking feature of the MARS distributions when using real data as the background in the MARS modeling is that the peaks are in the negative MARS value region. MARS is essentially a black box in this analysis, so it is uncertain why this would be the case. Also, the distributions are much wider than when MC protons are used at the background.

An argument for using real data as the background is that it is expected to be nearly 100% composed of real background events as opposed to real γ events. Also, statistics are much better for these background events. Sub-run 5268_0384 alone has almost four times the statistics of the MC proton data available. However, to get a better feel for what the background is really like, samples of real data from many different Milagro epochs should be used.

Conclusion

This work is still on-going. This note only showed some of the best results obtained with MARS on a choice set of parameters. Ideally, one could feed MARS any number of parameters without much thought and have it return optimal predictions. In actuality, throwing in random parameters seems to reduce the effectiveness of MARS, so one must put more consideration into individual parameters. Specific cut values are still handled by MARS though, saving on some human-work. Once optimal Q-factors are obtained for on and off the pond, separately, a total Q-factor can be calculated based on the efficiencies of events of each type falling on and off the pond. This yields a 2-dimensional Q-factor, which can then be applied to various analyses. The figures below show the total Q-factor, γ efficiency, and proton efficiency for the 5 parameter case using MC proton events as the background as an example of what the total distributions look like.

Using real data as the background source for the MARS models provides distributions quite different from the MC proton case. Larger Q-factors can be obtained with real data as the background, but much wider MARS distributions which peak well below a MARS value of zero also appear. An argument can be made for using real data anyway, but the discrepancy between real data and MC protons still needs to be understood and solved.

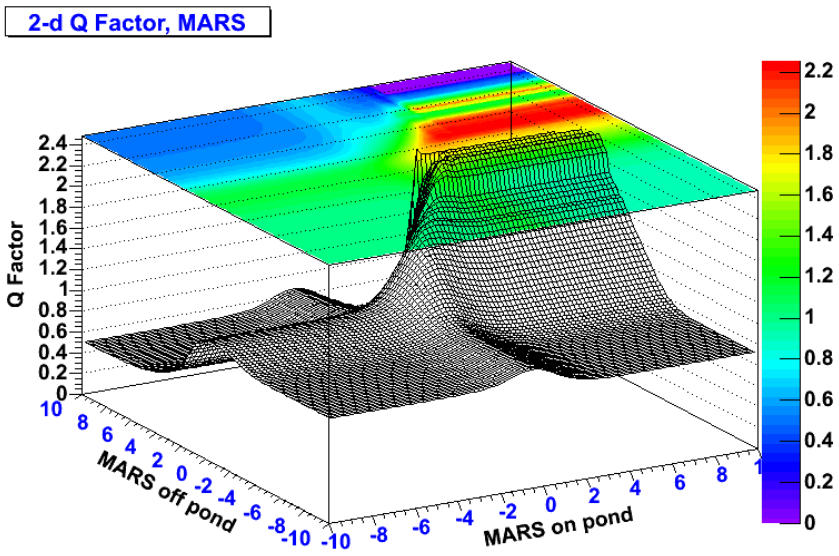

Figure 25

5 parameters, total Q-factor. This shows the combined Q-factor for a given pair of MARS values.

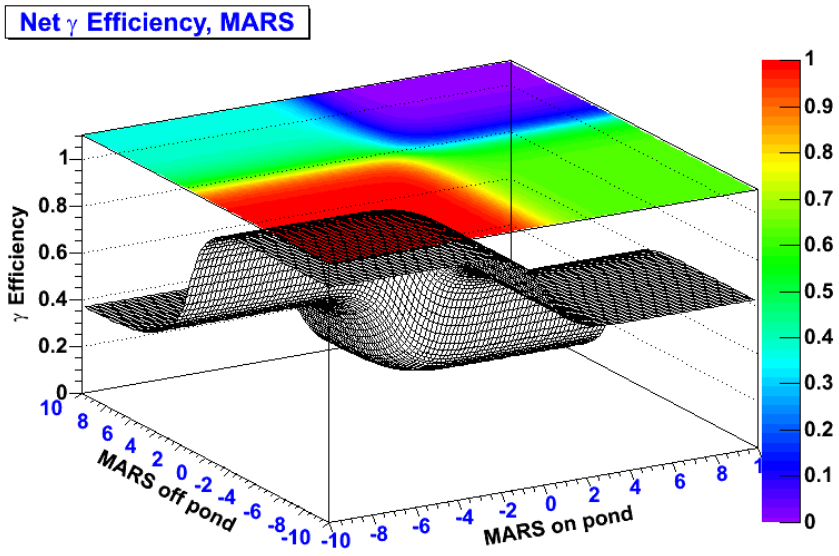

Figure 26

5 parameters, total γ efficiency. This gives the total γ efficiency for a given pair of MARS values.

