

What's wrong with MC prediction?

Stefano Lacaprara

INFN Padova

PD meeting,
PD, 7 March 2012



Intro

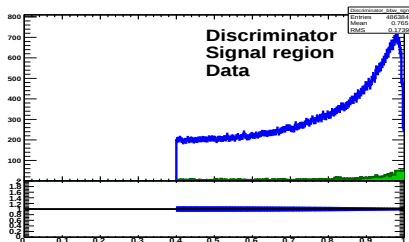
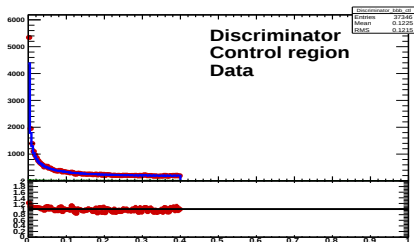
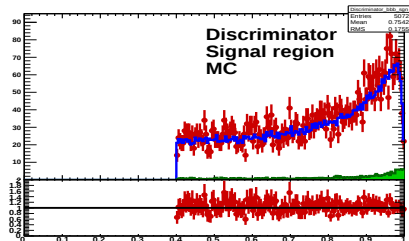
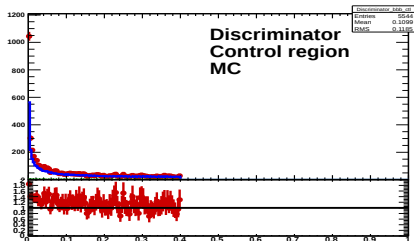


What is about

- Ok, so we have a problem in MC prediction (120 GeV)
- while data is much better
- it seems that the disagreement comes from Discriminator ~ 0
- So:
 - 1 Zoom discriminator around $D \sim 0$ and see what it looks like
 - 2 select only the very low value of Discriminator as a control region, and see how the distribution looks like
 - 3 try if it makes sense to remove the very first bin (in data and MC) and check how things are looking in this way
- try eff only and fb/c from MC truth, with full statistics.

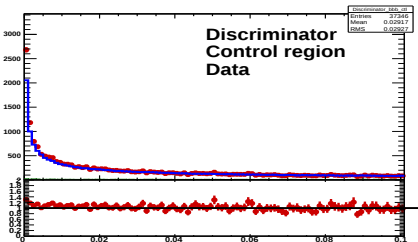
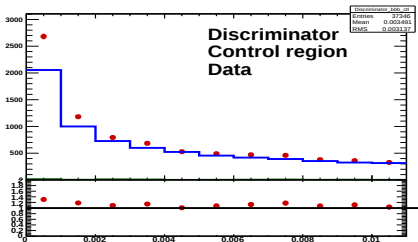
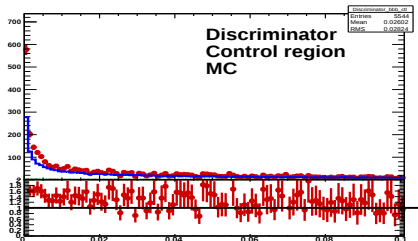
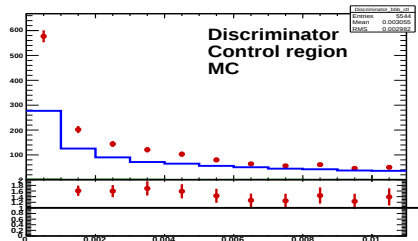


Discriminator





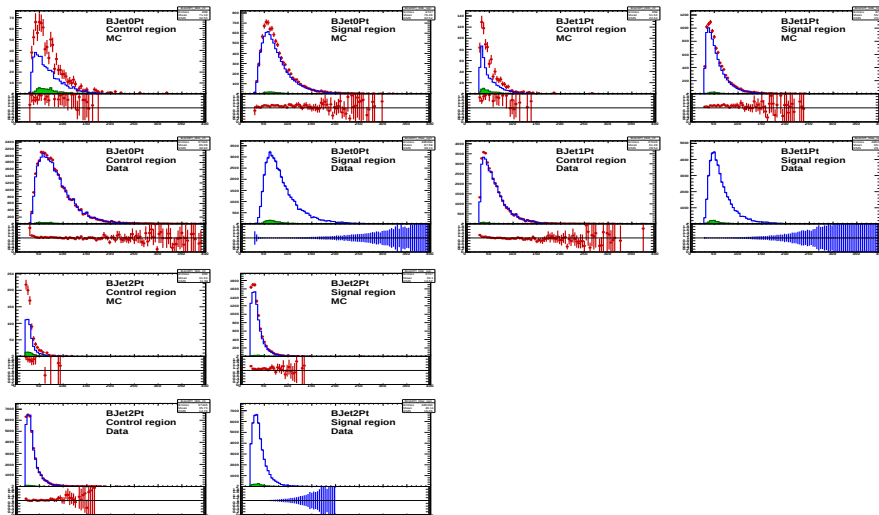
Discriminator Zoom

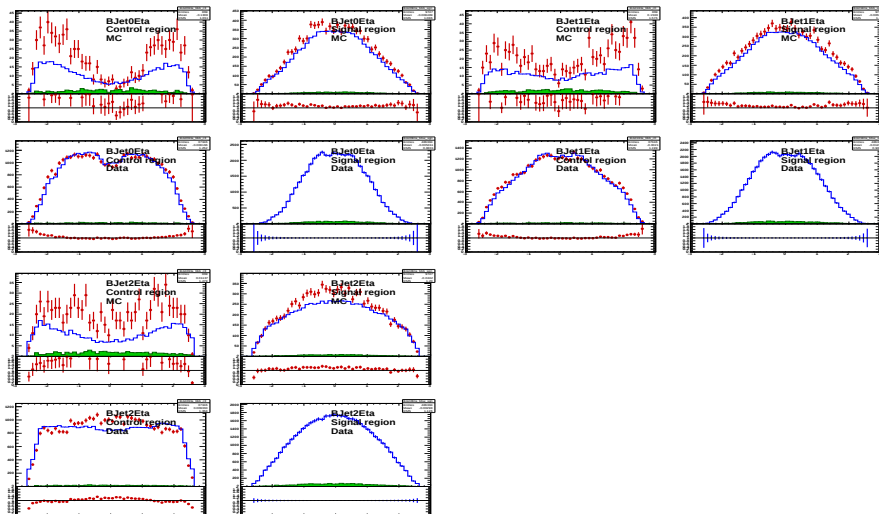


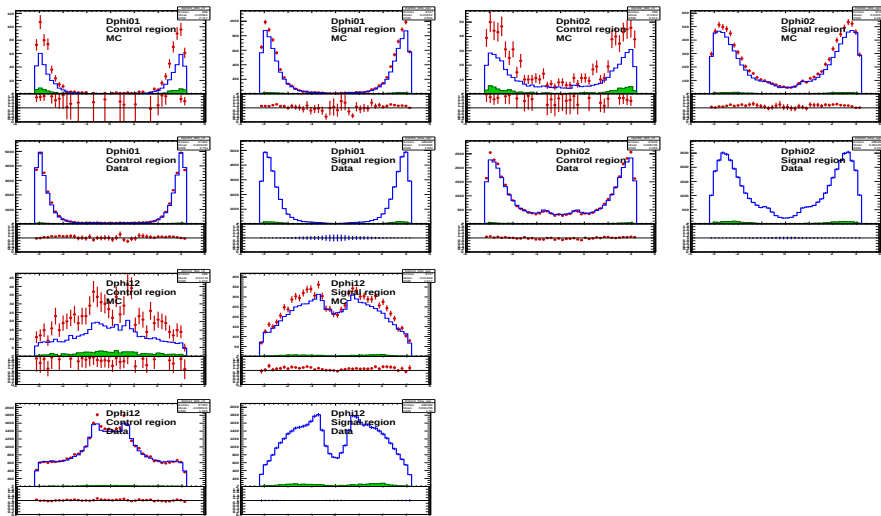
Second try

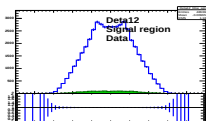
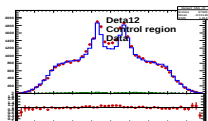
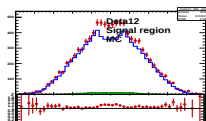
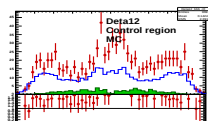
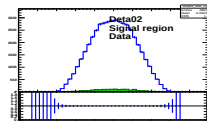
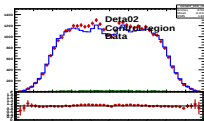
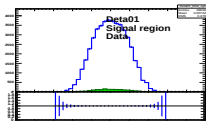
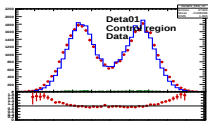
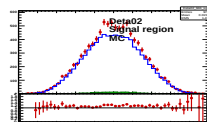
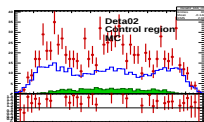
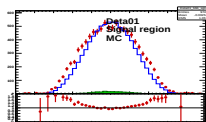
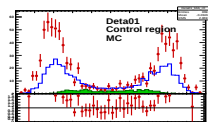
Ok, define (for MC only) control region as $D < 0.01$, so to see the distribution of all variables in the problematic region (control region) and everywhere else (signal region).

