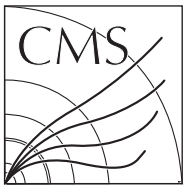
Available on the CMS information server

CMS NOTE AN-14-243



The Compact Muon Solenoid Experiment

CMS Draft Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

2014/12/01
Head Id: 269377
Archive Id: 266572:269470M
Archive Date: 2014/11/27
Archive Tag: trunk

Search for a Higgs boson decaying to invisible final states

Chayanit Asawatangtrakuldee[2], Jim Brooke[3], David Colling[1], Gavin Davies[1], Patrick Dunne[1], Anne-Marie Magnan[1], Alexander Nikitenko[1], and Joao Pela[1]

[1] Imperial College London (UK)
[2] Peking University, Beijing (China)
[3] University of Bristol (UK)

Abstract

In this note, investigations are made into improving the analysis for the search of a Higgs boson produced by Vector-Boson Fusion and decaying to invisible particles, compared to what was published in HIG-13-030 with the 8 TeV dataset, and in view of preparing for 13 TeV. The parked triggers are used instead of the prompt ones, which allow a small increase in statistics due to different requirements at HLT, in particular no requirement on the MET and looser thresholds on the jets pT. An improved cut-based selection is presented, with better rejection of the QCD multijet background. A BDT-based selection is also investigated. Both approaches are optimised in terms of expected 95%CL limits on the branching ratio of Higgs to invisible.

Contents

1 Introduction

The first version of the Higgs to invisible with vector-boson fusion (VBF) production has been published in the CMS paper HIG-13-030, CMS PAS HIG-13-013 and associated analysis notes (AN-2012/403 and AN-2013/205). We refer the reader to these documents for the theoretical motivations behind this measurement.

The published analysis was using the prompt trigger dataset from 2012 and so is referred to as "prompt analysis". A parallel stream of data, the parked data, was also put on disk for the runs B, C and D, and reconstructed only later. This note concentrates on reviewing the analysis strategy using the parked triggers, to improve on the results as well as prepare for the LHC RunII.

The note presents in details the data and MC samples used as well as object reconstructions and preselections in sections 2 to 5. Sections 7 to 10 concentrates on the different background control regions, for Top, W+jets, Z+jets and QCD multijets processes. Results, systematics and limits are given in sections 11 to 13.

The software used is a development of that used for the cross-check analysis to the prompt data VBF Higgs to invisible results presented in HIG-13-013 and HIG-13-030, it has been checked that the framework is able to reproduce the HIG-13-030 result. It has also been checked that the parked data yields in the various regions of the HIG-13-030 analysis are the same as those obtained when using the code which was the "main analysis" for HIG-13-030.

2 Data and MC samples

2.1 Data

The full dataset from 2012 with $\sqrt{s} = 8$TeV has been processed, using the golden JSON files and giving a total integrated luminosity analysed of 19.2 ± 0.5 fb^{-1}. The dataset names and integrated luminosity per dataset are detailed in Table 1.

Table 1: Datasets, JSON and corresponding integrated luminosity analysed.

Dataset/JSON	Int. Lumi [pb^{-1}]
/MET/Run2012A-22Jan2013-v1/AOD	889
/VBF1Parked/Run2012B-22Jan2013-v1/AOD	3871
/VBF1Parked/Run2012C-22Jan2013-v1/AOD	7152
/VBF1Parked/Run2012D-22Jan2013-v1/AOD	7317
Total analysed	19229
Cert_190456-208686_8TeV_22Jan2013ReReco_Collisions12_JSON.txt	19789

2.2 MC

The MC samples used are given in Table 2. The cross sections come mainly from two sources: the PREP CMS database (`http://cms.cern.ch/iCMS/prep/requestmanagement`), and the twiki `https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat8TeV`.

Table 2: MC processes, corresponding cross sections (at NLO or NNLO when available) and equivalent integrated luminosity analysed.

Dataset	σ [pb]	No. of Events	Equivalent $\int L$ [fb^{-1}]
$(Z \to \nu\nu)$ + jets ($50 < \mathrm{HT} < 100$ GeV)	381.2	4040980	10.6
$(Z \to \nu\nu)$ + jets ($100 < \mathrm{HT} < 200$ GeV)	160.3	4416646	27.6
$(Z \to \nu\nu)$ + jets ($200 < \mathrm{HT} < 400$ GeV)	41.49	5055885	122
$(Z \to \nu\nu)$ + jets ($400 < \mathrm{HT} < \infty$ GeV)	5.274	1006928	191
$(W \to l\nu)$ + jets (inclusive)	37509(NNLO)	76102995	2.03
$(W \to l\nu)$ + 1 jet	5400	23141598	42.9
$(W \to l\nu)$ + 2 jet	1750	34044921	19.5
$(W \to l\nu)$ + 3 jet	519	15539503	29.9
$(W \to l\nu)$ + 4 jet	214	13382803	62.5
$(Z/\gamma \to ll)$ + jets (Mll > 50)	3503.71(NNLO)	30459503	8.7
$(Z/\gamma \to ll)$ + 1 jets (Mll > 50)	561	24045248	42.9
$(Z/\gamma \to ll)$ + 2 jets (Mll > 50)	181	21852156	121
$(Z/\gamma \to ll)$ + 3 jets (Mll > 50)	51.1	11015445	216
$(Z/\gamma \to ll)$ + 4 jets (Mll > 50)	23.04	6402827	278
EWK $(Z/\gamma \to ll)$ + 2 jets	0.888	2978717	3354
EWK $(W^+ \to l\nu)$ + 2 jets	6.48	8996164	1388
EWK $(W^- \to l\nu)$ + 2 jets	4.09	5994018	1466
WW	54.838(NLO)	10000431	182
WZ	33.21(NLO)	10000283	301
ZZ	17.654(NLO)	9799908	555
$W\gamma$	461.6	4802358	10.4
tt + jets	245.8(NNLO)	6923750	28.2
t (t-channel)	56.4(NLO)	3758227	66.6
t (tW-channel)	11.1(NLO)	497658	44.8
t (s-channel)	3.79(NLO)	259961	68.6
$\bar{t}$ (t-channel)	30.7(NLO	1935072	63.0
$\bar{t}$ (tW-channel)	11.1(NLO)	493460	44.5
$\bar{t}$ (s-channel)	1.76(NLO)	139974	79.5
QCD ($30 < \mathrm{p\,T} < 50$ GeV)	66285328.0	6000000	0.00009
QCD ($50 < \mathrm{p\,T} < 80$ GeV)	8148778.0	5998860	0.00074
QCD ($80 < \mathrm{p\,T} < 120$ GeV)	1033680.0	5994864	0.0058
QCD ($120 < \mathrm{p\,T} < 170$ GeV)	156293.3	5985732	0.038
QCD ($170 < \mathrm{p\,T} < 300$ GeV)	34138.15	5814398	0.170
QCD ($300 < \mathrm{p\,T} < 470$ GeV)	1759.549	5978500	3.40
QCD ($470 < \mathrm{p\,T} < 600$ GeV)	113.8791	3964848	34.8
QCD ($600 < \mathrm{p\,T} < 800$ GeV)	26.9921	3996864	148
QCD ($800 < \mathrm{p\,T} < 1000$ GeV)	3.550036	3998563	1130
QCD ($1000 < \mathrm{p\,T} < 1400$ GeV)	0.737844	964088	1310
QCD ($1400 < \mathrm{p\,T} < 1800$ GeV)	0.03352235	2000062	60000
QCD ($1800 < \mathrm{p\,T} < \infty$ GeV)	0.001829005	977586	534000

Note that the equivalent integrated luminosity differs slightly from that used in AN-2012/403 for some samples due to grid job failures in one or the other analysis. A list of the full names of the datasets can be found in appendix A of AN-2012/403.

The different jet multiplicity samples for W+jets and DY+jets are combined such that there is no double counting, by applying a weight per parton multiplicity w_i such that the total number of reweighted events can then be normalised to the luminosity in the data using the number of

events in the inclusive sample (N_{incl}) and the NNLO inclusive cross section XS_{incl}^{NNLO}.

$$w_i = \frac{N_{incl} \times f_i}{N_{incl} \times f_i + N_i}, \tag{1}$$

$$f_i = \frac{XS_i^{LO}}{XS_{incl}^{NNLO}} \tag{2}$$

The number of events N_i and LO cross section from PREP XS_i^{LO} are used for the individual i = 1-,2-,3- and 4-jet samples, given in Table 2.

W events are separated into decays to electron, muon or tau using generator-level information. Events from tau decaying to an electron (muon) are counted as "W$\rightarrow$e+ν" ("W$\rightarrow \mu + \nu$"). Only hadronic decays enter the "W$\rightarrow \tau + \nu$" category.

3 Objects definition

The final state contains two jets with the VBF topology, that is with a large rapidity-gap and in the full η coverage accessible (including the forward calorimeters), and large missing transverse energy from the invisible Higgs particle decay products. To reduce backgrounds, any event with "veto" electrons or loose muons in the signal region are removed. Tight electrons and muons are used in the estimations of the W and Z backgrounds. Also hadronic taus are of interest to estimate the contribution from the W decaying to taus.

Most of the objects use the Particle Flow reconstruction algorithm.

3.1 Electrons and muons

3.1.1 Veto leptons

"Veto electrons" are defined using the EGamma POG's cut based electron ID at the Veto working point with the additional requirements that,

$$p_T > 10\,\text{GeV},\ |\eta| < 2.4, \tag{3}$$

$$\text{effective-area-corrected isolation} < 0.15, \tag{4}$$

$$d_{xy} < 0.04\,\text{cm},\ d_z < 0.2\,\text{cm}. \tag{5}$$

"Loose muons" are muons as defined by the Muon POG with the additional requirements that:

$$p_T > 10\,\text{GeV},\ |\eta| < 2.1, \tag{6}$$

$$\text{relative combined isolation} < 0.2. \tag{7}$$

Contribution from pile-up is subtracted from the isolation using the β corrections.

3.1.2 Tight leptons

"Tight electrons" are defined using the EGamma POG's cut based electron ID with the Tight working point (similar to the 2011 WP70) with the following additional requirements:

$$p_T > 20\,\text{GeV},\ |\eta| < 2.4, \tag{8}$$
$$\text{effective-area-corrected isolation} < 0.1, \tag{9}$$
$$d_{xy} < 0.02\,\text{cm},\ d_z < 0.1\,\text{cm}. \tag{10}$$

"Tight muons" are also muons as defined by the Muon POG but with the following additional requirements:

$$p_T > 20\,\text{GeV},\ |\eta| < 2.1, \tag{11}$$
$$\text{relative combined isolation} < 0.12, \tag{12}$$
$$d_{xy} < 0.045\,\text{cm},\ d_z < 0.2\,\text{cm}. \tag{13}$$

Tight electrons are required to be separated from any loose muon candidate by at least $\Delta R > 0.3$ to clean them from muons.

3.2 Hadronic taus

Tau leptons used in the $W \to \tau_{had}\nu$ background estimation are selected using the following discriminants:

- "decayModeFinding",
- "byTightCombinedIsolationDeltaBetaCorr3Hits",
- "againstMuonTight",
- "againstElectronTight",

and we require that they have:

$$p_T > 20\,\text{GeV}, d_z < 0.2\,\text{cm}\ \&\ |\eta| < 2.3. \tag{14}$$

The expected efficiency and fake rate of the discriminant "byTightCombinedIsolationDeltaBetaCorr3Hits" are shown in Figure 1, provided by the HiggsTauTau CMS analysis group from Imperial College London. The tight working point has an expected efficiency of 0.55 and a fake rate of 0.02(0.03) in the barrel(endcap).

3.3 Jets

The jets used in this analysis are anti-k_T Particle Flow jets, "PFJets", with cone size 0.5 and L1+L2+L3+L2L3Residual corrections applied in data and L1+L2+L3 corrections applied in MC. The global tag used for the Jet Energy Scale is FT53_V21A_AN6::All for data and START53_V27::All for MC.

We also require that the jets pass the loose PFJet ID criteria, detailed in CMS AN_2010_003, which are:

- Neutral Hadron Fraction < 0.99
- Neutral EM Fraction < 0.99
- Number of Constituents > 1

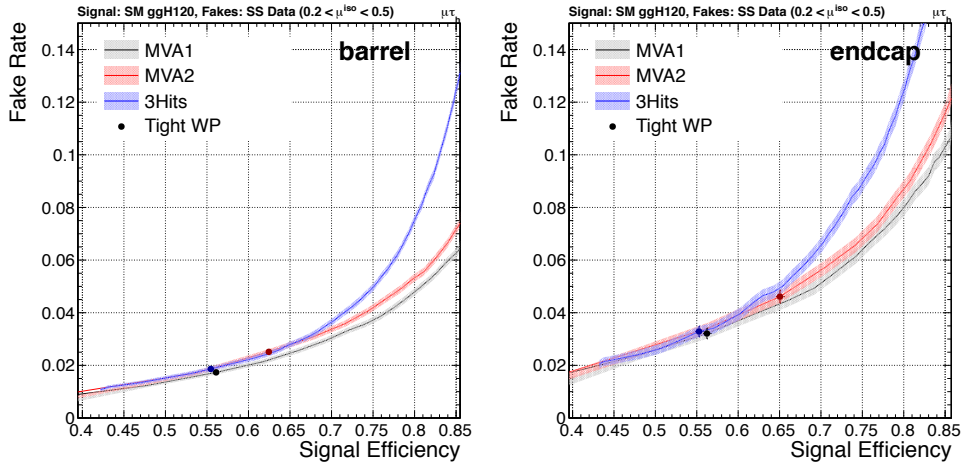


Figure 1: TauID efficiency and fake rate as a function of the discriminant in the barrel (left) and endcap (right) regions, using the CMS Higgs to $\tau\tau$ analysis in the $\mu + \tau_h$ channel.

and for $|\eta| < 2.4$ in addition apply

- Charged Hadron Fraction > 0
- Charged Multiplicity > 0
- Charged EM Fraction < 0.99.

Pileup ID criteria are also applied, using the full BDT method at the loose working point detailed at https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJetID, with the Dec2012 weights and the package recojets/jetproducers with version METPU_5_3_X_v4.

Finally cleaning from leptons has been performed by filtering out jets with a veto electron or a loose muon within $\Delta R < 0.5$.

3.4 MET

Particle Flow MET, "PFMET", has been used with type 0 (charged hadron subtraction for PU mitigation) and type 1 (propagating jet energy corrections) corrections applied. Additionally events are required to pass standard MET filters, this removes events due to anomalous HCAL signals, beam halo, the HCAL and ECAL laser calibration events, ECAL dead cells, bad EE supercrystals, and tracking failures. The proportion of events removed by each filter in the data is detailed in Table 3 for each run: A, B, C and D. The ECAL+HCAL laser filters numbers are the percentage of events removed by these filters after the other filters have already been applied.

The unusually large number of events cut out by the HBHE noise filter, the tracking failure filter and the CSCTightHaloFilter in run A appears to be due to a correlation between passing the trigger and this filter. The percentage of events in run A which pass the trigger but fail the filter is 0.99% for the HBHE noise filter 0.1% for the tracking failure filter and 0.4%for the CSCTightHaloFilter. The total fraction of events which pass the trigger removed by the MET filters in run A is 2.3%. For the ECAL+HCAL laser filters, the percentage is with respects to events passing all the other filters. The filters have not been applied in MC in this analysis.

We also use METnomu which is defined as the result of adding the transverse momentum vectors of any tight muons present in the event to the MET.

Table 3: Percentage of events removed by each filter in the 2012 runs A, B, C and D.

Filter	% rejected			
	Run 2012A	Run 2012B	Run 2012C	Run 2012D
HBHENoiseFilter	22.90	0.19	0.19	0.17
EcalDeadCellTriggerPrimitiveFilter	0.375	0.009	0.010	0.013
eeBadScFilter	0.007744	0	0.000014	0.000036
trackingFailureFilter	3.07	0.00037	0.00742	0.00030
manystripclus53X	0.0018	0.0013	0.0023	0.0013
toomanystripclus53X	0.00048	0.00117	0.00203	0.00114
logErrorTooManyClusters	0	0	0.000014	0.000036
CSCTightHaloFilter	10.3	0.40	0.40	0.51
Total	28.50	0.60	0.60	0.69
ECAL+HCAL laser filters	0.94	0.0087	0.00027	0

4 Data/MC correction factors

In addition to applying event-by-event weights to the Monte Carlo in order to make the Monte Carlo pileup distribution match that in data we also perform reweightings for trigger efficiency, and lepton reconstruction, isolation and identification efficiencies.