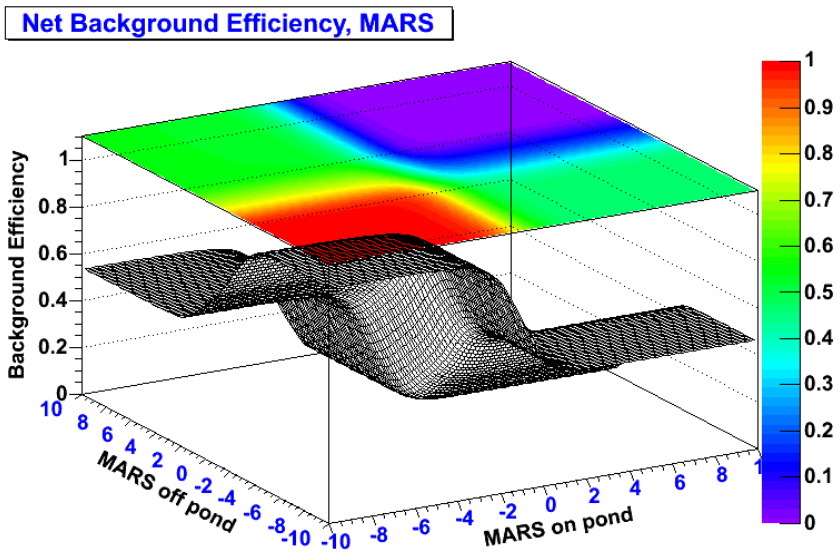

Figure 27

5 parameters, total proton efficiency. This gives the total proton efficiency for a given pair of MARS values.

New Results – Added 10/26/04

More parameters have been added to MARS to see if a plateau or peak in maximum Q-factors can be obtained. So far, up to 12 parameters have been used. Also, a tighter $\Delta(\theta,\phi)<0.7$ cut on MC γ events was used in the following.

MARS input: 12 parameters
The parameters are x2, mxPE, sumPEtop, sumPEbot, sumPEout, nb2, nb8, nOR/nAS – 0.05(NB2/cxPE), hitas, hitmu, hitor, and pchi2.

pchi2 = describes the lumpiness of events; this is in Milinda as part of GHStat.h; Brenda Dingus first implemented this parameter

$$pchi2 = \chi^2_{PE} = \left[\sum_i \frac{\left[PE_i - \overline{PE_{nn}} \right]^2}{PE_i + \overline{PE_{nn}}/nn} \right] \Big/ N \qquad (3)$$

PE_i is each muon layer PMT, $\overline{PE_{nn}}$ is the average number of PEs for the nearest neighbors of the i^{th} PMT, *nn* is the number of nearest neighbors, and N is the number of PMTs with the sum of the nearest neighbors PEs being greater than 2. The sum in Equation (3) is done only for PMTs with nearest neighbor PE sums greater than 2.

The following histograms use MC protons as background. Distributions and event efficiencies are shown for events either on or off the pond, on the pond only, and off the pond only.

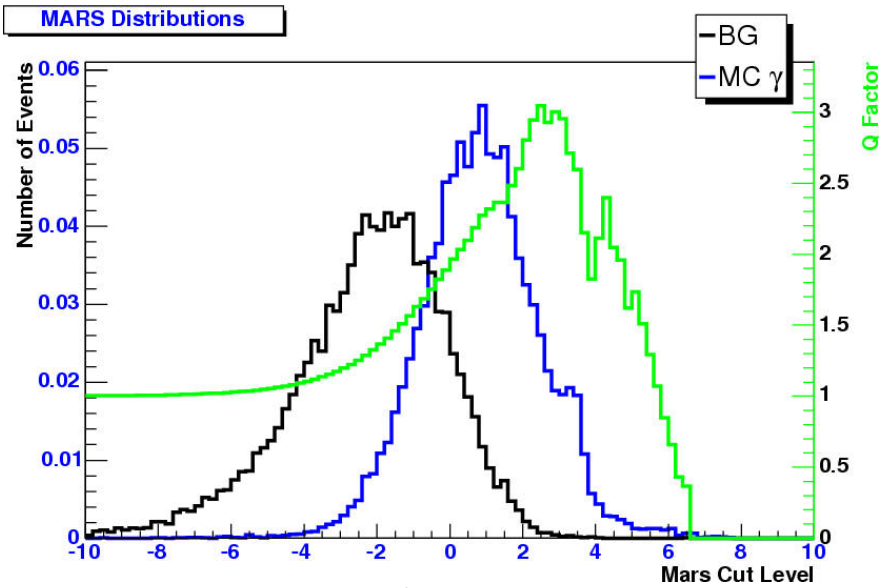

Figure 28

12 parameters, on+off the pond. The maximum Q-factor is now up to 3. The background is MC protons.