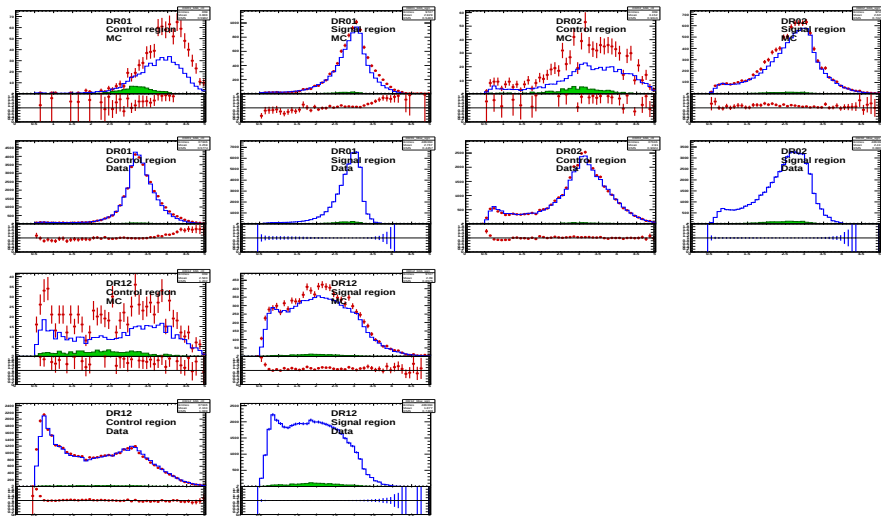
Same plots for $D < 0.001$ in the backup.

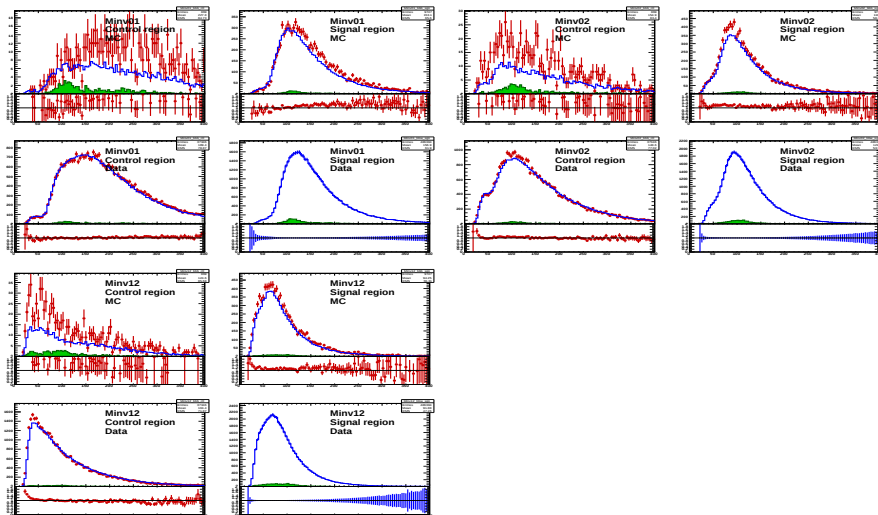


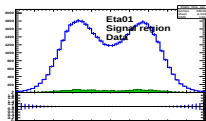
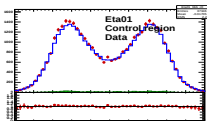
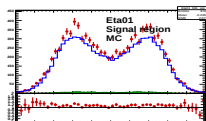
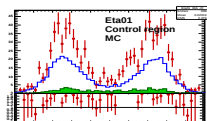
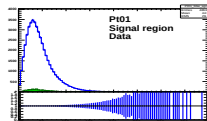
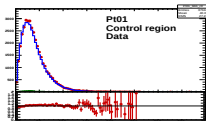
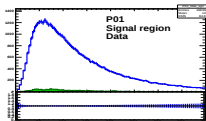
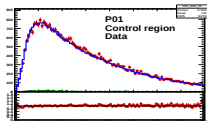
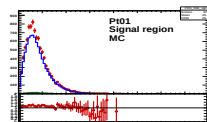
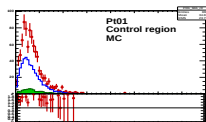
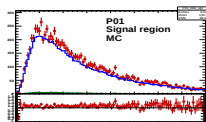
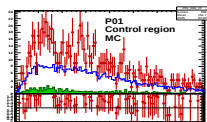






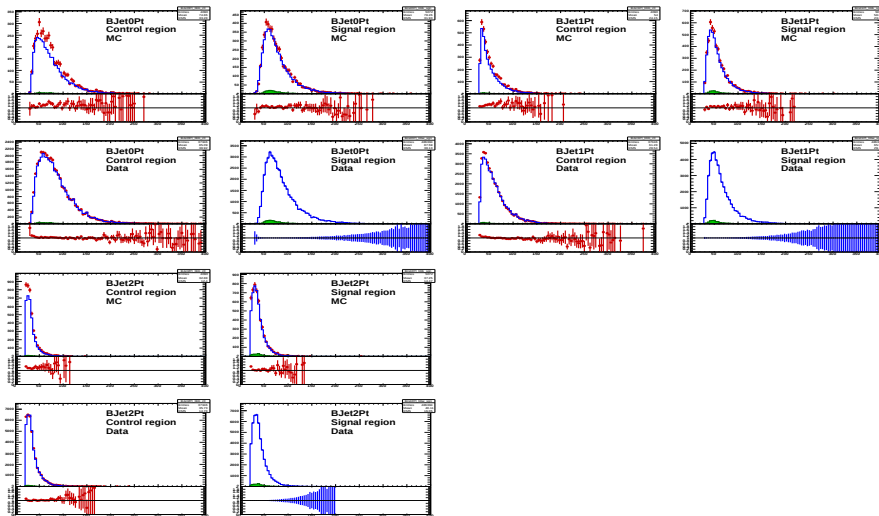


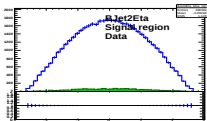
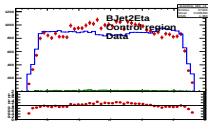
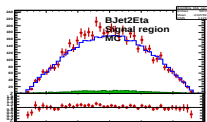
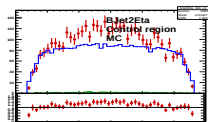
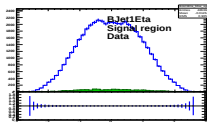
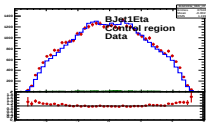
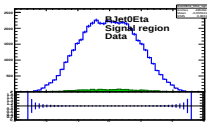
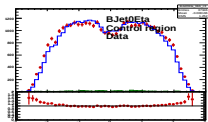
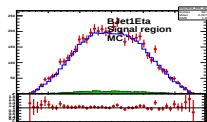
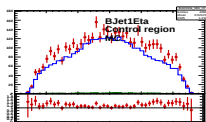
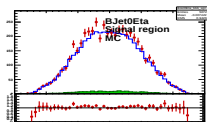
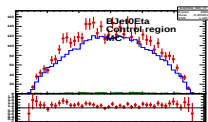