4.1 Trigger efficiency

The triggers used in the analysis are:

- HLT_DiPFJet40_PFMETnoMu65_MJJ800VBF_AllJets for run A,
- HLT_DiJet35_MJJ700_AllJets_DEta3p5_VBF for runs B and C, and
- HLT_DiJet30_MJJ700_AllJets_DEta3p5_VBF for run D.

All three of these HLT paths are seeded at level 1 by L1_ETM40 (in the case of the run B, C and D triggers there are also other L1 seeds, but we require that L1_ETM40 fired for all run periods). The turn-on curves of each trigger have been measured as a function of offline PFMETnoMu in bins of M_{jj} and jet 2 p_T, to obtain the variation of efficiency with the different variables involved in the trigger including correlations. The turn on curves in each bin are then fit using the following function:

$$\frac{\epsilon_{max}}{2}\mathrm{Erf}\left(\frac{x-x_0}{\sqrt{\Gamma}}\right)+1, \tag{15}$$

where ϵ_{max} is the maximum efficiency of the trigger in the bin, x_0 is the centre of the turn on and Γ is the width of the turn on. The results of these fits can be seen in Appendix A

Instead of applying the trigger in MC, we apply to each event a luminosity weighted average of the efficiency of the three triggers obtained for an event of that MET, M_{jj} and second jet p_T.

4.2 Lepton identification and isolation

Given we are using identification and isolation working points recommended by the EGamma and Muon POGs, we can use the tag-and-probe efficiencies measured by the POG as a function of the p_T and η of the leptons. For the tight lepton selections, MC events are reweighted using the data/MC scale factors per lepton. For the veto selection, the veto efficiencies are applied, that is the ratio:

$$\frac{1-\varepsilon_{data}}{1-\varepsilon_{MC}}$$

per lepton identified at generator level with the same acceptance, that is $p_T > 10$ GeV and $|\eta| < 2.1$ (2.4) for muons (electrons), but not reconstructed.

5 Selection

5.1 Preselection

We perform a preselection using the objects defined in section 3. The requirement to be above the trigger thresholds in all variables (described in section 4.1) gives the following conditions:

$$\begin{aligned} &\eta_{j1} \cdot \eta_{j2} < 0,\ \eta_{j1,2} < 4.7, \\ &\text{jet 1 } p_T > 50\,\text{GeV},\ \text{jet 2 } p_T > 40\,\text{GeV}, \\ &\Delta\eta_{jj} > 3.6,\ \text{GeV},\ M_{jj} > 800\,\text{GeV}, \\ &\text{METnomu} > 90\,\text{GeV}. \end{aligned} \tag{16}$$

In addition to this selection unless stated otherwise (e.g. in the W+jets and Z+jets control regions) we apply a lepton veto, requiring that there are no veto electrons or muons. We do not veto tau leptons as the identification efficiency (described in section 3.2) is too low for it to be useful in removing a significant amount of background events.

After applying this loose preselection, the QCD multijet process is the dominant background. The MET in QCD multijets is mostly coming from mismeasured jets, we therefore additionally require the following:

$$\text{min}\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{j}) > 1.0,\ \frac{METnomu}{\sigma_{METnomu}} > 3.0,\ M_{jj} > 1000\,\text{GeV}, \tag{17}$$

where $\text{min}\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{j})$ is the minimum azimuthal angle between any jet with $p_T > 30$ GeV and the MET ignoring muons.

Figure 2 left shows the significance of the MET with muons ignored, "MET significance", after the preselection. It is clear that the QCD contribution becomes much lower above 3. More discussion on the QCD is given in section 10. The QCD contribution in pink in all the figures shown is just for illustration of what the MC predicts out-of-the-box when requiring significant generator-level MET in the events (see section 10). The other backgrounds are also normalised in their control regions as explained later in sections 7 to 9. Figure 2 right shows $\text{min}\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{j})$, after requiring the MET significance to be greater than 3. Again it is clear that most QCD events have a high p_T jet close to the MET in ϕ. Finally in Figure 3 the dijet mass distribution is shown after the selection on both MET significance and $\text{min}\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{j})$ has been applied.

These three cuts along with the trigger driven selection define our "preselection" which is in summary:

- $\eta_{j1} \cdot \eta_{j2} < 0$, $\eta_{j1,2} < 4.7$, $\Delta\eta_{jj} > 3.6$
- jet 1 $p_T > 50\,\text{GeV}$, jet 2 $p_T > 40\,\text{GeV}$, $M_{jj} > 1000\,\text{GeV}$
- METnomu $> 90\,\text{GeV}$, $\frac{METnomu}{\sigma_{METnomu}} > 3.0$,
- $\text{min}\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{j}) > 1.0$.

A selection of control plots after the preselection are shown in Figure 4.

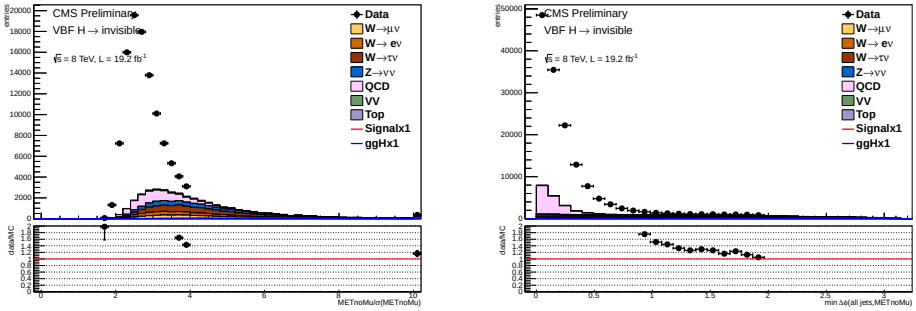


Figure 2: left: the MET significance after the trigger driven selection described in equation 16. right: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ after the trigger driven selection and requiring the MET significance to be greater than 3

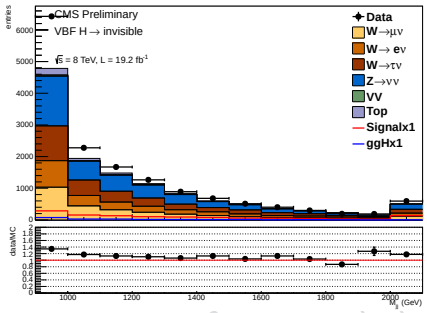


Figure 3: the dijet mass after the trigger driven loose preselection, requiring the MET significance to be greater than 3 and $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ to be greater than 1

5.2 Cut-based analysis selection

As can be seen in Figure 4 when we do not model QCD there is a significant difference between data and our background estimation. For this reason we chose to search for a region where QCD is negligible to use as the signal region for our cut based analysis. This means that we can accept a QCD contribution estimation with much larger uncertainties, whilst still not worsening our expected limit significantly.

In Figure 2 the ratio data/MC flattens for MET significance above 4 and $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ above 2, where the QCD contribution is expected to become negligible. Our estimate of the QCD contribution in this region also supports the fact that QCD can be neglected (see section 10).

Within this tightened region, we performed an optimisation of our cuts, using the 95% confidence level expected limit calculated as described in section 13 as our figure of merit. We varied our cuts on the MET significance, $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$, sub-leading jet p_T and dijet mass cuts, and the optimum expected limit that we obtained was for the following cuts:

- $\eta_{j1}\cdot\eta_{j2} < 0$, $\eta_{j1,2} < 4.7$, $\Delta\eta_{jj} > 3.6$
- jet 1 $p_T > 50\,\mathrm{GeV}$, jet 2 $p_T > 45\,\mathrm{GeV}$, $M_{jj} > 1200\,\mathrm{GeV}$
- METnomu $> 90\,\mathrm{GeV}$, $\frac{METnomu}{\sigma_{METnomu}} > 4.0$,
- $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) > 2.3$.

The region with these cuts applied will be referred to as the signal region.

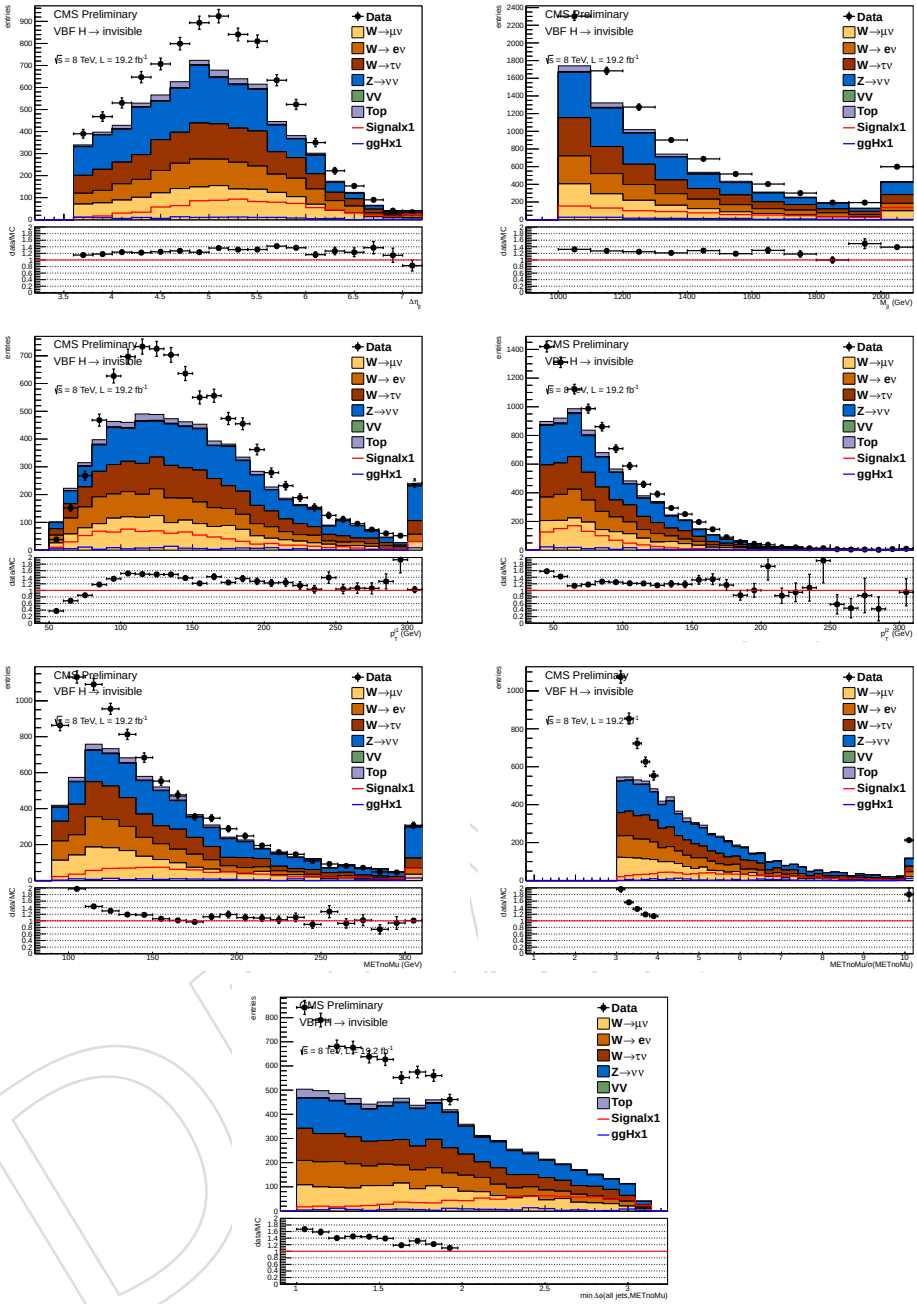


Figure 4: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum and METnomu after the preselection. It should be noted that there is no estimation of QCD in these histograms, we believe this accounts for the difference between the data and Monte Carlo. Furthermore the MET significance and $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ plots have been blinded in the region that we use as the signal region for our cut-based analysis

5.3 MVA study

Using the above defined signal region as a starting point we investigated whether we could improve our expected limit by using an MVA. The reaason for starting from the signal region is that we needed to pick a region where QCD is negligible, as otherwise the uncertainty on our QCD estimation technique is not acceptible. We used TMVA and trained both a BDT and a

Fisher discriminant using the variables showing the most shape difference between signal and background and the biggest differences in correlations between signal and background. We then reran the optimisation procedure above with the addition of a cut on the value of the MVA discriminant. The best improvement in expected 95% C.L. limit on invisible branching fraction was less than 1% so we decided not to use the MVA for our final selection as once systematics accounting for our understanding of the MVA were added we believe this improvement would be removed.

6 General method for extracting data-driven estimates of the main backgrounds

Due to the tight cuts of our analysis even from preselection level, we use data driven methods to estimate the contribution from our significant background processes. The general method used for all of these data driven estimates is to identify a control region which is dominated by the background process to extract a normalisation factor from, then to apply that factor to the Monte Carlo describing that background process in other regions. This can be expressed using the following formula for a process "A":

$$N_S^A = N_S^{A\,MC} \frac{N_C^{Data} - N_C^{bkg}}{N_C^{A\,MC}}, \tag{18}$$

where

- N_S^A is the estimation of the number of events from process A in the signal region.
- $N_{S/C}^{A\,MC}$ is the number of events predicted by process A Monte Carlo in the signal/control region.
- N_C^{Data} is the number of data events in the control region.
- N_C^{Bkg} is the number of events from other background processes in the control region.

The assumption of this method is that if the shape of the background contribution from Monte Carlo describes well the data in the control region for all the important kinematic variables, then we can trust that it will be also the case in the signal region and use the same normalisation factor. We will hence systematically check the shape agreement in the different control regions.

7 Top background estimation

Due to their decay to W bosons and b quarks, events with top quarks often have both:

- sufficient MET from the decay of the b quarks and/or W bosons,
- hard jets from either the b quarks or hadronic W bosons.

In cases where there are either no leptons, or the leptons are not reconstructed this can lead to top quarks contributing to our signal region. More significantly, as will be described later in section 8, if the leptons from the top decay are reconstructed this can lead to a top quark contribution to the W control regions of this analysis as well.

We use the method described in section 6 and equation 18 to obtain a data driven normalisation for this background. The control region that is chosen is the signal region with the lepton veto replaced with a requirement that there is one tight electron and one tight muon as defined in section 3. This region is chosen because top quark pair production where both resulting W

bosons decay leptonically can result in the production of two leptons of different flavour. Other processes leading to the same final state are from diboson production, estimated from Monte Carlo.

Unfortunately this region has only small numbers of data events, for this reason we loosen our cut on $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ to zero. As can be seen from Figure 5, we have very few events with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ above 2.3, however the data to Monte Carlo ratio we obtain in our control region remains unchanged to within statistical errors as the cut is loosened from 1 to 0, and as can also be seen from Figure 5 there is no significant trend in the data/MC ratio with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$.

To summarise the top control region is the signal region with the following modifications:

- the lepton veto is replaced with a requirement that there is one tight electron and one tight muon and no other veto leptons,
- the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ cut is removed.

Within the statistics available, a good agreement between data and MC is observed for all the variables studied. The extrapolation to the signal region is shown in Table 4.

Table 4: top estimate in the signal region using the control selection and the expectations from MC in both regions. The uncertainty given is only statistical.

	Signal region	Control region
N^{data}	XXX	21 ± 4.6(stat.)
N^{bkg}	N/A	0.3 ± 0.1(stat.)
N^{MC}	5.3 ± 1.3(stat.)	24.6 ± 4.0(stat.)
$\frac{N^{data}-N^{bkg}}{N_C^{MC}}$	0.84 ± 0.19(stat.)$\pm$0.14(MC stat.)	
Final estimate	4.4 ± 1.0(stat.)$\pm$1.3(MC stat.)	N/A

8 W background estimation

Whilst hadronically decaying W bosons do not often have sufficient MET and thus fail our selection, leptonically decaying W bosons often have both:

- sufficient MET to pass our selection due to the neutrino from the decay,
- the visible lepton not being reconstructed or falling outside the detector acceptance or definition of a veto lepton and thus not being removed by the lepton veto.

When the W bosons are produced in association with jets, this leads to events which pass our selection.

We use the method described in section 6 and equation 18 to obtain a data driven normalisation for this background. Monte Carlo estimates are taken from a combination of the samples generated with QCD production and electroweak production.

In the case where the W decays to an electron/muon (directly or through the leptonic decays of tau leptons) the control region that is chosen is that where the lepton veto is replaced with a requirement that there is only one tight electron/muon and no other veto or tight leptons as described in section 3.