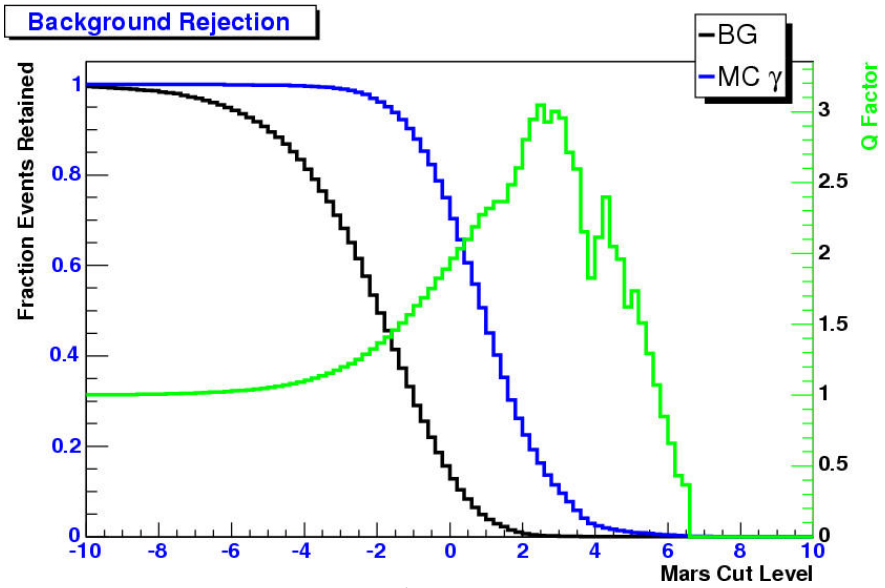

Figure 29

12 parameters, on+off the pond. At a Q-factor of 2.5, a good fraction of signal is still kept.

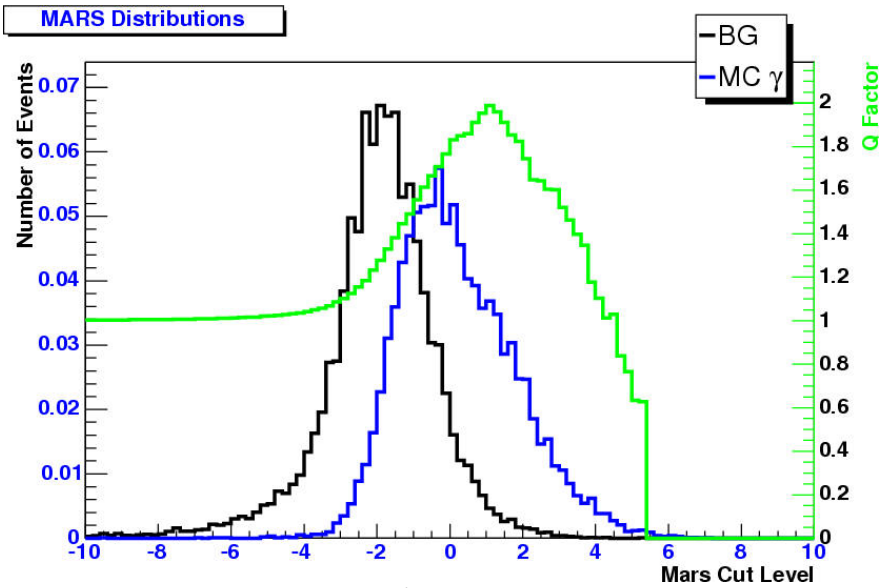

Figure 30

12 parameters, on the pond. Even for on-pond events, the Q-factor is very good compared to previous results.