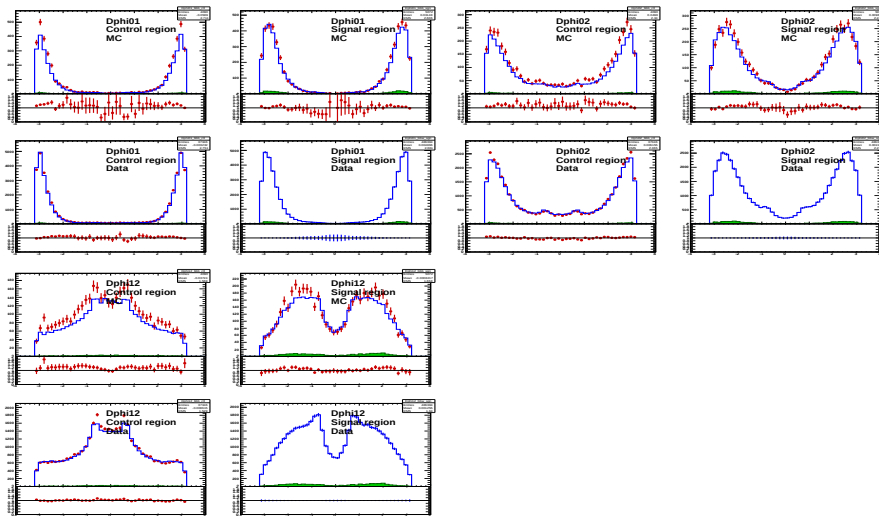


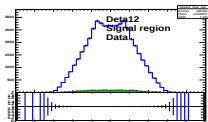
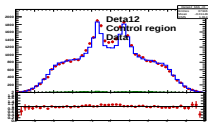
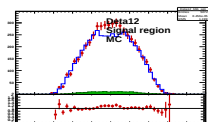
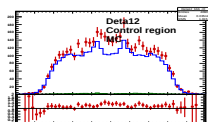
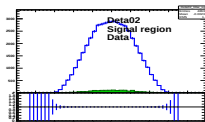
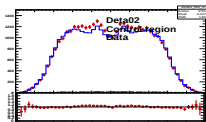
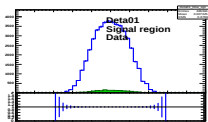
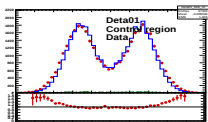
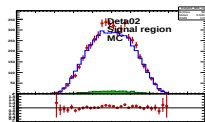
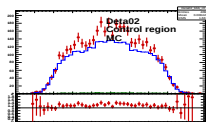
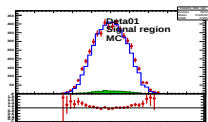
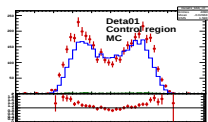
Third try

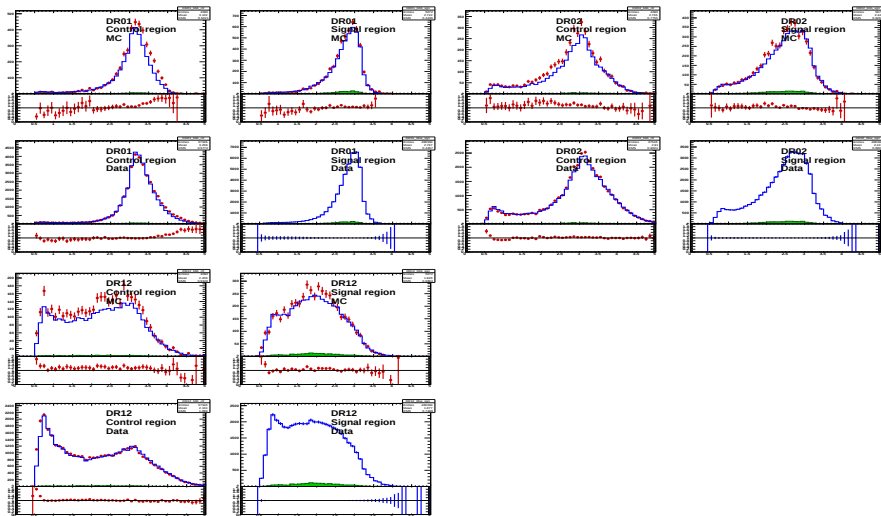
Ok, define (for MC only) control region as $0.01 < D < 0.04$ and signal region as $0.04 < D < 1$. and see what happen
Same plots for $0.05 < D < 0.04$ in the backup.

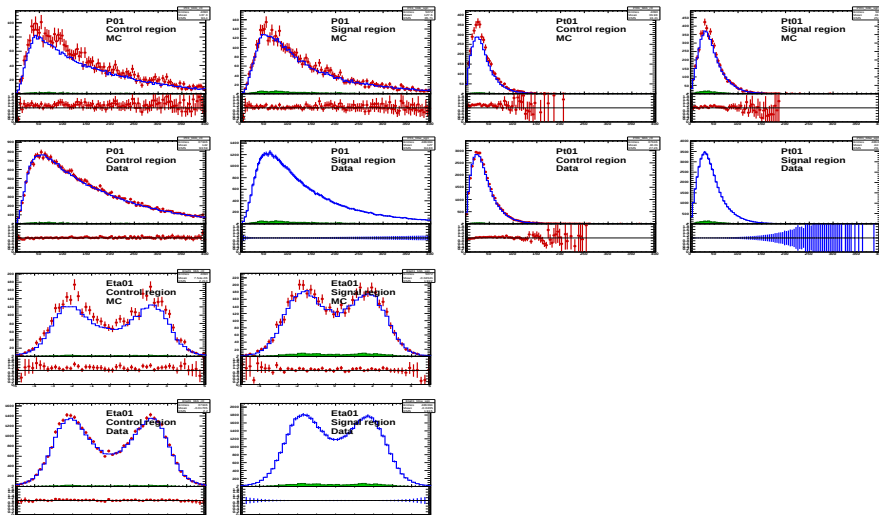












Going back to Fb/c

Ok, let's see where the MC non-closure is coming from.

First, try by applying ϵ_x **only**, and taking $F_{b,c}$ from MC truth, *jet per jet*.
Namely:

$$P_b^{3rd-j}(j) = \begin{cases} \epsilon_b & \text{if jet is B} \\ \epsilon_c & \text{if jet is C} \\ \epsilon_{light} & \text{otherwise} \end{cases}$$

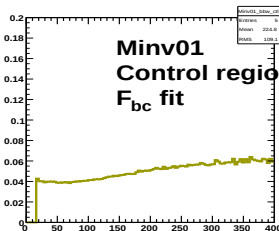
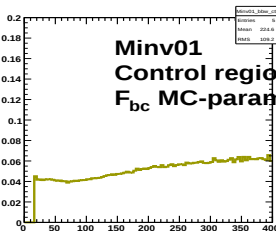
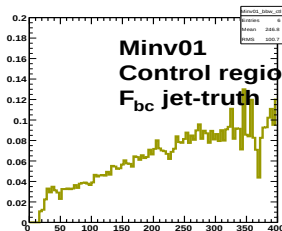
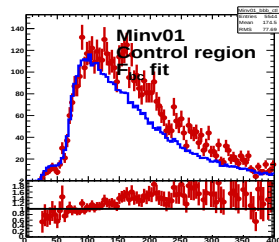
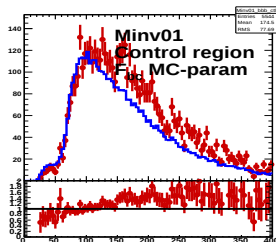
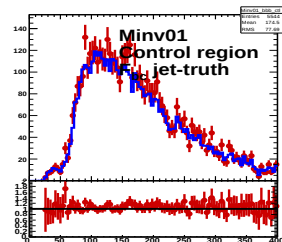
Compare, for MC control region only,

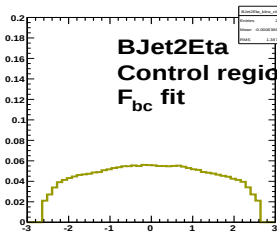
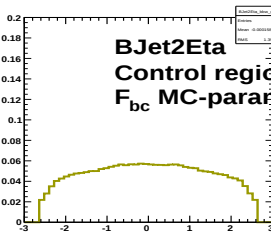
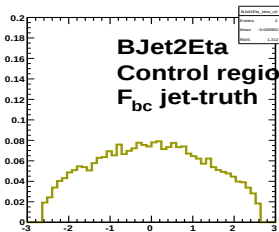
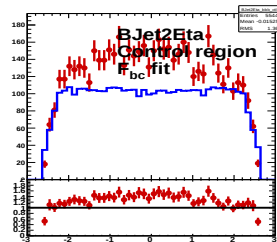
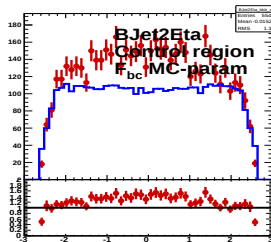
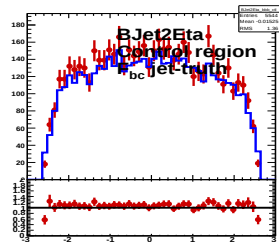
- the predictions with the above P_b (jet-truth)
- MC truth parametrized vs $(E_t^{(3)}, \Delta R_{min})$ (MC-param)
- F_{bc} from Fit.