For the W decays to hadronic taus, the control region that is chosen is that where we require a reconstructed hadronic tau in addition to the signal region conditions. As we do not veto

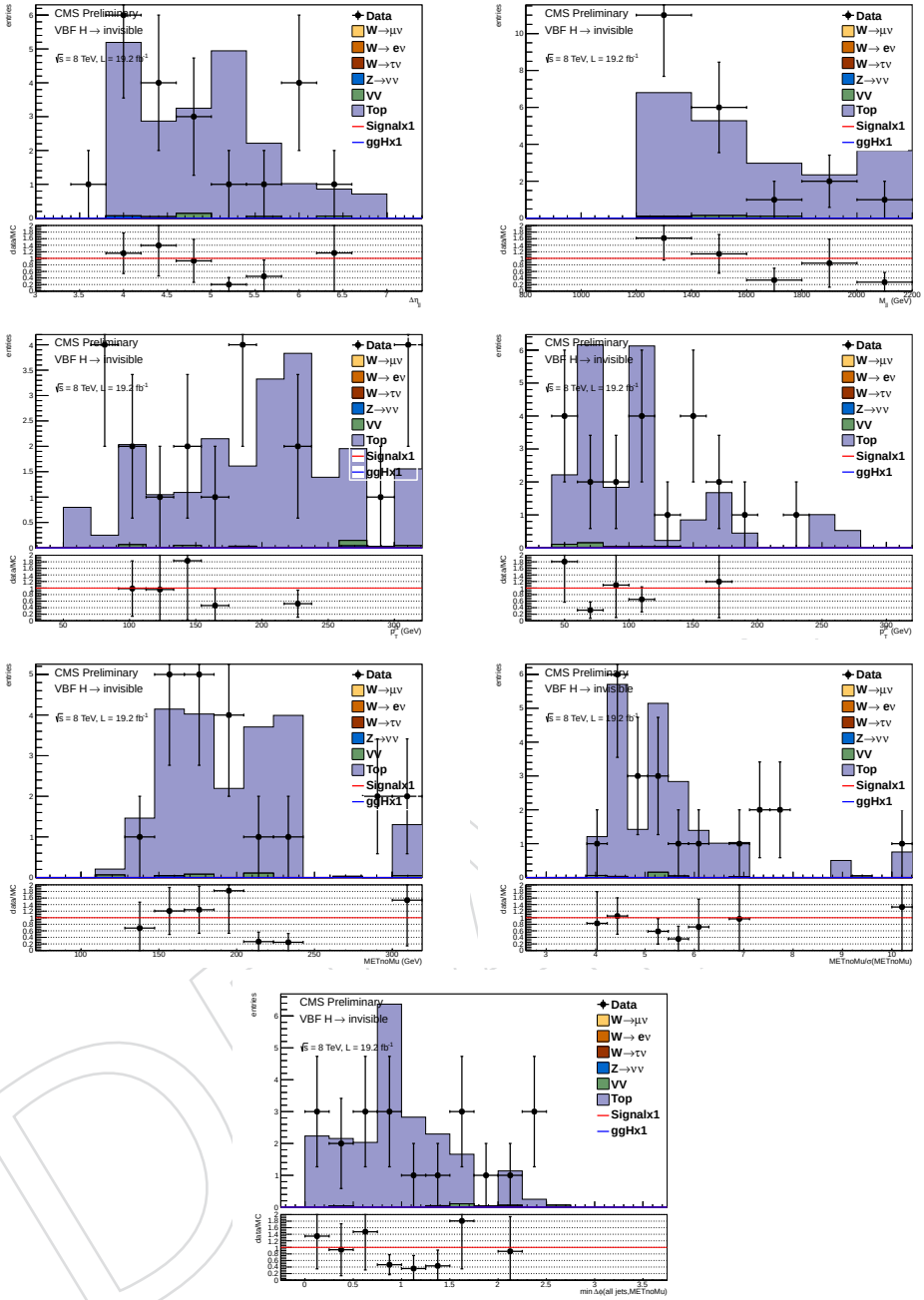


Figure 5: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum and METnomu in the top control region.

events with hadronically-decaying tau leptons in the signal region, due to the relatively high fake rate and low efficiency of tau-tagging algorithms, this control region is in fact a subset of the signal region.

However, as can be seen in section 8.3 the number of data events in that region was not sufficient to estimate the W to tau contribution, so the actual region used only has a small overlap with the signal region.

The dominant contribution to N_C^{Bkg} for all three W control regions is top quark production, this is because the decay of top quarks almost always produces genuine W bosons. For example in the tau control region it is estimated to account for 14% of the data events in the region. For this reason we use the data driven method described in section 7 to estimate the top quark contribution to each control region and propagate the error on the data driven weighting of the top contribution through to the error on the W background estimation. We also consider diboson production, using estimates from Monte Carlo.

8.1 $W \to e\nu$

As described above the $W \to e\nu$ control region is that where all of the signal region conditions are required except that the lepton veto is replaced with the requirement that there is one tight electron and no other veto leptons. The distributions of the variables we use in the control region are shown in Figure 6. Within the statistics available, a good agreement between data and MC is observed for all the variables studied.

The inputs to equation 18 and the final estimation of the $W \to e\nu$ background are shown in Table 5.

Table 5: $W \to e\nu$ estimate in the signal region using the control selection and the expectations from MC in both regions. The uncertainty given is only statistical.

	Signal region	Control region
N^{data}	XXX	68 ± 8.2(stat.)
N^{bkg}	N/A	3.1 ± 1.5(stat.)
N^{WMC}	114.6 ± 8.9(stat.)	129.6 ± 8.1(stat.)
$\frac{N^{data}-N^{bkg}}{N_C^{WMC}}$	0.50 ± 0.06(stat.)± 0.03(MC stat.)	
Final estimate	57.4 ± 7.3(stat.)± 5.9(MC stat.)	N/A

8.2 $W \to \mu\nu$

As described above the $W \to \mu\nu$ control region is that where all of the signal region conditions are required except that the lepton veto is replaced with the requirement that there is one tight muon and no other veto leptons. The distributions of the variables we use in the control region are shown in Figure 7. Again, within the statistics available, a good agreement between data and MC is observed for all the variables studied. The inputs to equation 18 and the final estimation of the $W \to \mu\nu$ background are shown in Table 6.

Table 6: $W \to \mu\nu$ estimate in the signal region using the control selection and the expectations from MC in both regions. The uncertainty given is only statistical.

	Signal region	Control region
N^{data}	XXX	300 ± 17.3(stat.)
N^{bkg}	N/A	12.7 ± 4.6(stat.)
N^{WMC}	142.1 ± 10.1(stat.)	401.1 ± 15.1(stat.)
$\frac{N^{data}-N^{bkg}}{N_C^{MC}}$	0.72 ± 0.04(stat.)± 0.03(MC stat.)	
Final estimate	101.8 ± 6.1(stat.)± 8.3(MC stat.)	N/A

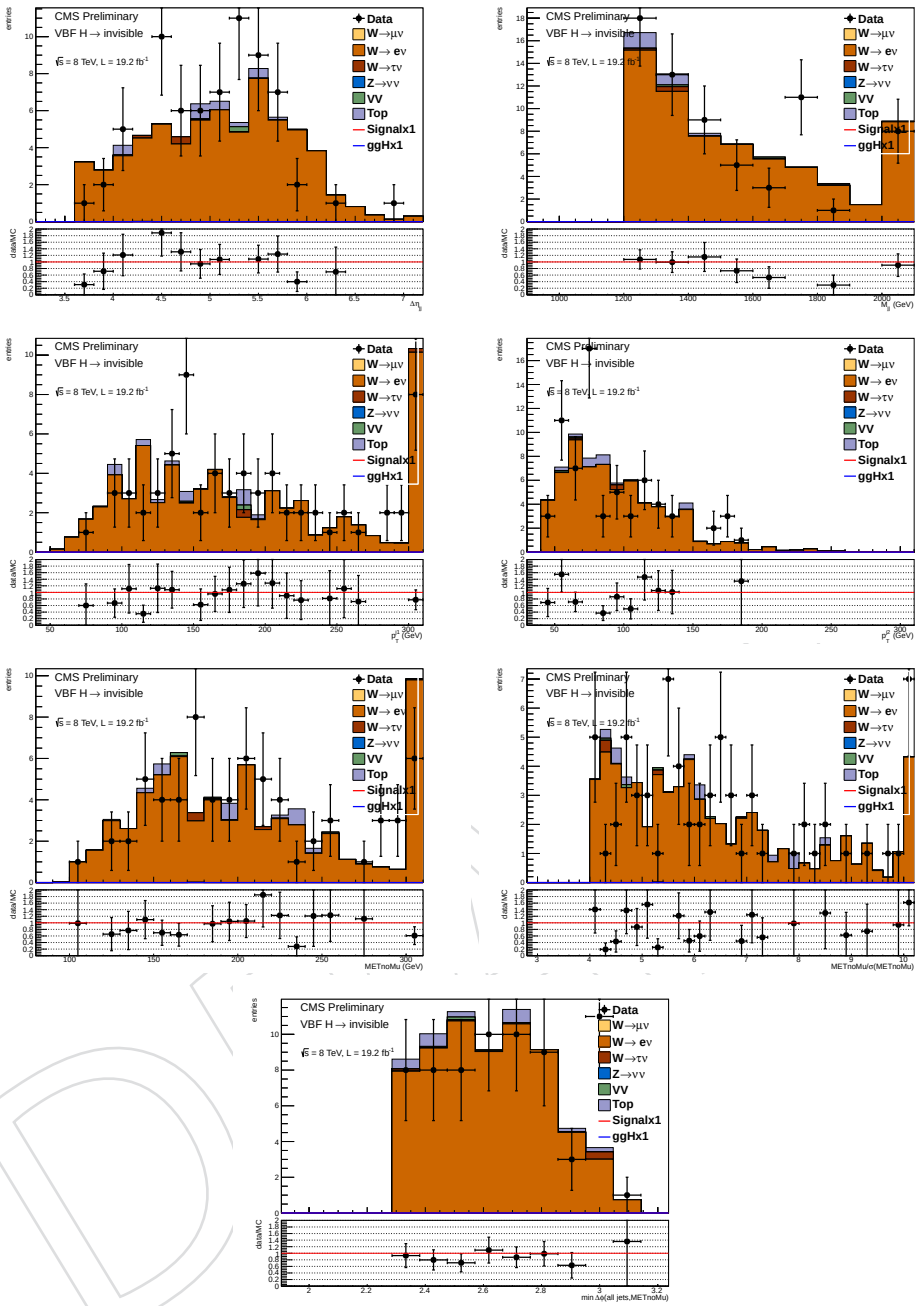


Figure 6: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum and METnomu in the $W \to e\nu$ control region.

8.3 $W \to \tau\nu$

If we ask for an additional hadronic tau in the signal region, we are left with a handful of data events. In order to obtain a data driven estimate of the $W \to \tau\nu$ background we must therefore loosen our conditions on this control region. We first remove the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ cut, however this leads to QCD contamination for the lower values of $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ cut as the requirement of one hadronic tau, unlike the requirement of two tight leptons in the top control region, does not render QCD negligible. In order to reduce the remaining QCD we place a cut on the transverse mass of the hadronic tau and MET system, m_T requiring that it is greater than

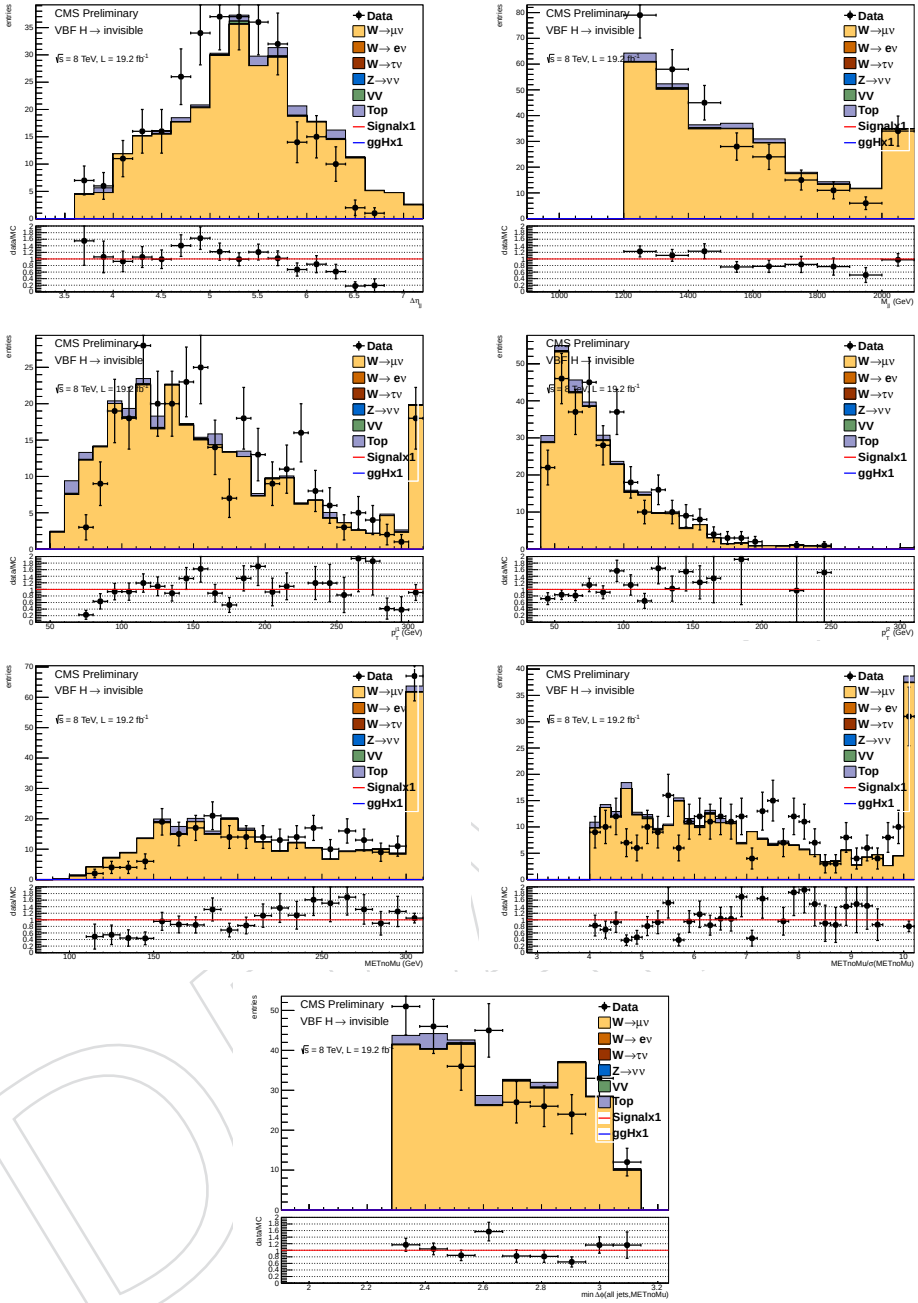


Figure 7: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum and METnomu in the $W \to \mu\nu$ control region.

20 GeV. We also place a cut on $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j1/j2})$, the minimum azimuthal angle separation between the two leading jets (p_T ordered) and the MET ignoring muons, requiring that it be greater than 1; this variable differs from $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ in that only the two leading jets are considered rather than all jets with $p_T > 30$ GeV.

In order to gauge the effect that these differences in cuts have on our estimation of the $W \to \tau\nu$ background we study the effect of making similar changes in the $W \to \mu\nu$ region where we do have sufficient data events. We observe that when the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ cut is loosened to 1 there is a 20% change in the data driven weight obtained. We therefore place a 20% systematic on

our $W \to \tau\nu$ background estimation to account for the differing cuts between the signal and the control region.

To summarise the $W \to \tau\nu$ control region is the same as the signal region except with the following modifications:

- In addition to the lepton vetos we require one hadronic tau in the event,
- We remove the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ cut
- We require $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j1/j2}) > 1.0$
- We require the transverse mass of the hadronic tau MET system to be greater than 20 GeV

The distributions of the variables we use in the control region are shown in Figure 8. Again, within the statistics available, a good agreement between data and MC is observed for all the variables studied. The inputs to equation 18 and the final estimation of the $W \to \tau\nu$ background are shown in Table 7.

Table 7: $W \to \tau\nu$ estimate in the signal region using the control selection and the expectations from MC in both regions. The uncertainty given is only statistical.