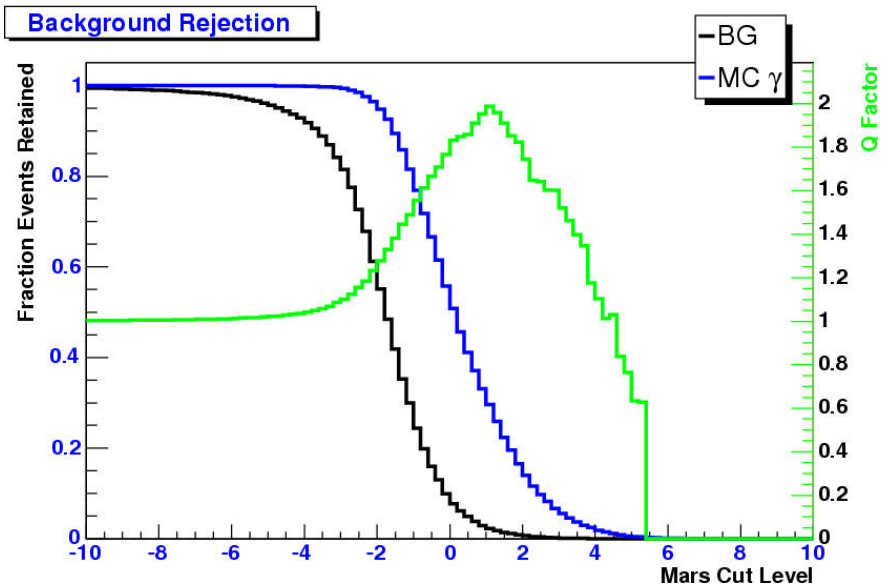

Figure 31

12 parameters, on the pond. The peak Q-factor for on-pond events still retains many signal events.

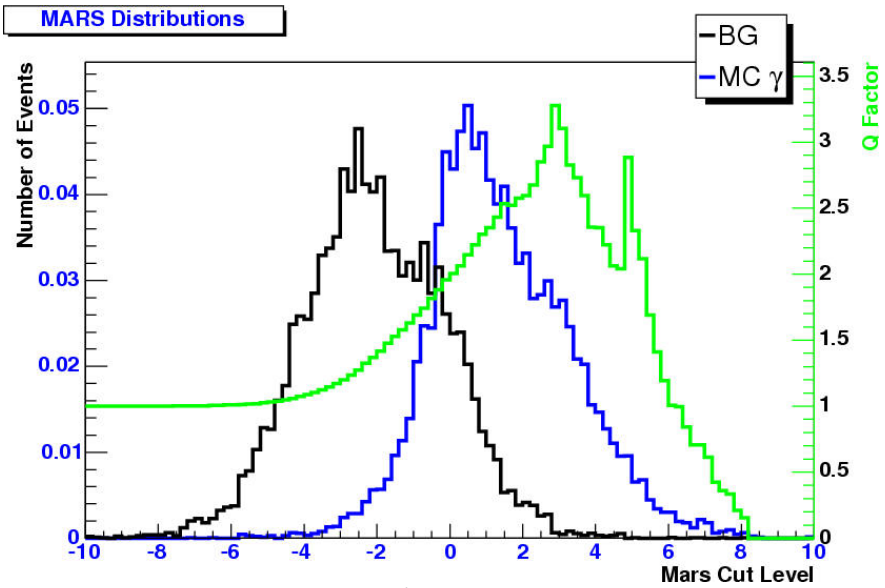

Figure 32

12 parameters, off the pond. Q-factors above 3 are seen for off-pond events.

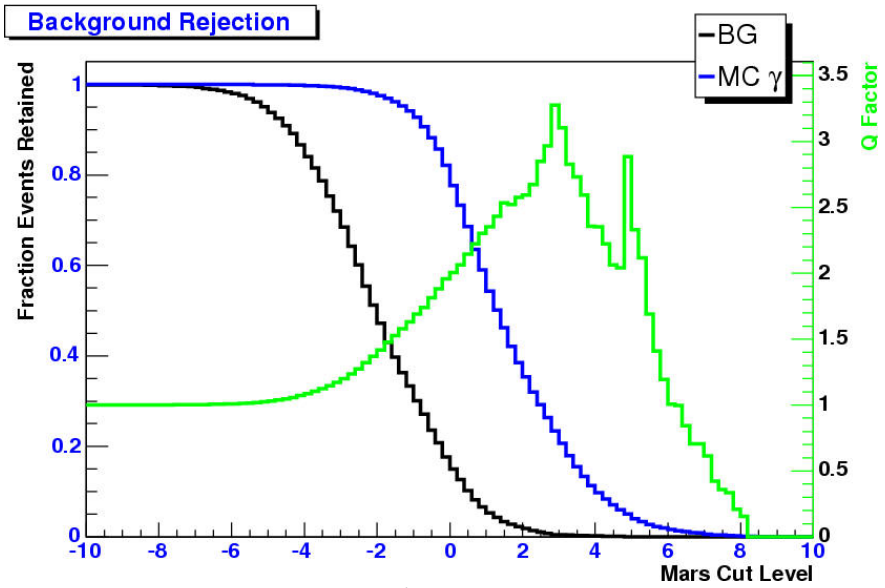

Figure 33

12 parameters, off the pond. Roughly 20% of the signal is still kept even for Q-factors above 3.

The 12 parameter case was also run with the older $\Delta(\theta,\phi)<1.2$ cut for MC γ events, though the results are not shown in this memo. The differences in Q-factors were substantial, with peak values increasing by roughly 0.5 to 1 for the three core location cases when using the tighter cut.