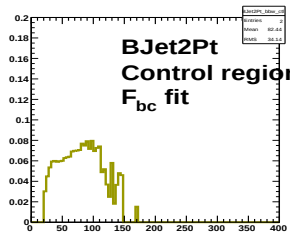
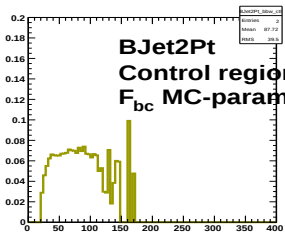
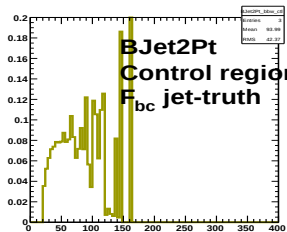
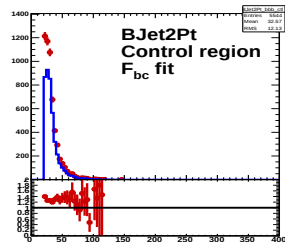
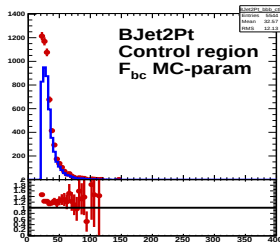
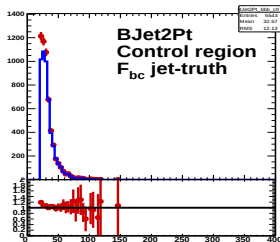
Plot P_b vs the chosen variable as well, and compare. Bottom row (yellowish line).

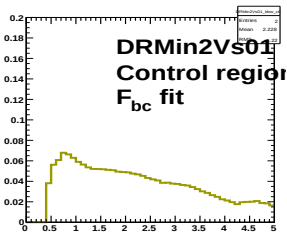
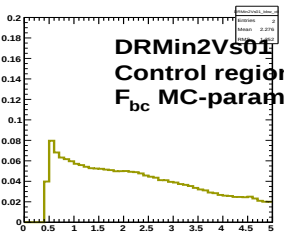
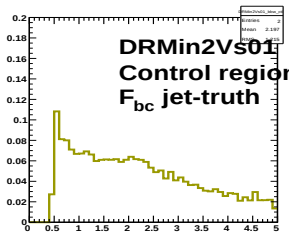
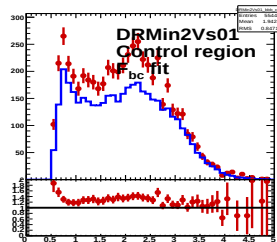
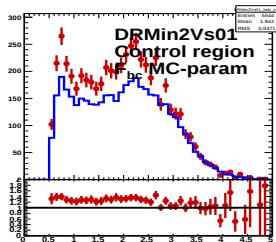
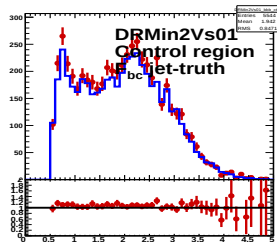


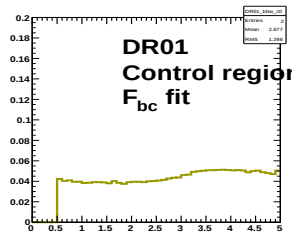
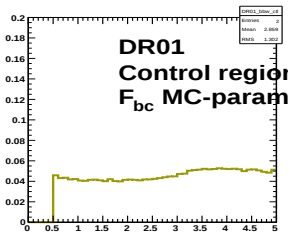
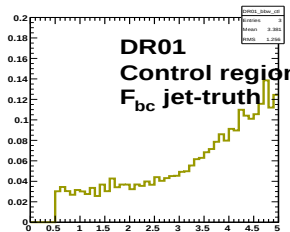
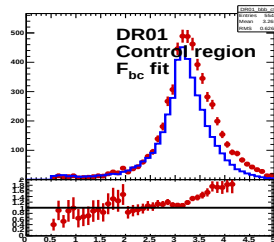
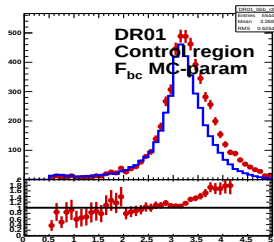
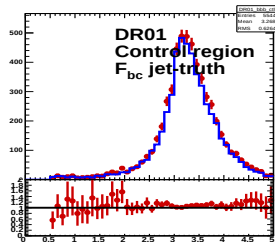
Comparing distribution from MC

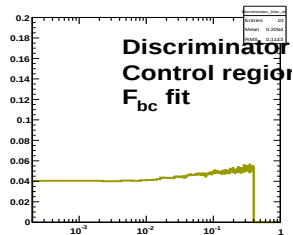
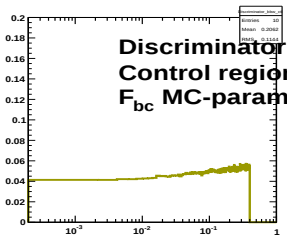
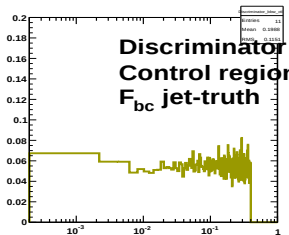
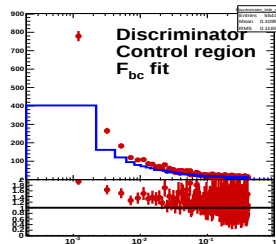
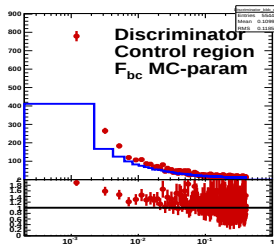
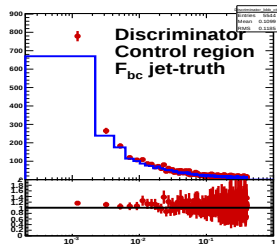














Update



Alert!

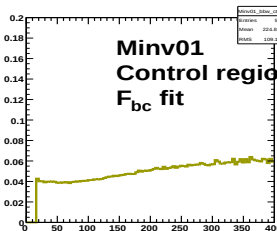
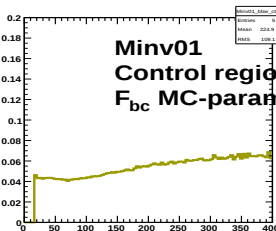
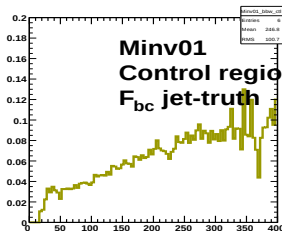
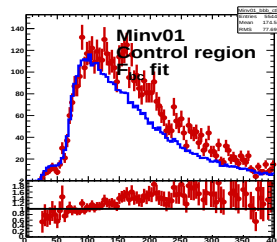
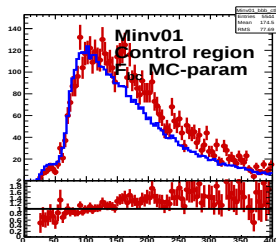
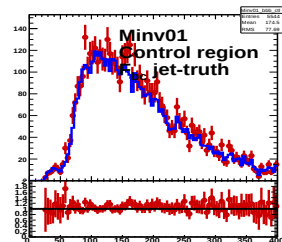
Found a bug, in the normalization of all jets for F_b , F_c : the FI was taken twice. F_b from 0.081 to 0.086 ($\sim 5\%$).

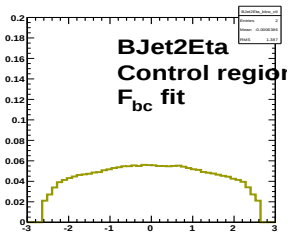
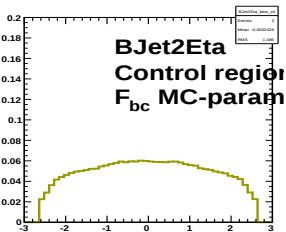
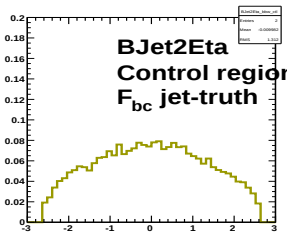
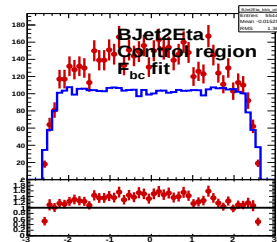
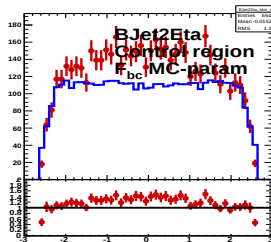
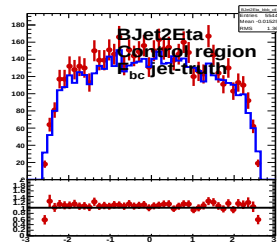
Re-run the btagging matrixes on the skim sample;
Check also with flat $F_{b,c}$, namely one hardcoded number for all events.
The ratio seen/predicted is like the following:

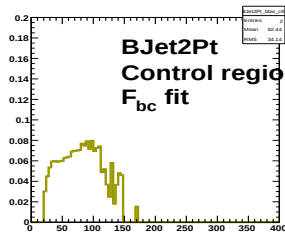
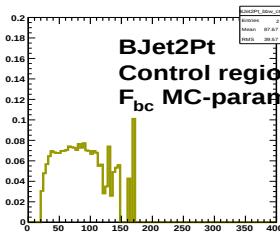
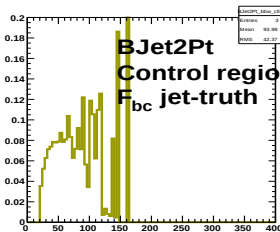
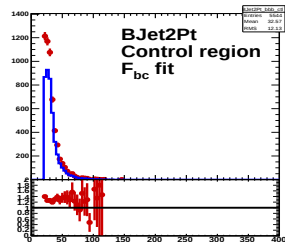
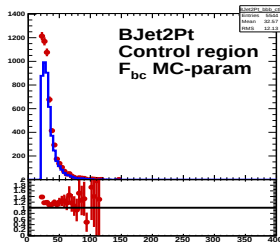
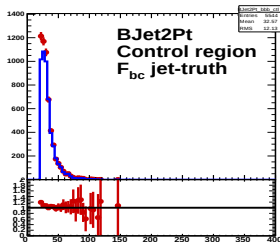
- Jet-truth (Seen) 5544/ (pred) 5148.19 = 1.07688
- Fit (Seen) 5544/ (pred) 4284.66 = 1.29392
- Param-truth (Seen) 5544/ (pred) 4402.88 = 1.25917 (with bug in)
- (Seen) 5544/ (pred) 4557.15 = 1.21655 (w/o bug)
- (Seen) 5544/ (pred) 4858.45 = 1.1411 w/ $f_B=8.59\%$; $f_C=4.67\%$; f_L 86.7%;
- (Seen) 5544/ (pred) 4596.07 = 1.2062 (w/o bug and using skim)

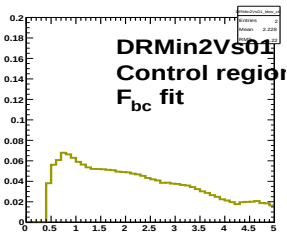
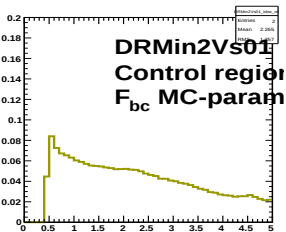
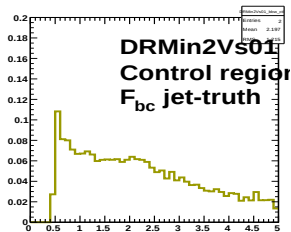
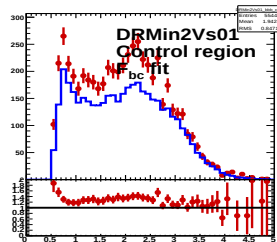
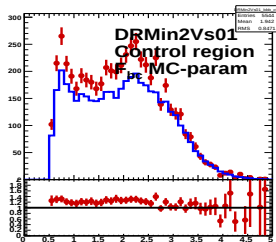
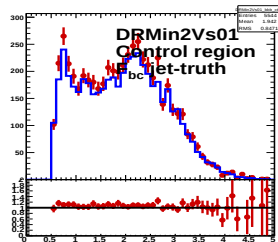


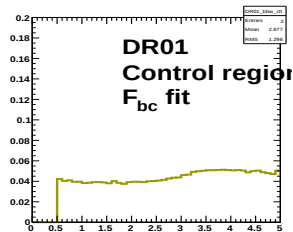
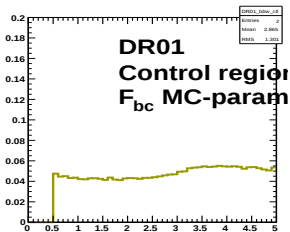
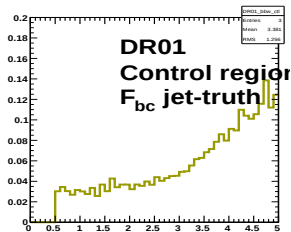
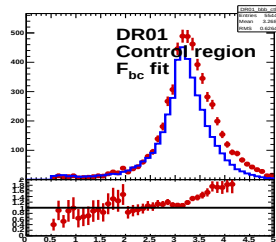
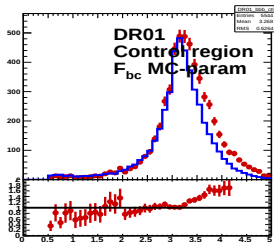
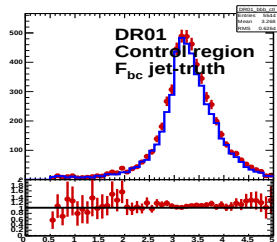
Comparing distribution from MC with improvement

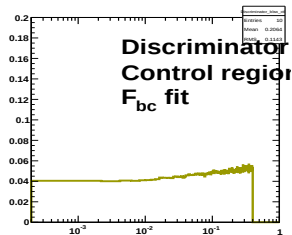
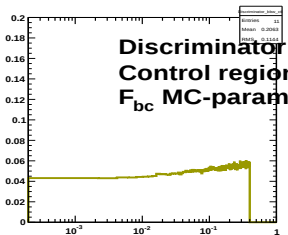
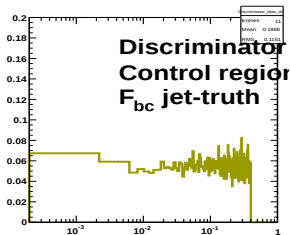
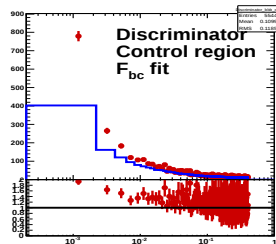
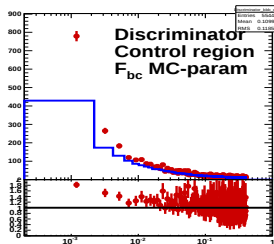
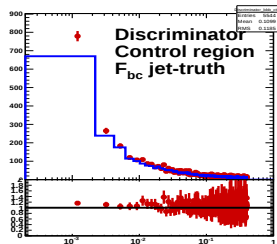














More Update



Alert!

For MC truth, I was evaluating $F_{b,c}$ in all phase space, not only in control region. Fixed.

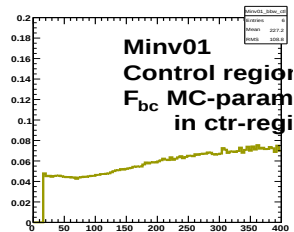
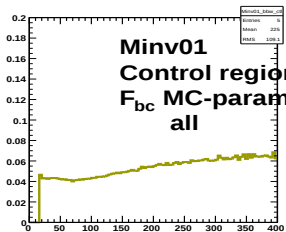
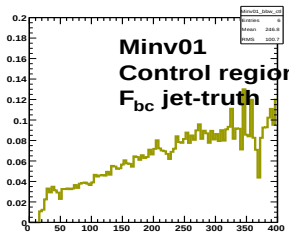
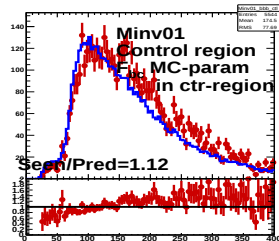
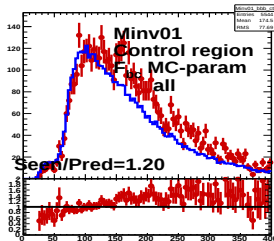
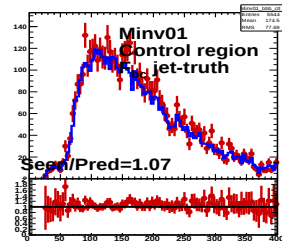
Seen/Pred:

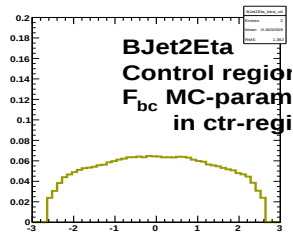
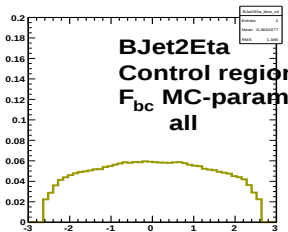
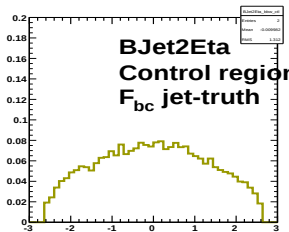
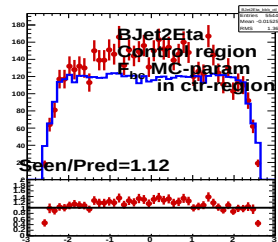
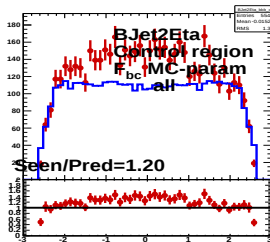
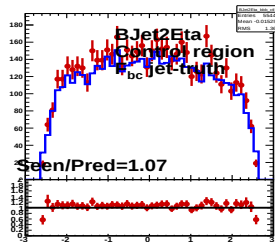
- Jet-truth $5544/5148.19 = 1.07688$
- Fit $5544/4284.66 = 1.29392$
- Flat $5544/4858.45 = 1.1411$ w/ fB=8.59%; fC=4.67%; fL 86.7%;
- Param in control reg. only $5544/4914.52 = 1.12809$
- Previous param, whole reg. $5544/4596.07 = 1.2062$

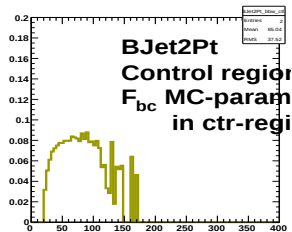
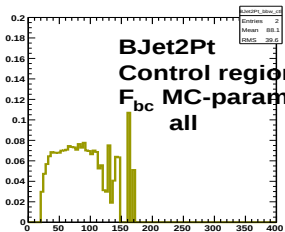
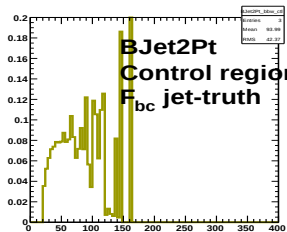
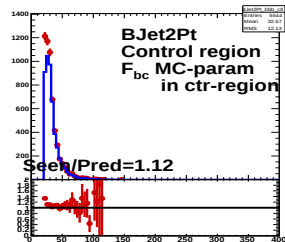
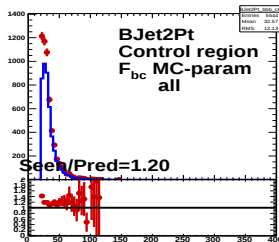
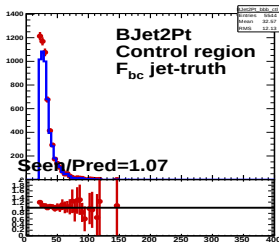
Closer and closer ...

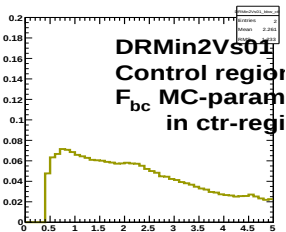
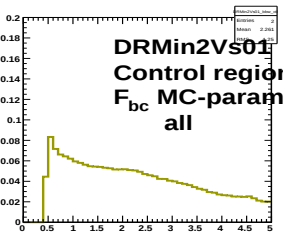
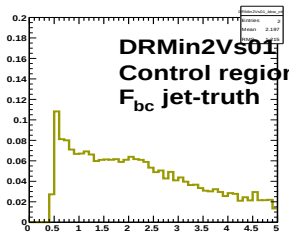
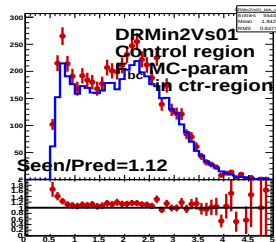
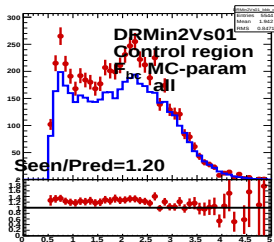
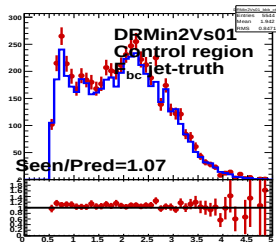


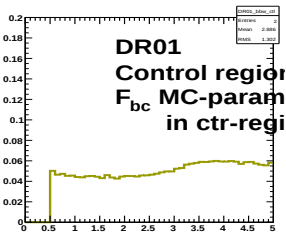
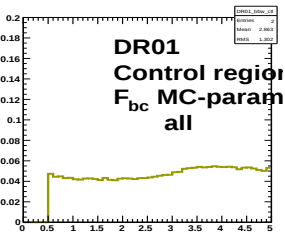
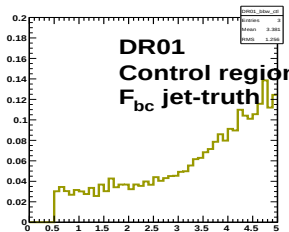
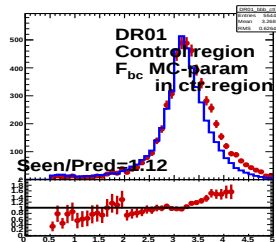
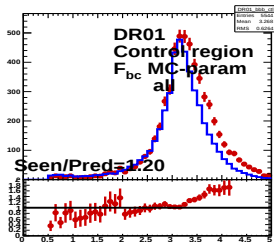
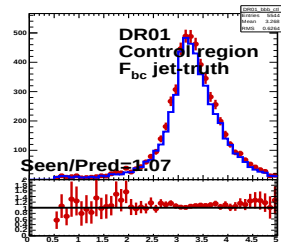
Comparing distribution from MC with further improvement

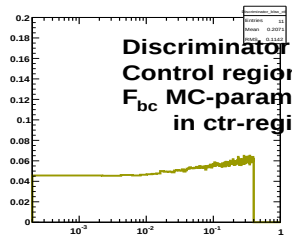
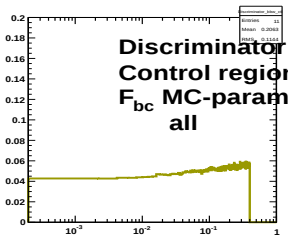
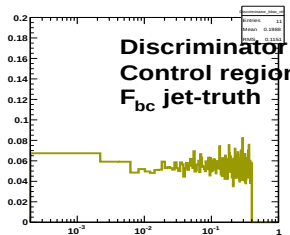
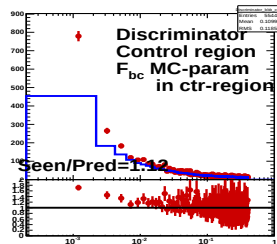
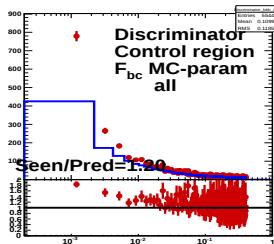
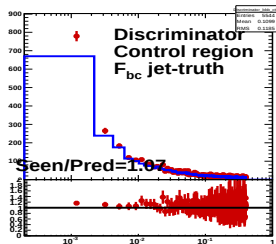