	Signal region	Control region
N^{data}	XXX	76 ± 8.7(stat.)
N^{bkg}	N/A	11.3 ± 4.6(stat.)
N^{WMC}	122.6 ± 8.8(stat.)	81.0 ± 6.4(stat.)
$\frac{N^{data}-N^{bkg}}{N_C^{MC}}$	0.80 ± 0.11(stat.)$\pm$0.08(MC stat.)	
Final estimate	98.0 ± 13.2(stat.)$\pm$12.6(MC stat.)	N/A

9 Z background estimation

The decay of Z bosons to neutrinos provides genuine MET, so the production of Z bosons in association with jets, which satisfy our other signal region requirements, provides an irreducible background process. The estimation technique used for this background differs slightly from that used in sections 7 and 8. This is because the control region chosen is that where we replace our lepton veto with a requirement that there are two tight muons consistent with the Z mass hypothesis (i.e. $60 < M_{\mu\mu} < 120\,\mathrm{GeV}$) and no other leptons and we must therefore take into account the fact that the cross-section for the decay of a Z boson to neutrinos is not the same as that for the decay to muons. The equation used is therefore:

$$N_S^{Z\to\nu\nu} = \left(N_C^{Data} - N_C^{bkg}\right) \cdot \frac{\sigma\left(Z \to \nu\nu\right)}{\sigma\left(Z \to \mu\mu\right)} \cdot \frac{\epsilon_S^{ZMC}}{\epsilon_C^{ZMC}}, \tag{19}$$

where:

- $N_S^{Z\to\nu\nu}$ is the expected $Z \to \nu\nu$ contribution in the signal region
- N_C^{Data} is the number of data events in the control region.
- N_C^{Bkg} is the number of events from other background processes in the control region. We consider diboson, top quark, and W+jets backgrounds, using the respective data-driven normalisations for the latter two processes.
- $\sigma\left(Z \to \nu\nu\right)$ is the sum of the electroweak and QCD contributions to the $Z \to \nu\nu$ process, which is 6602 pb. The origin of this cross-section is discussed below.

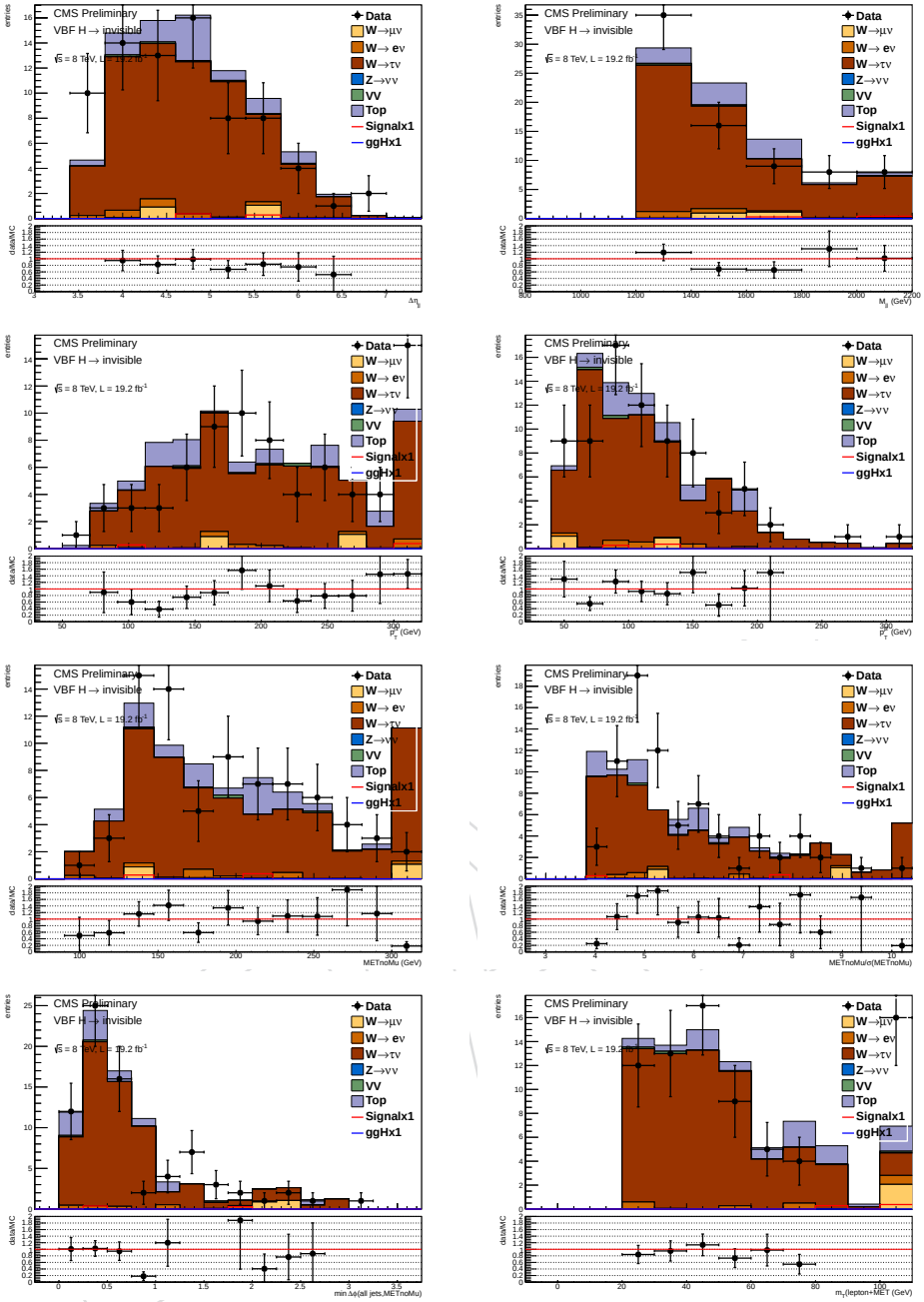


Figure 8: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum and METnomu in the $W \to \tau\nu$ control region.

- $\sigma\left(Z/\gamma^* \to \mu\mu\right)$ is the sum of the electroweak and QCD contributions to the $Z/\gamma^* \to \mu\mu$ process, which is 1168 pb, the origin of this cross-section is discussed below.
- $\epsilon_{S/C}^{ZMC}$ are the the efficiencies for passing our signal/control region cuts in the Z to muons Monte Carlo. Z boson production occurs via both QCD and electroweak processes, and both of these must be taken into account. We therefore use the following

formula to appropriately weight the two production mechanisms:

$$\epsilon_S^{ZMC} = \frac{\sigma\left(Z \to \nu\nu, EWK\right) \frac{N_S^{MC}(EWK)}{N_{Gen}(Z_{mass},EWK)} + \sigma\left(Z \to \nu\nu, QCD\right) \frac{N_S^{MC}(QCD)}{N_{Gen}(Z_{mass},QCD)}}{\sigma\left(Z \to \nu\nu, EWK\right) + \sigma\left(Z \to \nu\nu, QCD\right)}, \quad (20)$$

$$\epsilon_C^{ZMC} = \frac{\sigma\left(Z/\gamma^* \to \mu\mu, EWK\right) \frac{N_C^{MC}(EWK)}{N_{Gen}(EWK)} + \sigma\left(Z/\gamma^* \to \mu\mu, QCD\right) \frac{N_C^{MC}(QCD)}{N_{Gen}(QCD)}}{\sigma\left(Z/\gamma^* \to \mu\mu, EWK\right) + \sigma\left(Z/\gamma^* \to \mu\mu, QCD\right)}, \quad (21)$$

where

- $\sigma\left(Z/\gamma^* \to \mu\mu, EWK\right)$ is the electroweak production cross-section in the phase space of our Z+jets Monte Carlo samples for a Z/γ^* decaying to two muons. We use 303 fb which is the NLO cross-section from VBFNLO
- $\sigma\left(Z/\gamma^* \to \mu\mu, QCD\right)$ is the QCD production cross-section in the phase space of our Z+jets Monte Carlo samples for a Z/γ^* decaying to two muons. We use 1168 pb which is one third of the NNLO cross-section for $Z \to \ell\ell$.
- $\sigma\left(Z \to \nu\nu, EWK\right)$ is the electroweak production cross-section for a Z boson decaying to two neutrinos. We use 1380 fb which is the NLO cross-section from VBFNLO
- $\sigma\left(Z \to \nu\nu\, QCD\right)$ is the QCD production cross-section for a Z boson decaying to two neutrinos. We use 6600 pb which is $\sigma\left(Z/\gamma^* \to \mu\mu, QCD\right)$ multiplied by the ratio of the $Z \to \nu\nu$ and $Z/\gamma^* \to \mu\mu$ cross-sections obtained from MCFM at NLO, which we find to be 5.651. Comparing this ratio to its electroweak equivalent it can be seen that they are not the same. We believe this difference is due to the presence of multi-peripheral diagrams in electroweak production.
- $N_S^{MC}(QCD/EWK)$ are the numbers of events in the signal region in the QCD/EWK produced Z+jets Monte Carlo. Given that we do not have sufficient statistics in our $Z \to \nu\nu$ Monte Carlo we use $Z \to \ell\ell$ Monte Carlo using the met recalculated without muons and without vetoing muons. We also place a generator level cut on the Z mass of $80 < M_Z^{Gen} < 100\,\text{GeV}$ to reduce the effects of photon interference.
- $N_C^{MC}(QCD/EWK)$ are the numbers of events in control region in the QCD/EWK produced Z+jets Monte Carlo.
- $N_{Gen}(QCD/EWK)$ is the number of events in the QCD/EWK produced Z+jets Monte Carlo at generator level.
- $N_{Gen}(Z_{mass}, QCD/EWK)$ is the number of events in the QCD/EWK produced Z+jets Monte Carlo at generator level after a generator level cut on the Z mass has been made such that $80 < m_Z^{Gen} < 100\,\text{GeV}$.

The distributions of the variables we use in the control region are shown in Figure 9. Again, within the statistics available, a good agreement between data and MC is observed for all the variables studied. The inputs to equations 19, 20 and 21 and the final estimation of the $Z \to \nu\nu$ background are shown in Table 8.

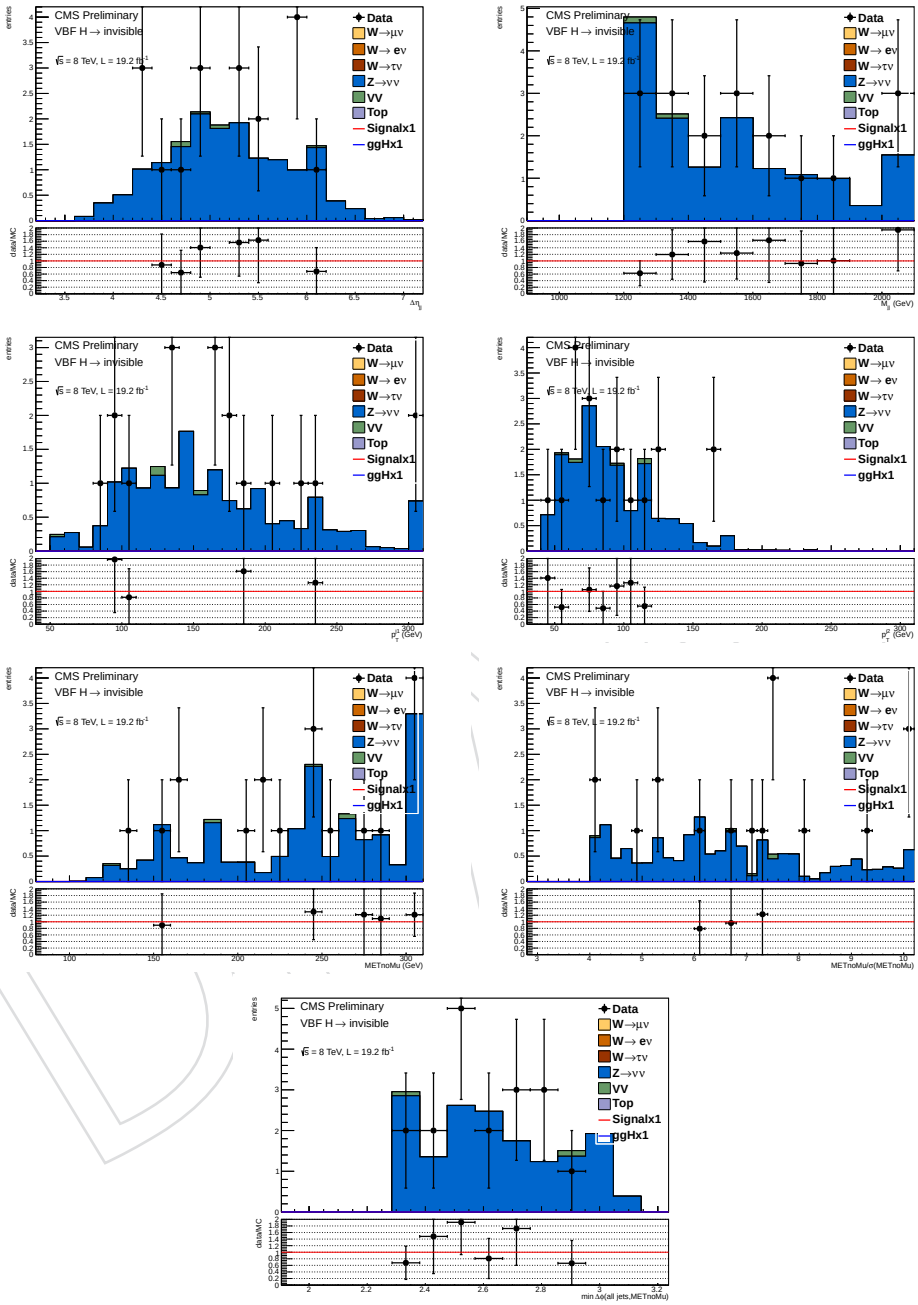


Figure 9: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum and METnomu in the $Z \to \mu\mu$ control region.

Table 8: $Z \to \mu\mu$ estimate in the signal region using the control selection and the expectations from MC in both regions. The uncertainty given is only statistical.

N_{Gen}(EWK)	5781.9	
N_{Gen}(Z mass, EWK)	4226.5	
N_{Gen}(QCD)	22789300.0	
N_{Gen}(Z mass, QCD)	20334900.0	
	Signal region	Control region
N^{data}	XXX	$18 \pm 4.2(stat.)$
N^{bkg}	N/A	$0.2 \pm 0.1(stat.)$
$N^{MC}(EWK)$	$7.9 \pm 0.2(stat.)$	$6.0 \pm 0.2(stat.)$
$N^{MC}(QCD)$	$29.5 \pm 3.0(stat.)$	$20.7 \pm 2.5(stat.)$
$\frac{N^{data}-N^{bkg}}{N^{MC}(EWK)+N^{MC}(QCD)}$	$0.67 \pm 0.16(stat.) \pm 0.06(MCstat.)$	
$Final N^{Z\to\nu\nu} estimate$	$157.3 \pm 37.6(datastat.) \pm 18.2(MCstat.)$	N/A

10 QCD background estimation

Multijet events that pass our selection have two very distinct origins:

- events with fake MET: MET comes from mismeasured jets in the events
- events with real MET: MET comes from decay of hadrons involving neutrinos, in particular heavy-flavour decays.

Because the purely fake MET contribution is very hard to model, we have oriented the basic selection of our analysis towards selecting true MET. This was achieved with a combination of requirements on the MET significance (> 3) and on the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j1/j2})$ ($> 1.$). In the following, we call this selection together with the trigger driven selection described in section 5.1, eq. 16 the QCD preselection. After such selection, we see about a factor 1.5 discrepancy between data and the sum of all other backgrounds (V+jets, top, VV), so we expect about a third of the events to be from multijet origin in the sample composition.

We then pursued three approaches to try and understand how to model the QCD contribution. First, a MC sample was produced using generator-level filters to enhance the real MET contribution. Second, additional jets in the events were used to select another jet pair that pass the selection, to provide an independent data-driven sample with QCD properties. Third, events with an additional jet $\mathrm{p}_T > 30$ GeV close to the MET in ϕ were used as data-driven QCD sample.

10.1 Monte-Carlo VBF-enriched QCD sample

Keeping the same $\hat{p_T}$ bins as for the inclusive QCD samples, the following generator-level filters were added:

- genMET $= ||\sum \vec{E_T}(\nu)|| > 40$ GeV
- At least 2 jets with $\mathrm{p}_T > 20$ GeV and $|\eta| < 5.0$
- At least one jet pair with $\mathrm{M}_{jj} > 700$ GeV, $|\Delta\eta_{jj}| > 3.2$.

The filters were tuned to allow private production of samples in a reasonable time but with enough statistics for the analysis.

Private samples have been produced using the same configuration as for the inclusive samples. The DBS and other characteristics of these samples are summarised in Table 9. To highlight the two population of events present in the QCD process, Fig. 10 shows the reconstructed PFMET as a function of the generator-level MET for all QCD inclusive events with $80 < \hat{p_T} < 600$ GeV. The red line shows the generator-level cut at 40 GeV applied in the enriched sample, and the blue line the PFMET cut at 130 GeV that was applied in the published prompt analysis (HIG-13-030).

Distributions after applying the QCD preselection are shown in Fig. 11 and Fig. 12. MC events are normalised to the data luminosity using the inclusive QCD LO cross sections.

The two leading jets, passing the VBF selection, are light quarks or gluons mainly. The leading jet not passing the VBF selection is mostly heavy-flavoured (Fig. 11), and at the origin of the large MET and MET significance. Most events have three jets with $\mathrm{p}_T > 30$ GeV, with a jet $\mathrm{p}_T > 30$ GeV aligned with the MET in ϕ (Fig. 12), this jet being as expected the heavy-flavoured jet.

Data-MC agreement is shown in Fig. 13 and 14 after the QCD preselection. As can be seen, the MC QCD is biased by the genMET cut, and does not reproduce the data for most variables. The

Table 9: DBS links and filter efficiencies for the private production of VBF-enriched QCD samples.

Database URL	https://cmsdbsprod.cern.ch:8443/cms_dbs_ph_analysis_01_writer/servlet/DBSServlet
QCD-Pt-80to120	/VBFQCD_Pt_80to120_MET40_step1_v1/pela-VBFQCD_Pt_80to120_MET40_step3_v1-3664d28163503ca8171ba37083c39fc9/USER
QCD-Pt-120to170	/VBFQCD_Pt_120to170_MET40_step1_v1/pela-VBFQCD_Pt_120to170_MET40_step3_v1-3664d28163503ca8171ba37083c39fc9/USER
QCD-Pt-170to300	/VBFQCD_Pt_170to300_MET40_step1_v1/pela-VBFQCD_Pt_170to300_MET40_step3_v1-3664d28163503ca8171ba37083c39fc9/USER
QCD-Pt-300to470	/VBFQCD_Pt_300to470_MET40_step1_v1/pela-VBFQCD_Pt_470to600_MET40_step3_v1-3664d28163503ca8171ba37083c39fc9/USER
QCD-Pt-470to600	/VBFQCD_Pt_470to600_MET40_step1_v1/pela-VBFQCD_Pt_470to600_MET40_step3_v2-3664d28163503ca8171ba37083c39fc9/USER

Sample	Ev. Gen.	Filter Eff.	Events	XS $[pb]$	Eq. Lumi. $[fb^{-1}]$
QCD-Pt-80to120	39376000000	0.000049	1614416	1033680	38.09
QCD-Pt-120to170	7000000000	0.000283	2051000	156293.3	44.79
QCD-Pt-170to300	1375000000	0.000987	1391500	34138.15	40.28
QCD-Pt-300to470	80000000	0.002659	207840	1759.549	45.47
QCD-Pt-470to600	25000000	0.004127	104675	113.8791	219.53

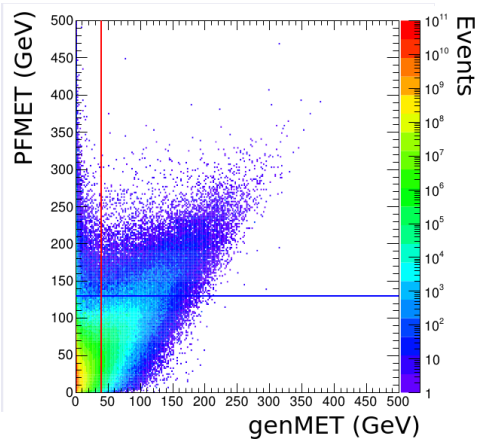


Figure 10: Reconstructed PFMET as a function of generator-level MET in the inclusive QCD samples $80 < \hat{p_T} < 600$ GeV before any selection.

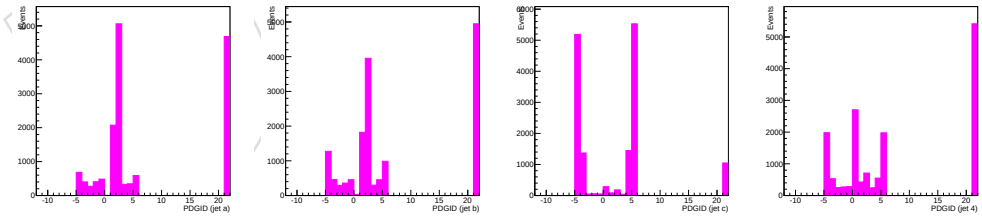


Figure 11: PDGID of the first 4 jets in the events.

MC tends to overpredict heavy-flavours compared to the data. Hence it cannot be used as is, and data-driven methods are investigated instead.

10.2 Data-driven QCD based on combinatorial

Given that it is not a fundamental property of QCD multijet events to produce the two leading jets with the VBF properties, it seems natural to try and use other jet pairs in the event to mimic the leading pair.

In the following, shapes are compared for different variables when using different pairs passing

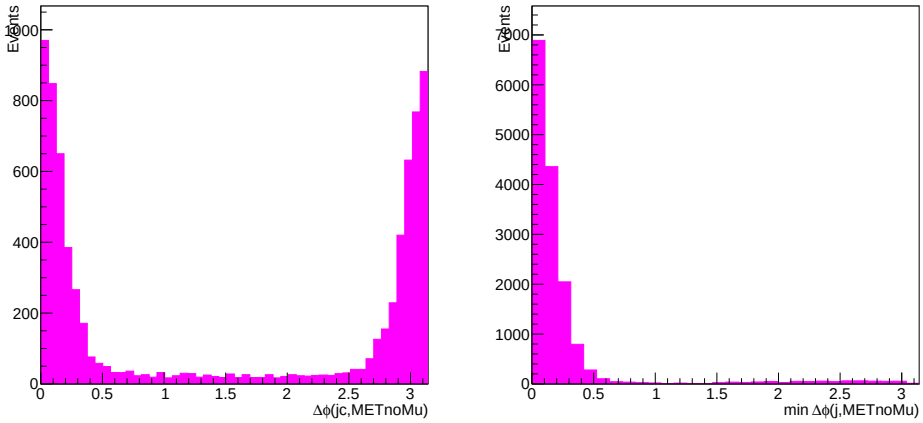


Figure 12: Left: $\Delta\phi$ between the MET and the 3rd jet. Right: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$.

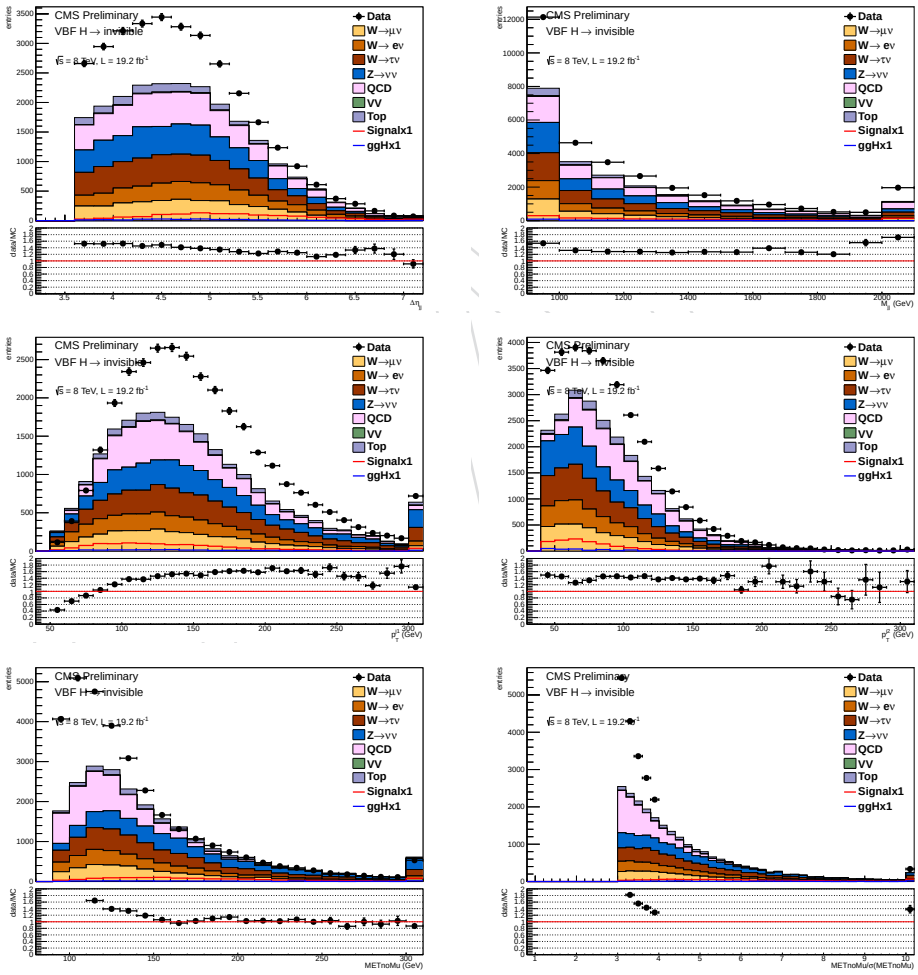


Figure 13: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum, METnomu and MET significance after the QCD preselection.

the preselection (removing possible overlaps). We call j1j2, j1j3 and j2j3 the different sort of events, where j1, j2 and j3 represent the first 3 p_T-ordered jets in the event.

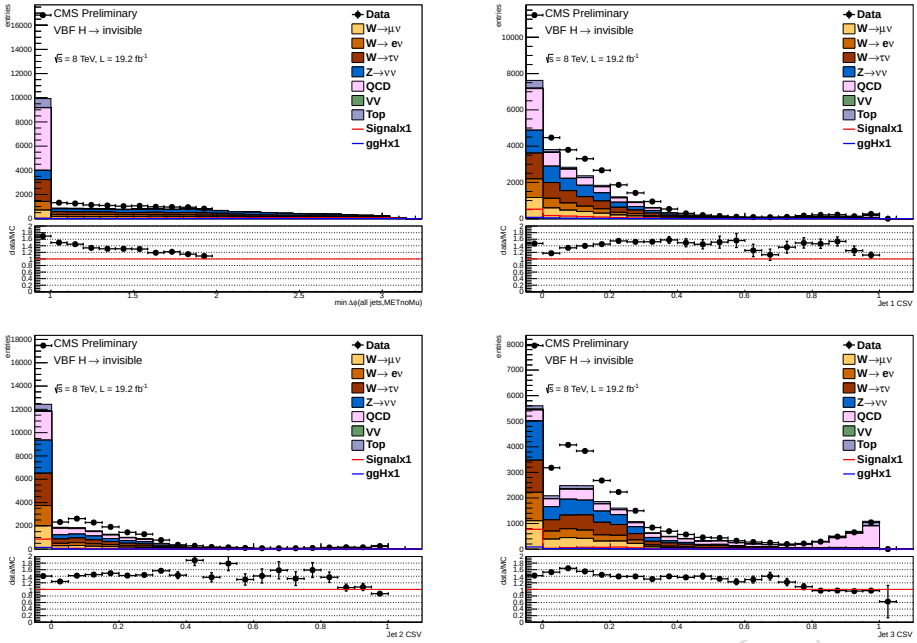


Figure 14: min$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ and CSV b-tagging discriminant for the three leading p_T jets.

Fig. 15 show the sample composition expected for the different selections after the QCD pre-selection. The VBF-enriched QCD MC is shown in pink, normalised to the data minus other backgrounds (V+Top+VV). The number of events in each sample are summarised in Table 10.

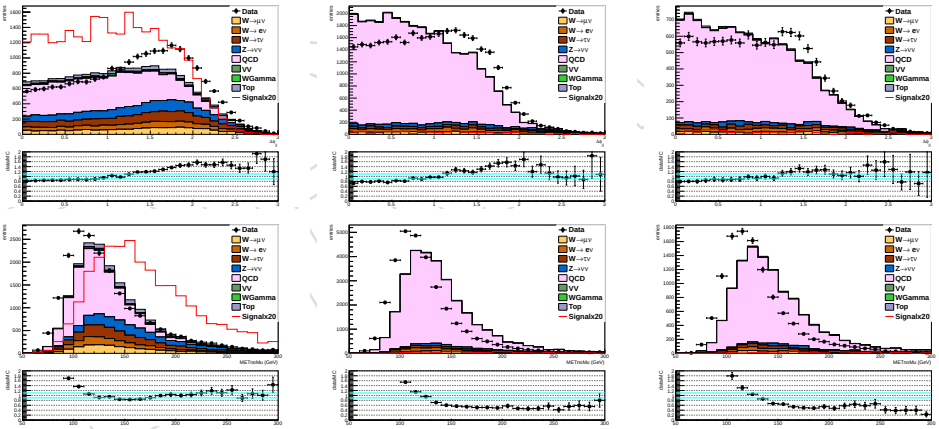


Figure 15: Sample compositions for j1j2 (left), j1j3 (middle) and j2j3 (right) pair selections, as a function of $\Delta\phi_{jj}$ (top) and METnoMuons (bottom).

Table 10: Sample compositions, with number of events after QCD preselection, duplicated with j1j2 or j1j2+j1j3 selections, and expected MC from V+jets, Top and VV backgrounds.

Pair	nPresel	nDupl	nSel	nMC
j1j2	19980	0	19980	9268.81
j1j3	30069	594	29475	4110.59
j2j3	11114	1689	9425	1735.98

To compare the shapes, in order to see whether the j1j3 and j2j3 samples could be used to model

the shape of the QCD in the j1j2 selection, backgrounds from V+jets, Top and VV are subtracted, after being normalised to their control regions as explained in sections 8 and 9, using the QCD preselection.

The results are shown in Fig. 16 for several variables. For each set of two plots, the left plot shows the data compared to QCD+Bkg normalised to the data luminosity (the data-driven QCD being normalised to data-Bkg), and the right plot shows the shape agreement (distributions normalised to 1) after background subtraction for the different QCD samples (MC, j1j3, j2j3) compared to the data-Bkg j1j2.

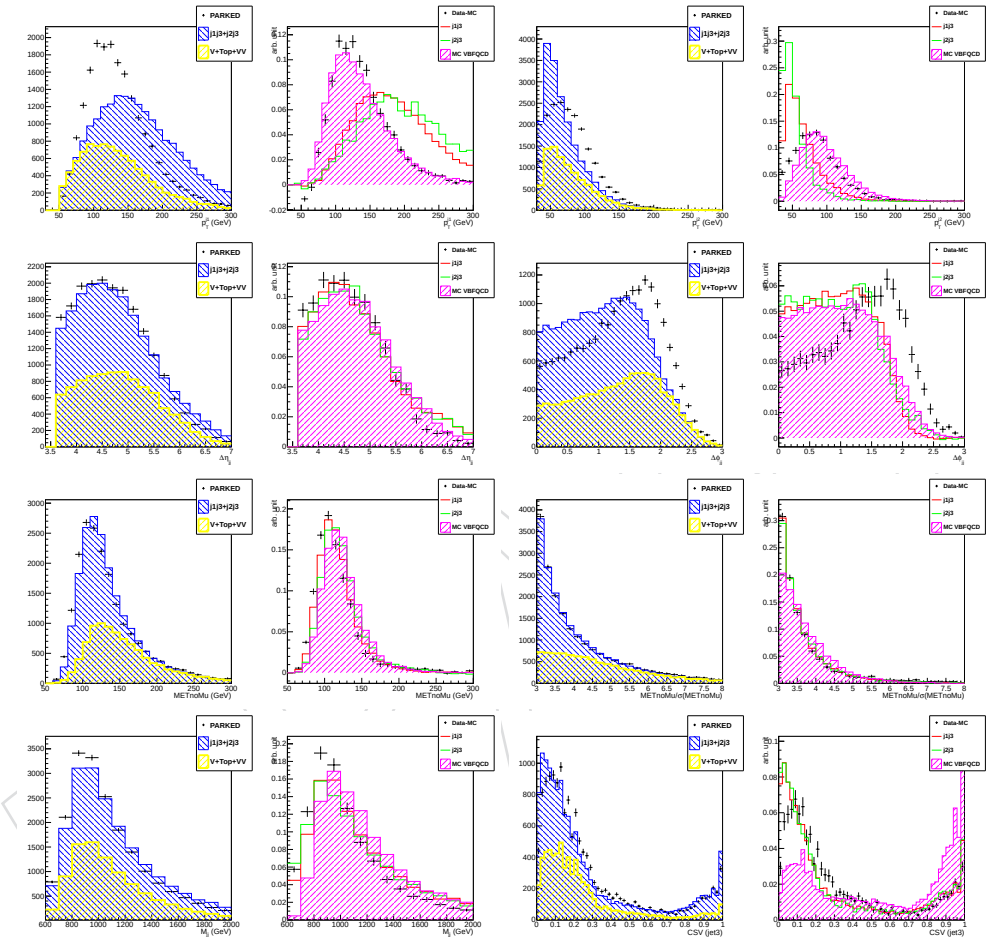


Figure 16: For each set of two plots, the left plot shows the data compared to QCD+Bkg normalised to the data luminosity (the data-driven QCD being normalised to data-Bkg), and the right plot shows the shape agreement (distributions normalised to 1) after background subtraction for the different QCD samples (MC, j1j3, j2j3) compared to the data-Bkg j1j2.