Conclusion

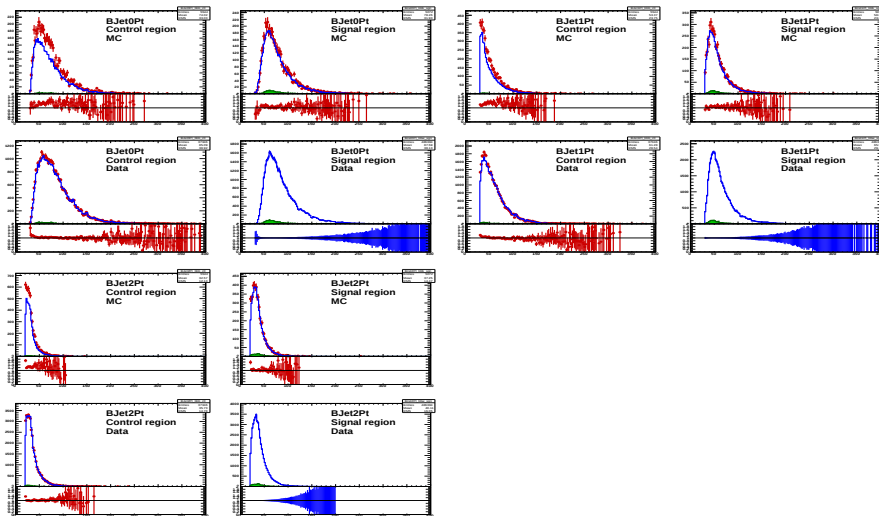
The bad news

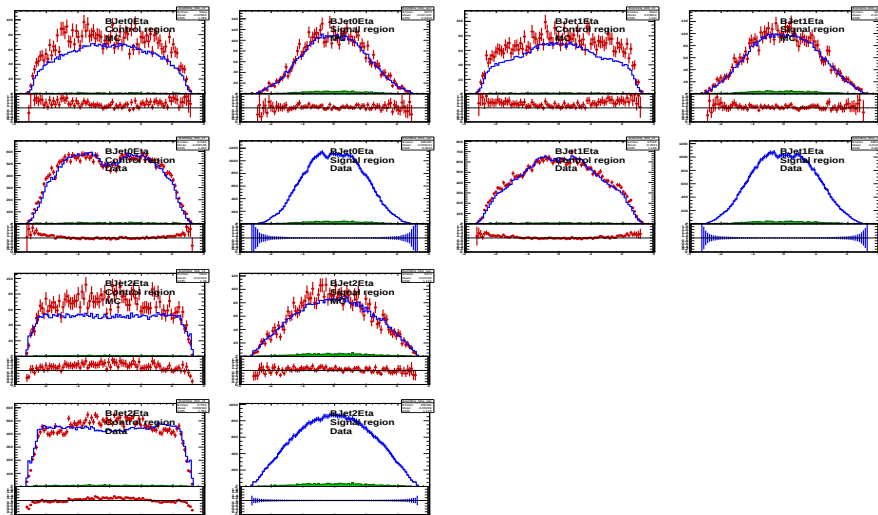
- clear underestimate of prediction for $D < 0.01$ or even $D < 0.001$.
- in DR01 the event excess is clearly visible in the falling edge of the distribution;
- However, excluding these events improve the situation but does not solve the problem

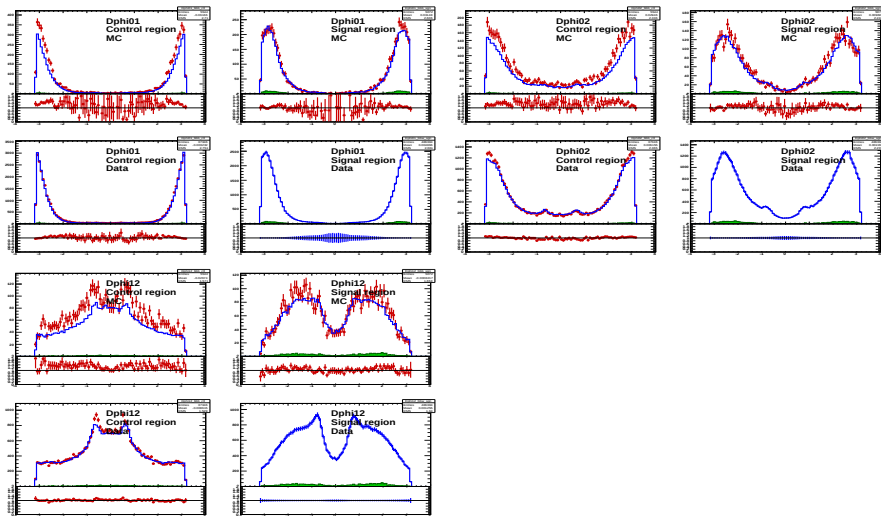
The good(?) news

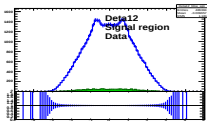
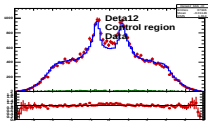
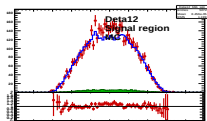
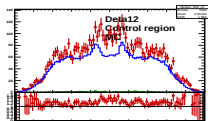
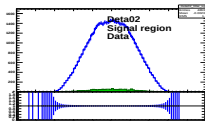
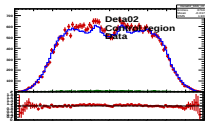
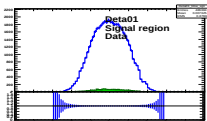
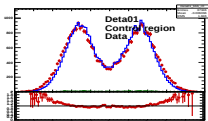
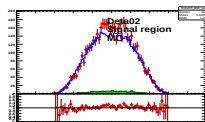
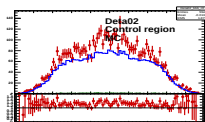
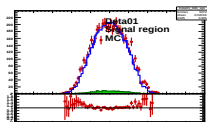
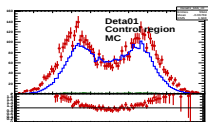
- Using F_x from MC truth, *jet per jet* the agreement is very good, thus ϵ_x are good and well parametrized.
- Using the MC truth, but parametrized, it's the same as using the Fit;
- the probability distribution is clearly different, showing that the problem is maybe in F_x parametrization.

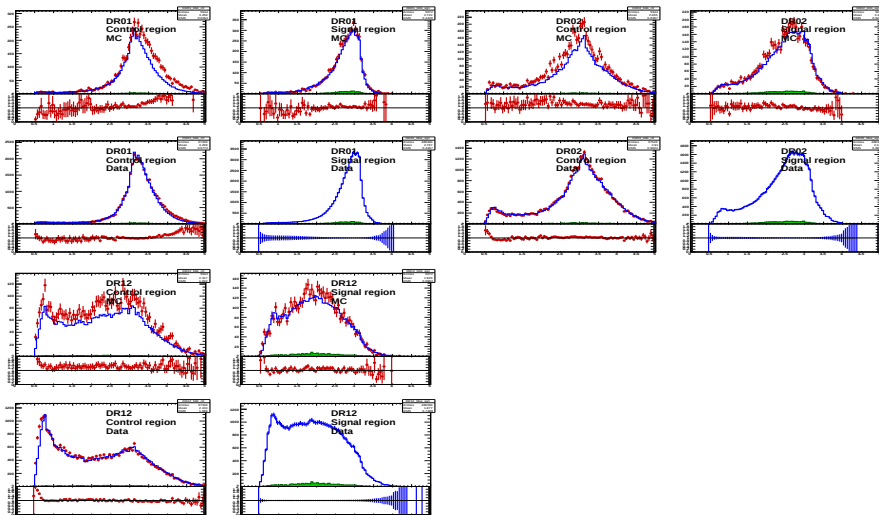
BACKUP

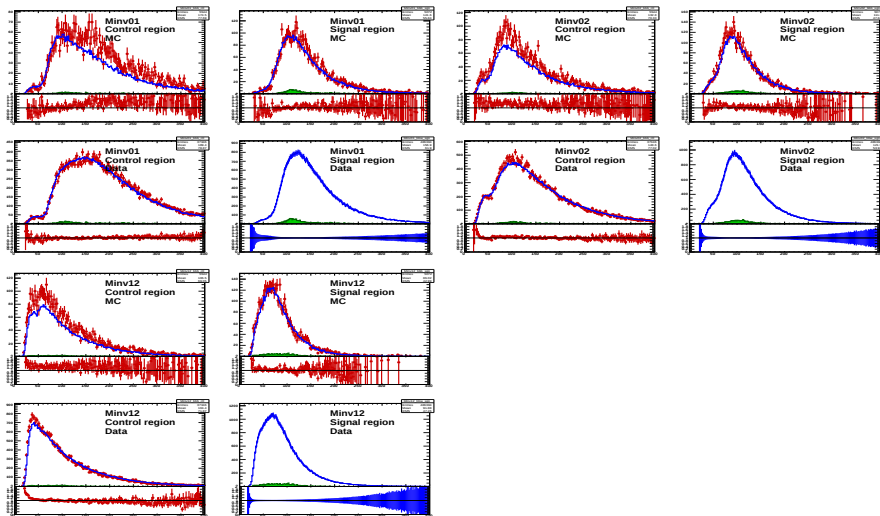


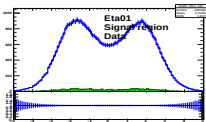
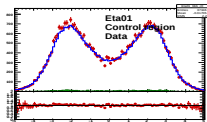
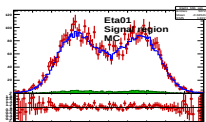
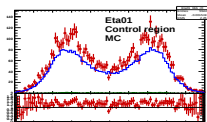
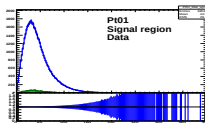
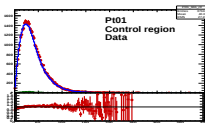
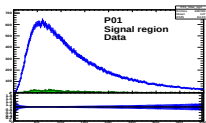
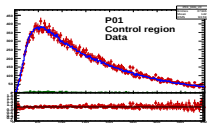
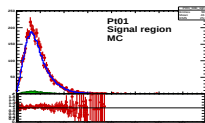
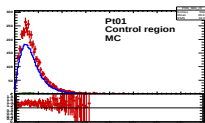
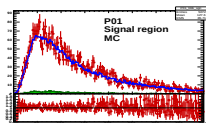
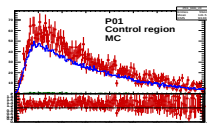