The MET variables and heavy-flavour content are much better reproduced using the data-driven samples than the MC QCD, but the jet kinematics is very different, which leads to a different kinematic of the jet pair, in particular for $\Delta\phi_{jj}$. Other ordering of the jet pairs have been tried (angular based on $\Delta\phi_{jj}$ for example), with no particular success in reproducing the shapes of both the jet and MET kinematics. Reweighting either the leading jet p_T or the $\Delta\phi_{jj}$ distributions also does not solve the discrepancies on the other variables. This approach has hence been abandoned.

10.3 Data-driven QCD based on inverted selection on $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$

From the previous two subsections, we have learned that the QCD background is mainly composed of events with a lower p_T third jet having a high neutrino content, and hence mostly aligned in ϕ with the MET. Whether that jet is reconstructed or not will lead to non-isolated or isolated MET respectively. Given the difficulties in finding a data-driven modeling of its shape, we hence decided to benefit from this property to further reject this process. Fig. 17 shows the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ distribution for events passing the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j1/j2}) > 1.0$ QCD preselection.

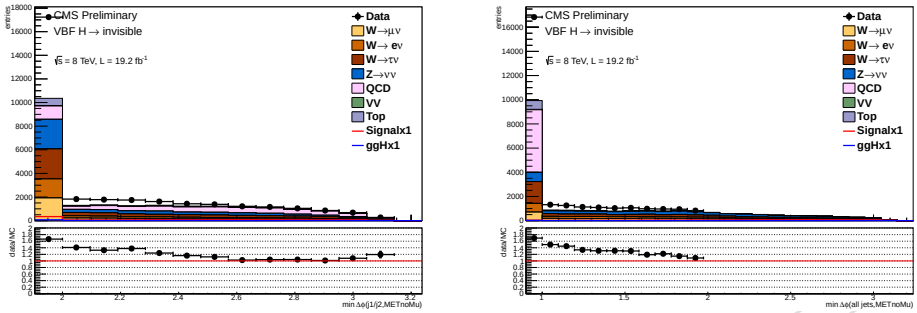


Figure 17: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j1/j2})$ (left) and $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ (right) distributions, for events passing the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j1/j2}) > 1.0$ QCD preselection.

Most QCD events have $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) < 2.0$. After some optimisation based on the expected limits (see section 5.2), the signal region is hence decided at $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) > 2.3$ (see section 5.1, eq. 17). All MC QCD generated $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) < 0.5$, when the data-MC discrepancy shows that there are QCD events expected also above still.

In the following, we investigate whether we can use events with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) < 1.0$, but with the two leading jets far from the MET (i.e. $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j1/j2}) > 2.3$), to model the shape of events with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) > 2.3$. In other words, we use the non-isolated MET events to model the isolated MET events, keeping all other cuts as in the final signal region. Data is compared with MC in Fig. 18 and Fig. 19 for this inverted selection. V+jets and top backgrounds are normalised using their control regions. A very good agreement between data and the MC VBF-enriched sample is observed, although a normalisation factor of 0.5 is applied to the QCD to match the data. It seems that the true MET component of the MC is well modeled when there is a third jet $p_T > 30$ GeV in the ϕ vicinity of the MET.

To investigate further the use of these events to model the QCD contribution for large $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$, we loosen a bit the selection to $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) > 1.0$ to have some QCD left to compare to. To further enrich in QCD and kill the signal contributions, a third jet is explicitly asked. The data-MC agreement is shown in Fig. 20, where the pink distribution is from the inverted selection normalised to the data-Bkg, and V+jets and Top Bkg are normalised to their control regions as usual.

Although the agreement is not perfect, the inverted sample does seem to give an appropriate data-driven model, with data-MC agreement within ±20% in the QCD-dominated regions. The scale factor obtained in this 3-jet selection can however not be applied in the 2-jet bin! And the signal contamination in the 2-jet bin is too high to allow for a normalisation of the QCD.

In conclusion, these studies showed us several things:

- the MC is not able to reproduce multijet events with large and isolated MET, but it does reproduce the data for non-isolated MET.

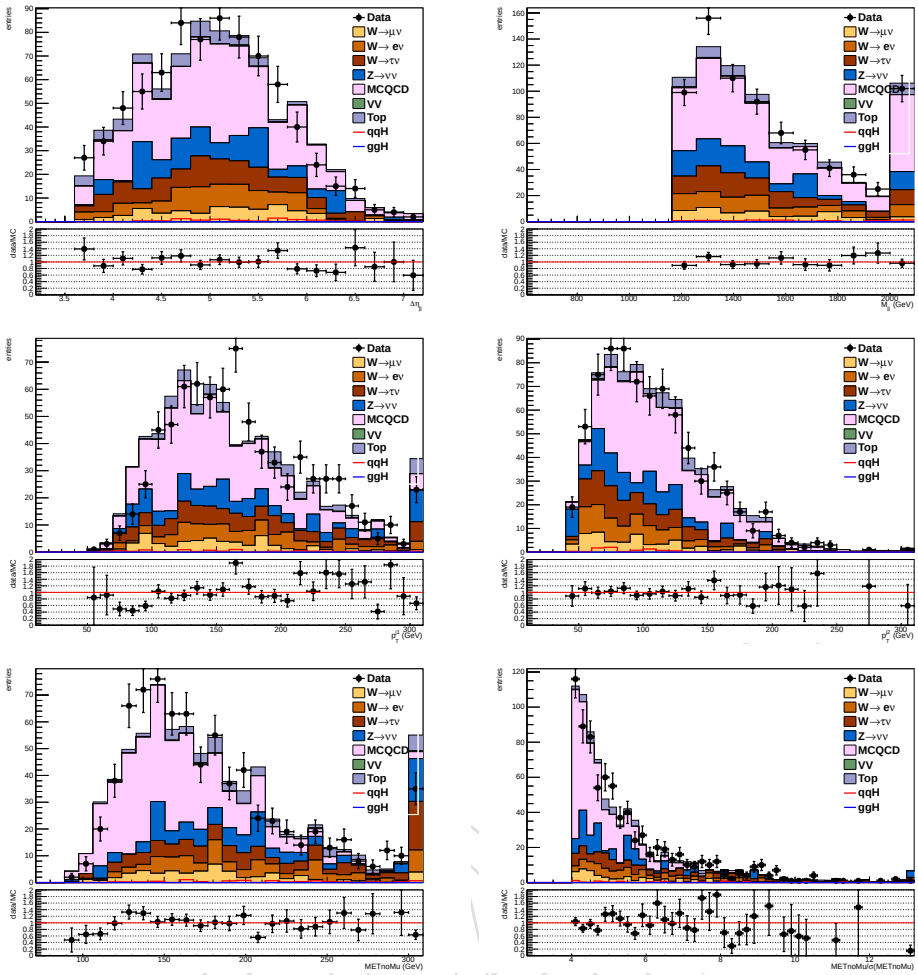


Figure 18: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum, METnomu and MET significance after the inverted selection.

- even if we find a data-driven sample for the shape, it does not seem possible with the statistics available and the trigger requirements to find a QCD control region with enough QCD having the same properties as the signal, but negligible signal contribution, to provide the required normalisation.

The next section is hence dedicated at finding a solution to extrapolate the normalisation from a looser selection to the signal region.

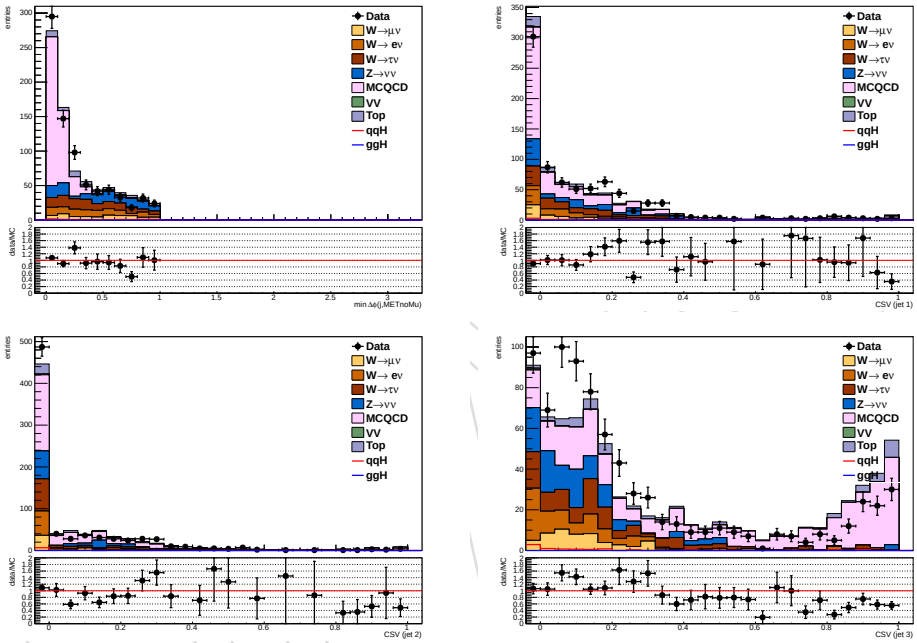


Figure 19: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ and CSV b-tagging discriminant for the three leading p_T jets after the inverted selection.

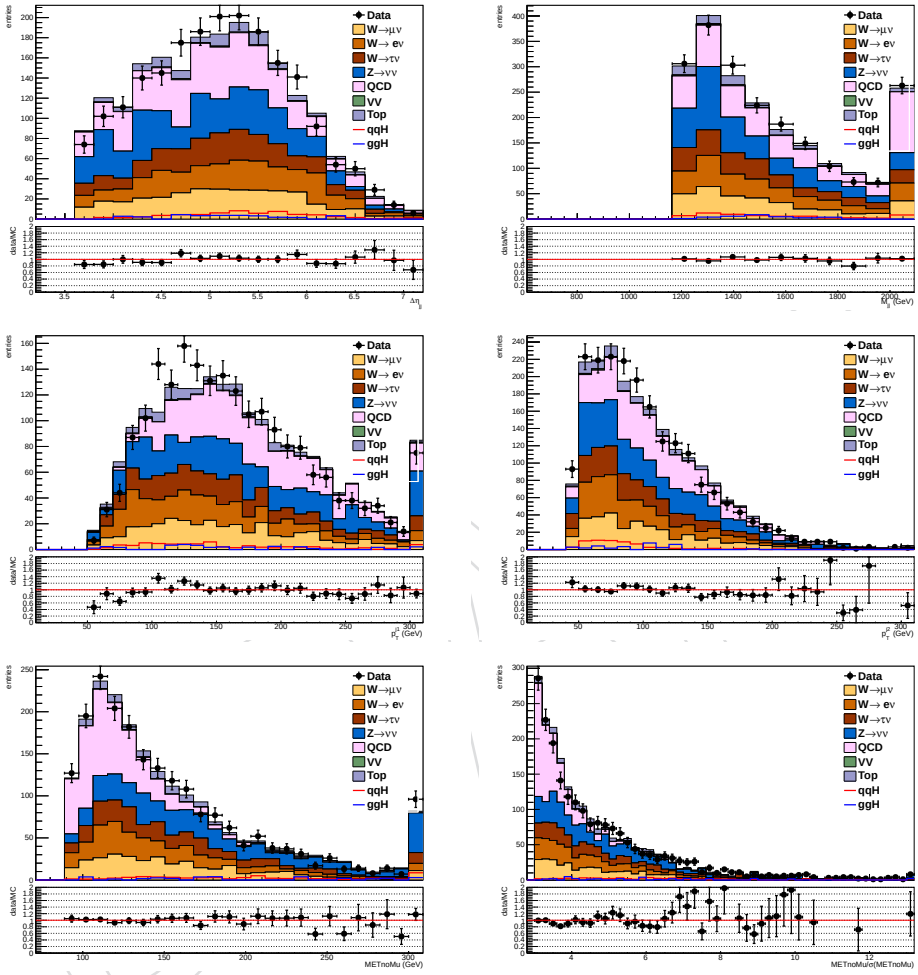


Figure 20: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum, METnomu and MET significance for events with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) > 1.0$ and 3 jets $p_T > 30$ GeV. The QCD sample is modeled by data with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) < 1.0$ but $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j1/j2}) > 1.0$.

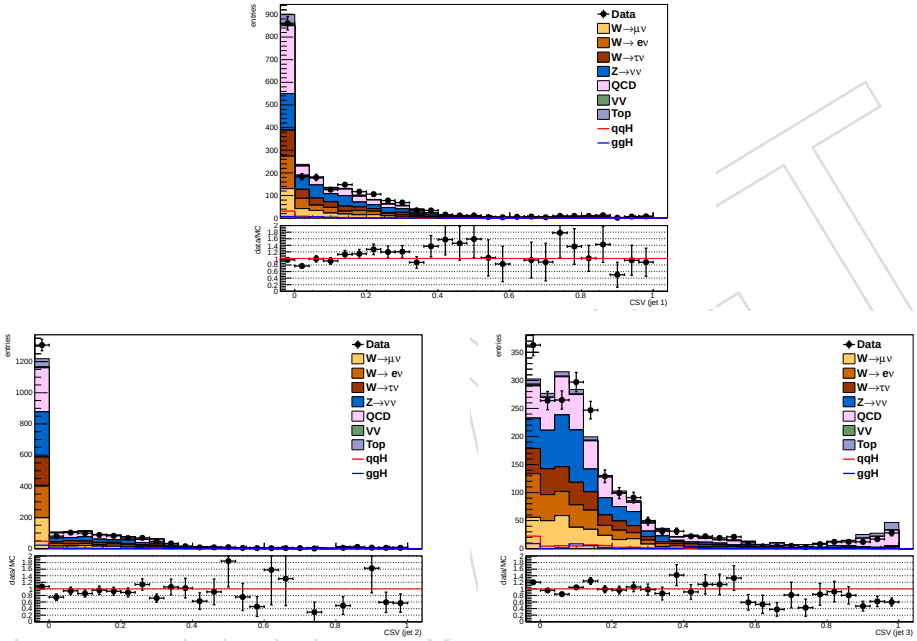


Figure 21: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ and CSV b-tagging discriminant for the three leading p_T jets for events with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) > 1.0$ and 3 jets $\mathrm{p}_T > 30$ GeV. The QCD sample is modeled by data with $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j}) < 1.0$ but $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j1/j2}) > 1.0$.

10.4 Final estimate in signal region

The different regions used in the final estimate are summarised in Fig. 22. The sample compositions for the different regions are given in Table 11.

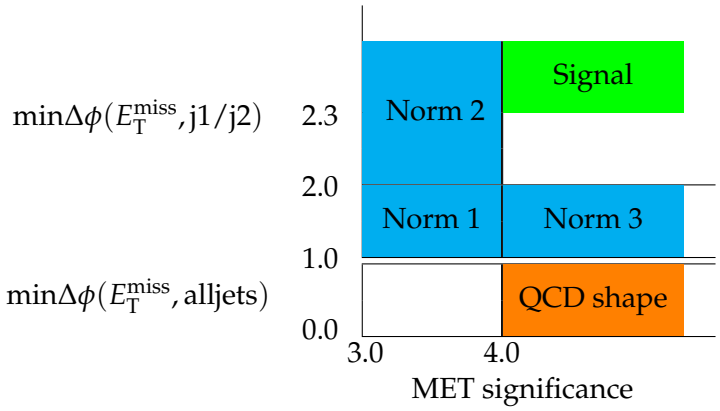


Figure 22: Definition of the signal and normalisation regions.

Table 11: Sample composition for the different regions. The number of QCD events is the number of data events passing the inverted selection before any normalisation, but after background subtraction. V+jets and top backgrounds are normalised to their respective control regions.

Region	Signal	Bkg	Data	signal frac	QCD
Norm1	50±5	1188±29	1586	3%	2290±55
Norm2	51±4	297±14	411	12%	1954±48
Norm3	132±8	1300±34	1517	9%	438±31
Signal	295.9±11.3	420±16	XXXX	-	362±36

We use the QCD shape region described in section 10.3 and Figure22 to estimate the shape of the QCD contribution to our signal region. This shape is normalised to data with a scale factor extracted for each region from the min$\Delta\phi(E_T^{miss}$, j) or met significance distributions using the formula: SF=$\frac{\int_a^b n_{Data-Bkg} dx}{\int_a^b n_{QCD} dx}$, as shown in Fig. 23, 24 and 25, 26. The SF are summarised in Table 12. A fit function (exponential or linear) is used to extrapolate to the signal region, independently for each variable. Because the variables are correlated, only one scale factor should be used. As the signal contamination is large in norm2 and norm3 regions, we use only the estimates from norm1, norm2 and 3 being used for cross-check in the no-signal hypothesis. A good consistency is found between all estimates and both variables, except for norm3 estimates. Norm3 is however a very weird region in which we ask for very good MET (metsig$>$ 4) but no jets recoiling against the MET in phi (min$\Delta\phi(E_T^{miss}$, j) $<$ 2). It also has poor statistics.

In the end, the SF used for estimating the QCD in the signal region is taken as the middle value of the envelope given by the two norm1 extrapolations, and the envelop as systematic uncertainty, that is SF= 0.048 ± 0.040. Given we expect 362±36 events in the QCD sample (Table 11), the final estimate is 17 ± 14 QCD events in the signal region.

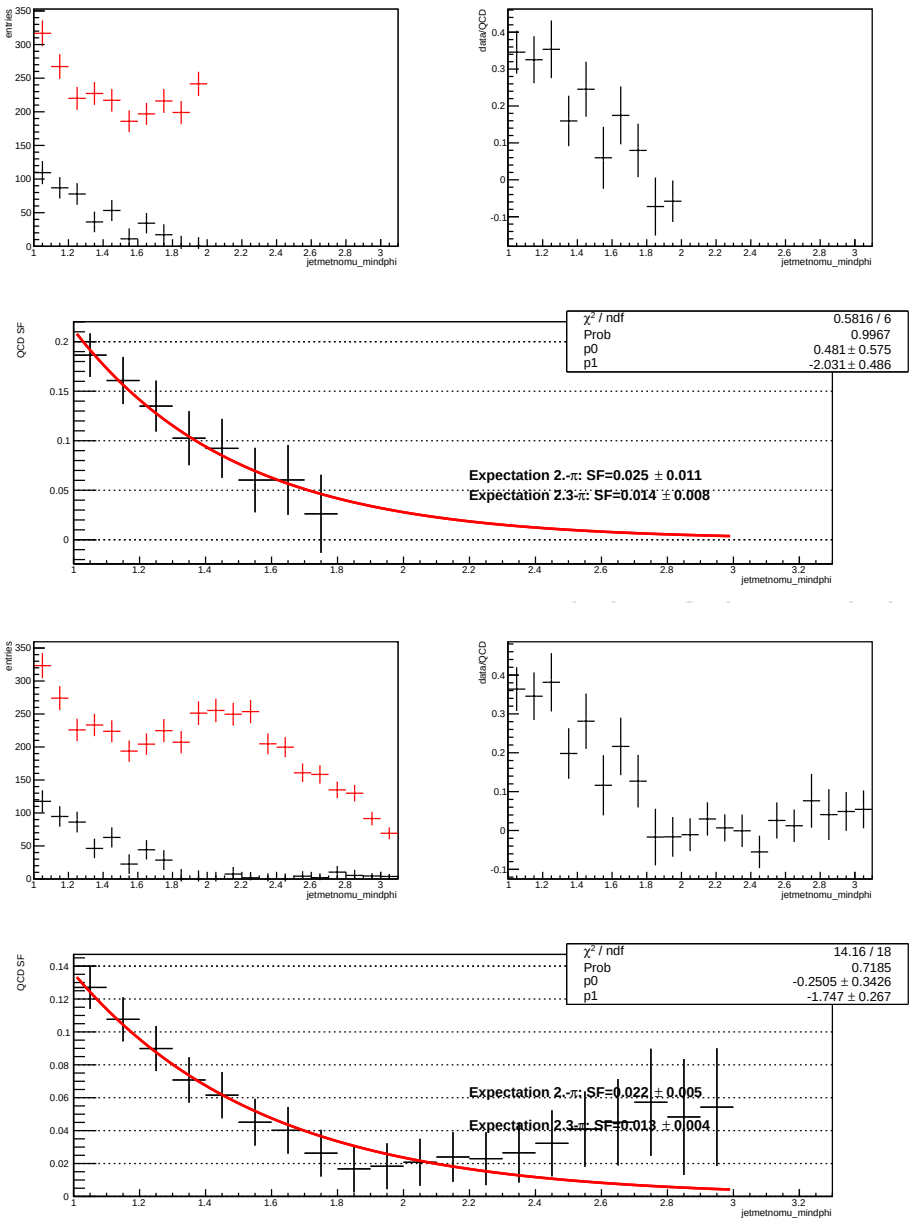


Figure 23: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ variable. For each set of three plots: top-left is raw number of data and QCD, both background-subtracted; top-right is ratio data/QCD; bottom is scale factor SF (see text). Top 3 plots: norm1 region. Bottom: norm1+norm2.

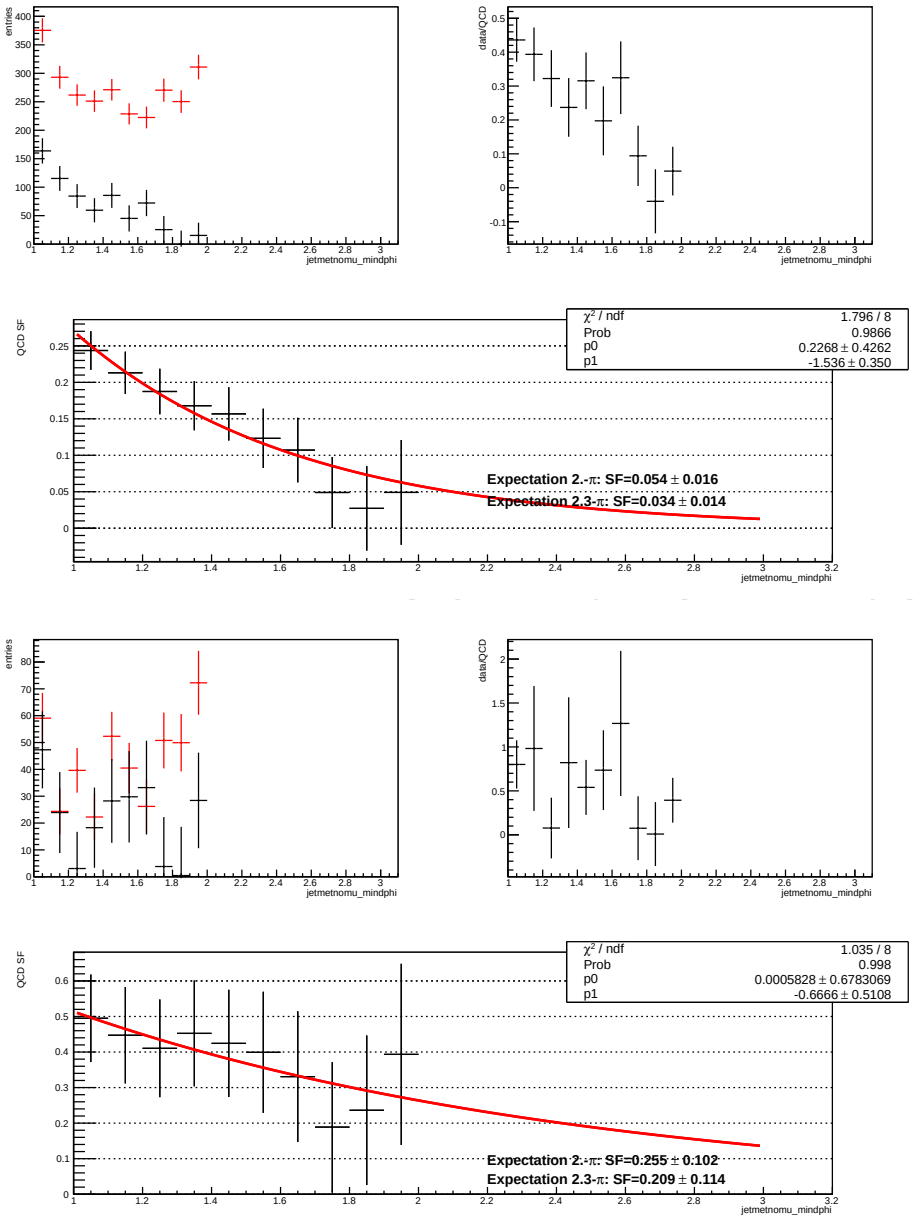


Figure 24: Same as Fig. 23, but for norm1+norm3 (top) and norm3 (bottom).

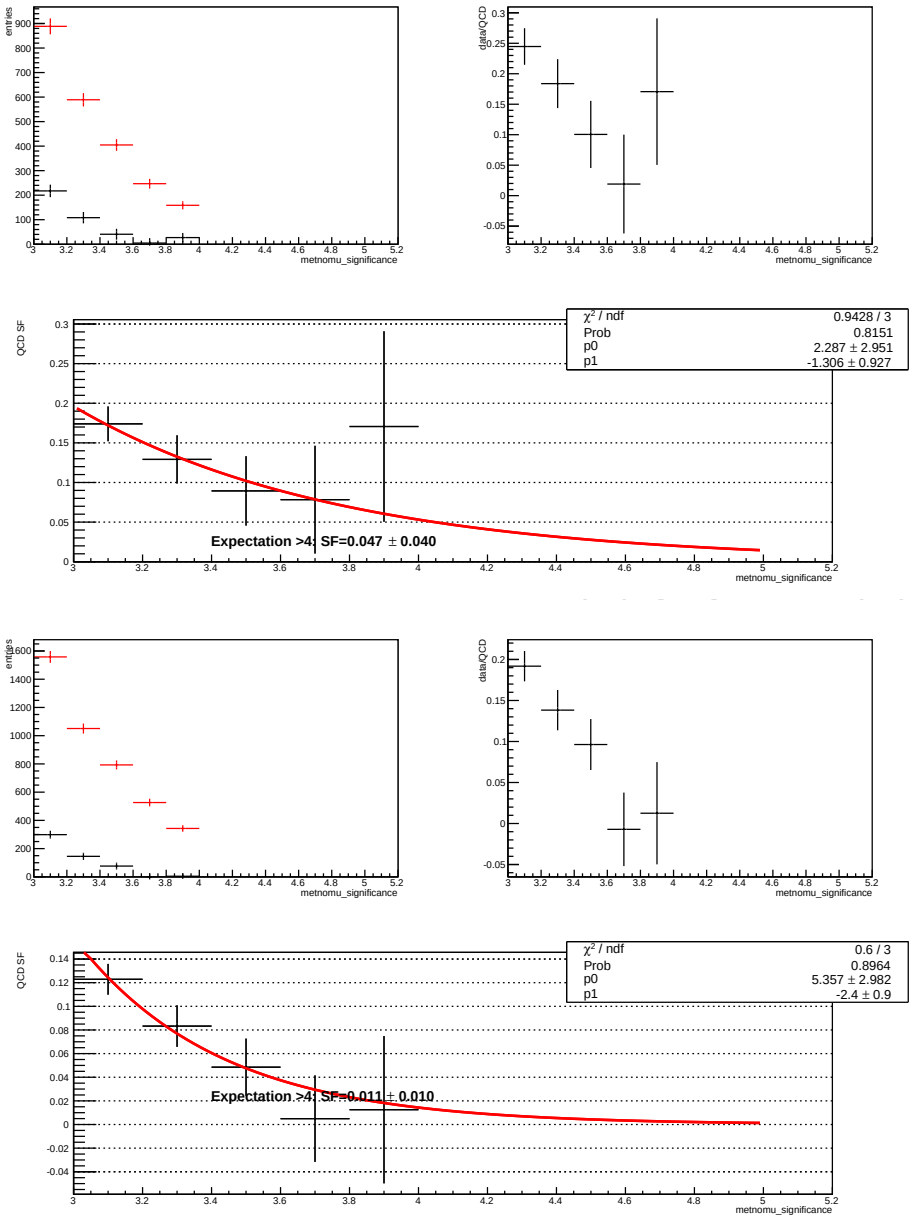


Figure 25: Same as Fig. 23 but for the met significance variable. Top: norm1. Bottom: norm 1+2.

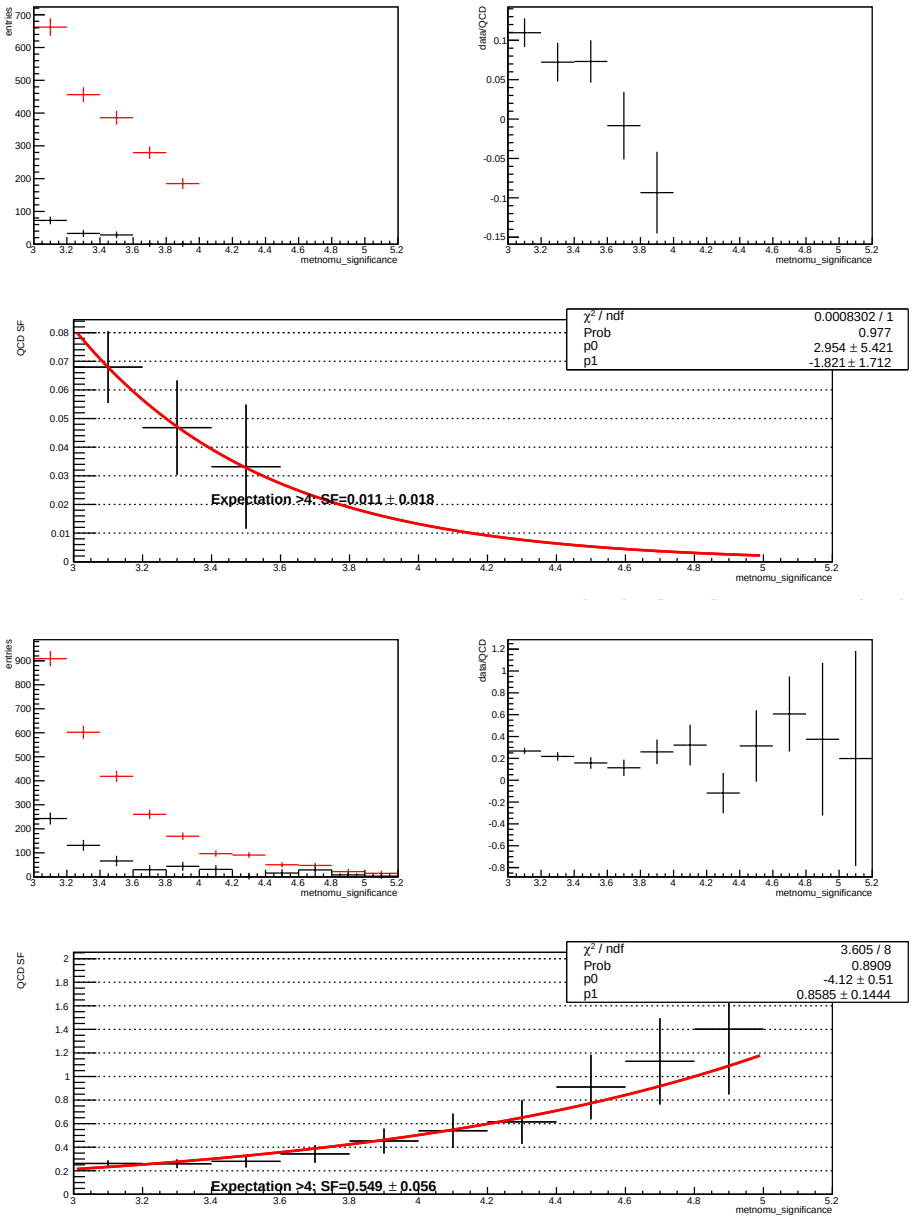


Figure 26: Same as Fig. 23 but for the met significance variable. Top: norm2. Bottom: norm 1+3.