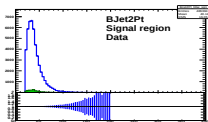
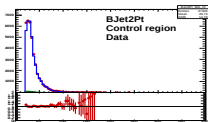
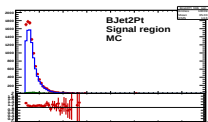
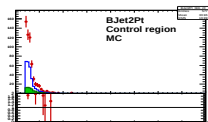
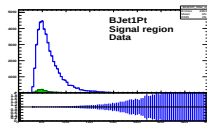
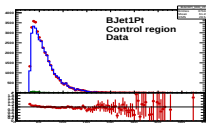
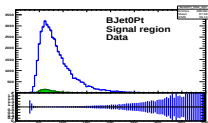
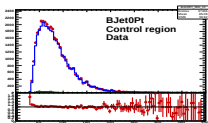
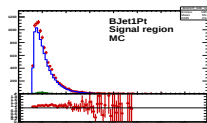
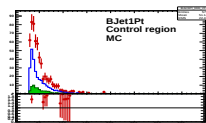
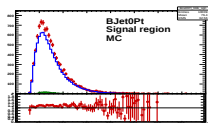
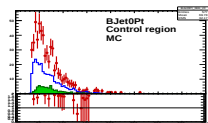


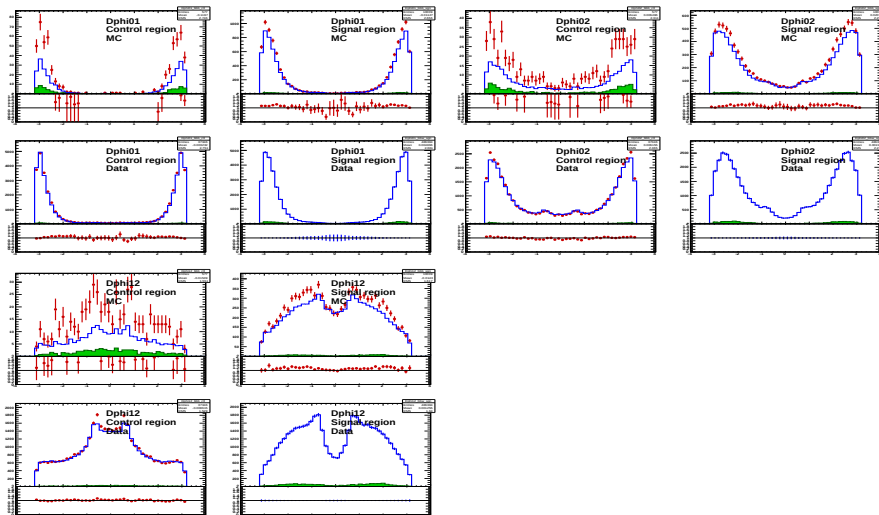


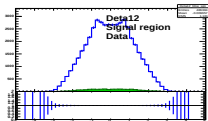
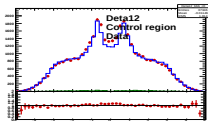
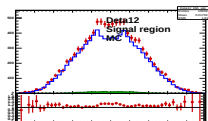
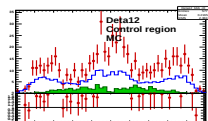
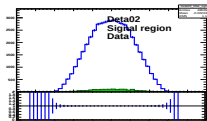
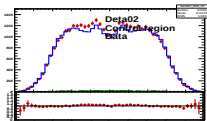
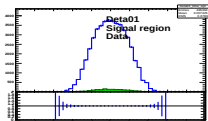
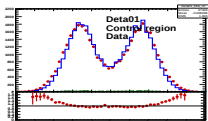
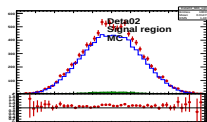
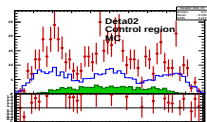
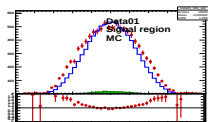
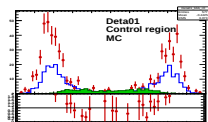


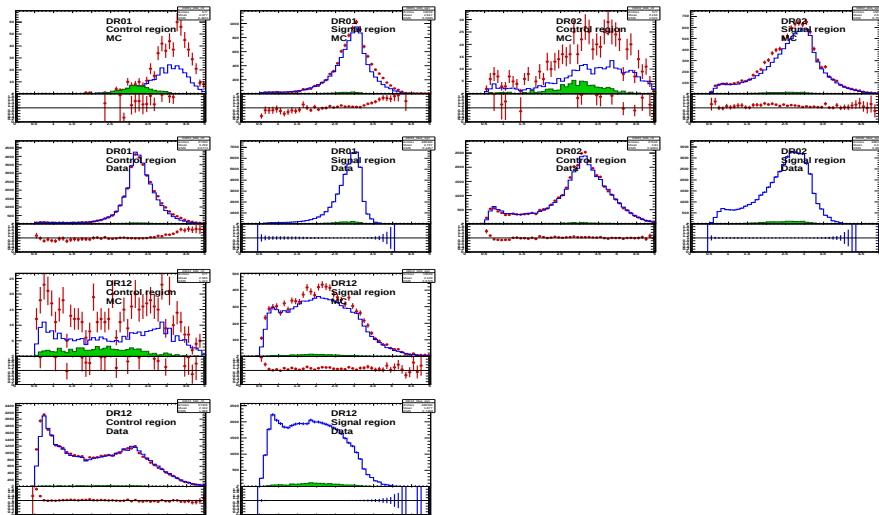
First try

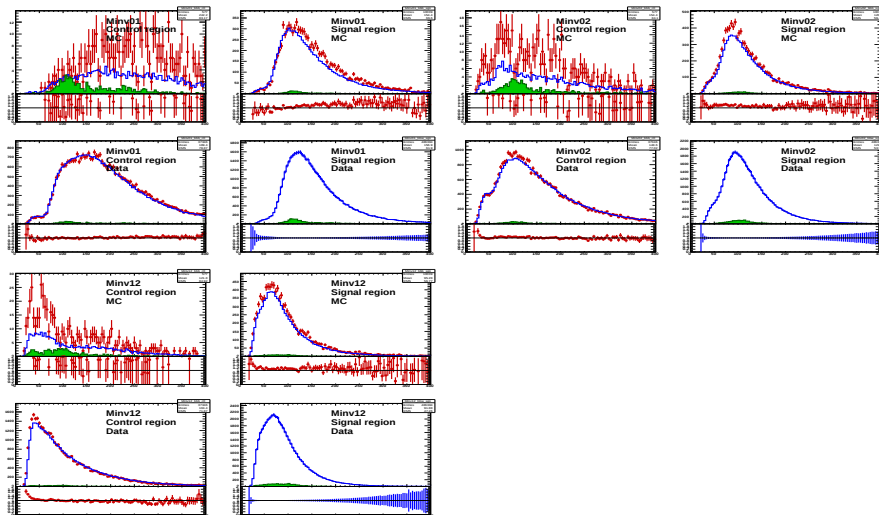
Ok, define (for MC only) control region as $D < 0.001$, so to see the distribution of all variables in the problematic region (control region) and everywhere else (signal region).

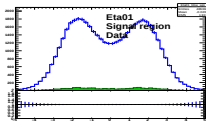
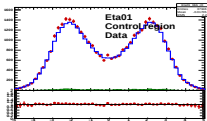
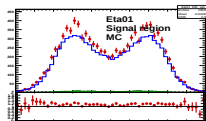
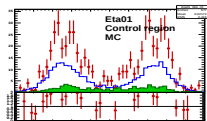
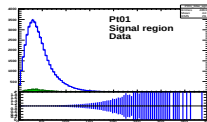
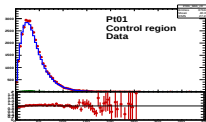
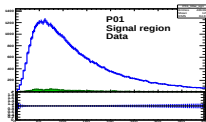
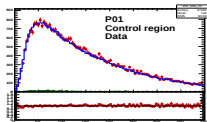
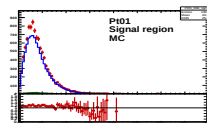
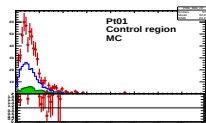
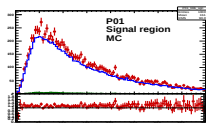
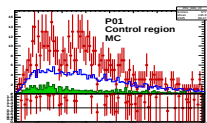












Third try

Ok, define (for MC only) control region as $0.05 < D < 0.04$ and signal region as $0.04 < D < 1$. and see what happen