Table 12: Summary of the QCD scale factors for the different regions ans extrapolations to the signal region independently for both variables.

Region	Factor	Extrapolation mindphi$>$ 2.3	Extrapolation metsig$>$ 4
Norm 1	0.17 ± 0.02	0.014 ± 0.008	0.05 ± 0.04
Norm 1+2	0.12 ± 0.01	0.013 ± 0.004	0.01 ± 0.01
Norm 1+3	0.24 ± 0.03	0.03 ± 0.01	0.55 ± 0.06
Norm 2	0.06 ± 0.01	-	0.01 ± 0.02
Norm 3	0.5 ± 0.1	0.209 ± 0.114	-

11 Results

The results of all of the background and signal estimation procedures with their associated statistical and systematic errrors are shown in table 13. Details of the sources of the systematic uncertainties can be found in section 12. Blinded plots of all variables that make up the signal region selection are shown in figure 27.

Table 13: Summary of estimated backgrounds and observed yield in the signal region.

Background	$N_{est} \pm (stat) \pm (syst)$
$Z \to \nu\nu$	$157.3 \pm 37.6 \pm 38.3$
$W \to \mu\nu$	$101.8 \pm 6.1 \pm 11.9$
$W \to e\nu$	$57.4 \pm 7.3 \pm 7.0$
$W \to \tau\nu$	$98.0 \pm 13.2 \pm 25.4$
top	$4.4 \pm 1.0 \pm 1.4$
VV	$3.8 \pm 0.0 \pm 0.7$
QCD multijet	$17 \pm 0 \pm 14$
Total Background	$439.7 \pm 41.0 \pm 55.8$
Signal(VBF)	$273.4 \pm 0.0 \pm 31.2$
Signal(ggH)	$22.6 \pm 0.0 \pm 15.6$

12 Systematics

Our dominant systematic uncertainties come from the statistical uncertainty on the number of data events in each of our control regions.

In order to estimate the effect of the jet energy scale (JES), unclustered energy scale (UES) and jet energy resolution (JER) on both the signal and background processes we vary each separately up and down by 1 standard deviation and repeat the full background estimation. The size of the variation is taken from the JET-MET POG recommendations.

As described in Section 4.2 we use the lepton identification efficiencies from the EGamma and Muon POGs, these come with associated uncertainties and we estimate their impact by repeating the background estimation process with the electron and muon identification efficiencies varied up and down by 1 standard deviation. In the case of hadronically decaying tau leptons we use the tau POG recommended uncertainty of 8% on the identification efficiency for the tau reconstruction algorithm that we use, which we place as a systematic on our $W \to \tau\nu$ background estimation.

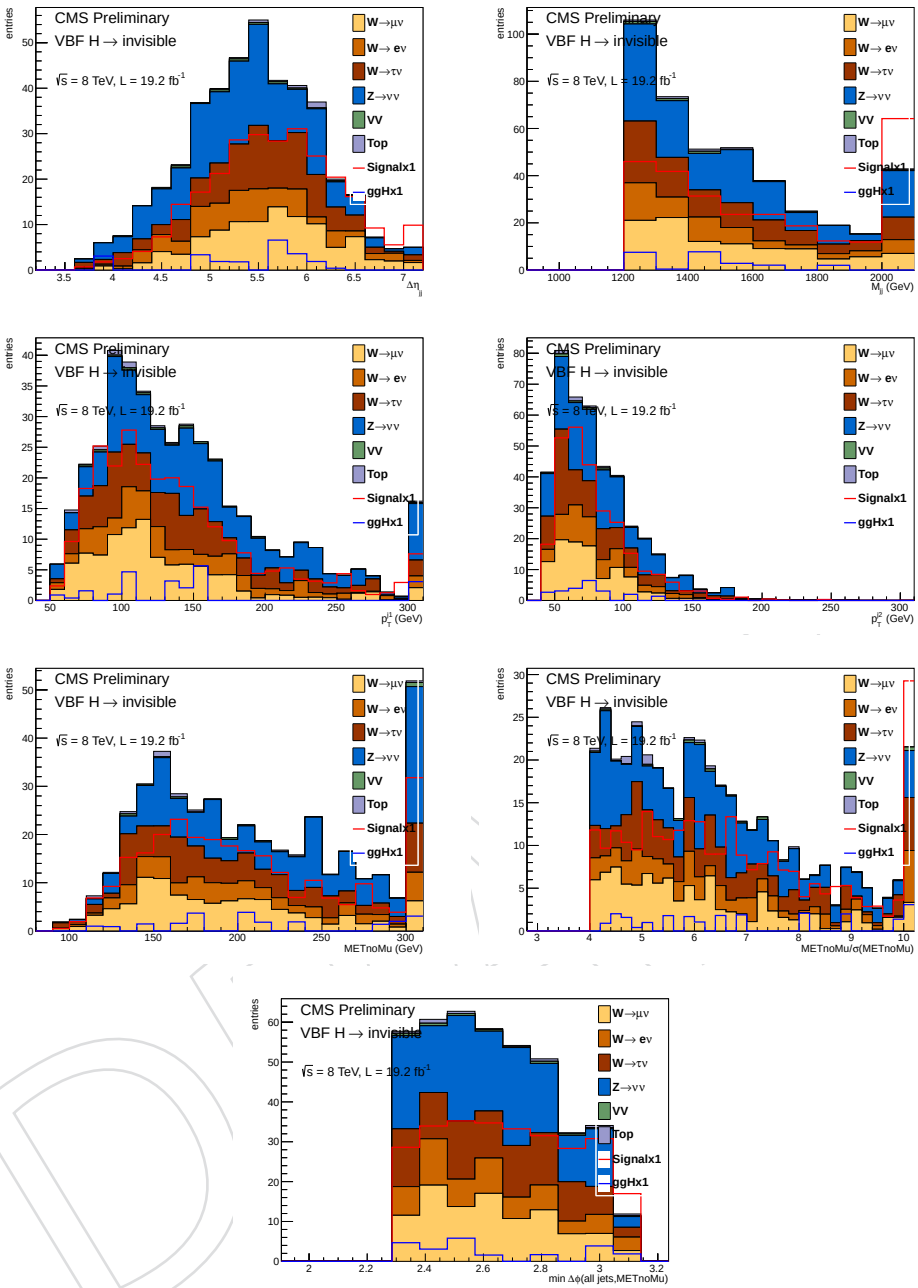


Figure 27: Dijet pseudorapidity difference, dijet mass, leading and subleading jet transverse momentum, METnomu, METnomu significance and min$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j})$ in the signal region. As this analysis is still blind the data are not shown

We also consider the effect of the uncertainty in our pileup reweighting (described in Section 4), this is done by varying the event-by-event weights that we apply up and down by one standard deviation and repeating the background estimation procedure. Another source of uncertainty is the uncertainty on the integrated luminosity analysed, we apply a 2.6% uncertainty on our estimate of all processes not normalised from data (i.e. the signal processes and diboson production) as recommended by the Lumi POG. We also have uncertainties on the production cross-section of the diboson process, which is the only background processes not normalised from data, we take this uncertainty from the uncertainty on the CMS published cross-section

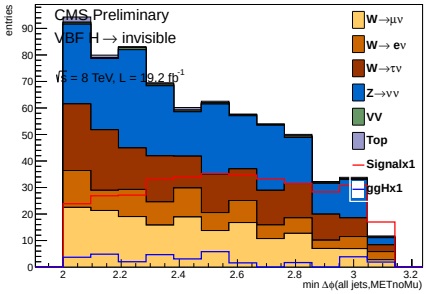


Figure 28: $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ in the signal region with the $\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{j})$ cut loosened to 2. The final background normalisations have been used. As this analysis is still blind the data are not shown

measurements of the relevant processes [1–3].

Further uncertainties come from the theoretical uncertainty on the cross-section ratio used to extrapolate from QCD produced $Z/\gamma^* \to \mu\mu$ to $Z \to \nu\nu$ in the $Z \to \nu\nu$ background estimation. We estimate this uncertainty by calculating the ratio of yields obtained from both MCFM and MadGraph in a VBF dominated region for both processes. The MCFM NLO result is found to be 1.14, while the MadGraph result is 1.2 ± 0.2 and we place a 20% uncertainty on the $Z \to \nu\nu$ estimate to account for this difference. Furthermore as mentioned in Section 8.3 we apply a 20% uncertainty to our estimation of the $W \to \tau\nu$ background to account for the difference in selection between the $W \to \tau\nu$ control region and our signal region.

Finally we have theoretical uncertainties on the signal cross-sections. We take the parton density function uncertainties on the total VBF and gluon fusion production cross-sections from the LHC Higgs Cross Section Working Groups Yellow Report 3. We also take the QCD scale uncertainty on the VBF production cross-section from Yellow Report 3. For the remaining gluon fusion production cross-section uncertainties, we use an estimate carried out using the Higgs to tau tau analysis framework and samples in the phase space of our analysis. This estimate gives a total uncertainty on the gluon fusion process from QCD scale, pdf and showering uncertainties of 59%.

We also studied the effect of the uncertainties in our modelling of the trigger efficiency. As described in Section 4.1 we measure the trigger turn on by fitting an error function to the trigger efficiency as a function of METnomu in bins of jet 2 p_T and dijet mass, after our selection only the bins with the highest values of dijet mass, and the two highest values of jet 2 p_T are used. It can be seen in Appendix A that for these bins the maximum is the only parameter of the error function with an uncertainty of any significance. As the effect of changing the maximum is to uniformly scale all MC events by the same factor this uncertainty will only effect processes taken directly from MC.

To estimate the maximum possible effect of this uncertainty we took the bin from each run period with the largest uncertainty and assumed that the luminosity weighted average of these uncertainties was the uncertainty for all bins. This worst case scenario estimate gave a 2.3% uncertainty, assuming the VV contribution to be negligible (it accounts for less than 1% of the total background), we applied this uncertainty to all signal processes, and it did not affect the expected limit. Furthermore this 2.3% uncertainty is far larger than the average uncertainty on the trigger efficiency bins used in our analysis. For these two regions we have decided not to consider this effect further.

13 Extraction of limits

We set 95% C.L. upper limits on the Higgs boson production cross section times invisible branching fraction using an asymptotic CLs method[4–6], following the standard CMS Higgs combination technique implemented using the standard CMS Higgs combination tool. The tool takes as input datacards with the yields for each process and "nuisance parameters" which represent the systematic uncertainties in each yield and their correlations. The datacards are set up to consider 6 background processes ($Z \to \nu\nu$, $W \to \mu\nu$, $W \to e\nu$, $W \to \tau\nu$, top quark production and diboson production) and 2 signal processes (VBF and gluon fusion production of a Higgs boson decaying invisibly). Each systematic described in section 12 is included with its own log-normal probability density function (PDF), except where there are multiple uncorrelated systematics affecting a single process, where they are combined into one datacard entry.

Using this procedure and assuming standard model Higgs boson kinematics, the expected 95% C.L. limit that we obtain on the invisible branching fraction of a SM 125 GeV Higgs boson is 38%. The 95% C.L. limit on the invisible branching fraction of a SM Higgs boson and the 95% C.L. limit on the cross-section times invisible branching fraction are shown as a function of Higgs boson mass in figure 29

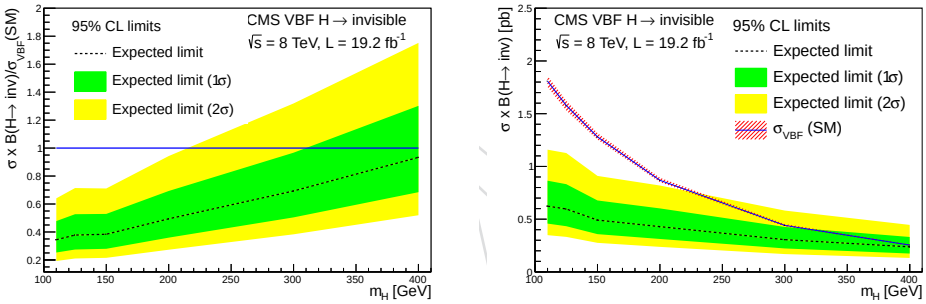


Figure 29: The 95% C.L. limit on the invisible branching fraction of a SM Higgs boson (left) and the 95% C.L. limit on the cross-section times invisible branching fraction (right)as a function of Higgs boson mass

14 Conclusion

We have reviewed the VBF Higgs to invisible analysis using the parked triggers, and in view of preparing for Run II. Using a selection driven by enhancing the contribution from MET coming from genuine invisible particles rather than mismeasured energy, we are able to improve the expected 95% C.L. limit by more than 20% compared to the published prompt analysis, with a new expected limit on the branching ratio of a standard model Higgs to invisible at 38% (compared to 49%). The improvement comes mainly from a better set of variables for the cut-based analysis allowing us to reject more efficiently the QCD multijet backgrounds whilst increasing the signal efficiency slightly. The use of the parked triggers also allows us to loosen the kinematic selection, in particular on the dijet mass, subleading jet p_T and MET.

References

[1] CMS Collaboration, "Measurement of WZ production rate", CMS Physics Analysis Summary CMS-PAS-SMP-12-006, (2012).

Table 14: The median expected limit taking all nuisances into account, and ignoring all nuisances are shown at the top of the table. Also shown is the relative change in the expected limit when each nuisance is removed from an uncertainty model with all nuisances considered, and the relative change in the expected limit when each nuisance is added to an uncertainty model with no other nuisances considered.

Median expected limit with:	All Nuisances: 37.8%	No Nuisances: 14.3%
Nuisance	Removal effect	Addition effect
lumi_8TeV:	0.0%	0.0%
CMS_eff_e:	0.0%	0.6%
CMS_eff_m:	-0.5%	6.8%
CMS_scale_j:	-6.5%	12.2%
CMS_res_j:	0.0%	0.0%
CMS_scale_met:	0.0%	0.0%
CMS_VBFHinv_puweight:	0.0%	0.6%
CMS_VBFHinv_zvv_norm:	-2.6%	24.5%
CMS_VBFHinv_zvv_stat:	-12.4%	65.8%
CMS_VBFHinv_wmu_norm:	-1.0%	6.1%
CMS_VBFHinv_wmu_stat:	-0.5%	3.4%
CMS_VBFHinv_wel_norm:	-0.5%	3.4%
CMS_VBFHinv_wel_stat:	-0.5%	4.8%
CMS_VBFHinv_tau_eff:	-0.5%	5.5%
CMS_VBFHinv_tau_extrapfacunc:	-4.1%	24.5%
CMS_VBFHinv_wtau_norm:	-1.5%	12.2%
CMS_VBFHinv_wtau_stat:	-1.5%	13.6%
CMS_VBFHinv_zvv_extrapfacunc:	-8.3%	53.5%
CMS_VBFHinv_top_norm:	0.0%	0.0%
CMS_VBFHinv_top_stat:	0.0%	0.0%
CMS_VBFHinv_qcd_norm:	-1.5%	7.5%
CMS_VBFHinv_vv_norm:	0.0%	0.0%
CMS_VBFHinv_vv_xsunc:	0.0%	0.0%
CMS_VBFHinv_qqH_norm:	0.0%	0.0%
QCDscale_qqH:	0.0%	0.0%
pdf_qqbar:	0.0%	0.0%
CMS_VBFHinv_ggH_norm:	0.0%	0.0%
QCDscale_ggH2in:	0.0%	0.0%
pdf_gg:	0.0%	0.0%
UEPS:	0.0%	0.0%

[2] CMS Collaboration, "Measurement of WW production rate", CMS Physics Analysis Summary CMS-PAS-SMP-12-013, (2012).

[3] CMS Collaboration, "Measurement of the pp→ZZ production cross section and constraints on anomalous triple gauge couplings in four-lepton final states at $\sqrt{s}$ =8 TeV", CMS Physics Analysis Summary CMS-PAS-SMP-13-005, (2013).

[4] A. L. Read, "Presentation of search results: the CLs technique", *J. Phys. G: Nucl. Part. Phys.* **28** (2002) 2693, doi:10.1088/0954-3899/28/10/313.

[5] T. Junk, "Confidence level computation for combining searches with small statistics", *Nucl. Instrum. Meth.* **A434** (1999) 435, doi:10.1016/S0168-9002(99)00498-2.

[6] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, R. Tanaka (Eds.), "Handbook of LHC Higgs Cross Sections: Differential Distributions", CERN Report CERN-2012-002, (2012).

A Trigger Efficiency Fits

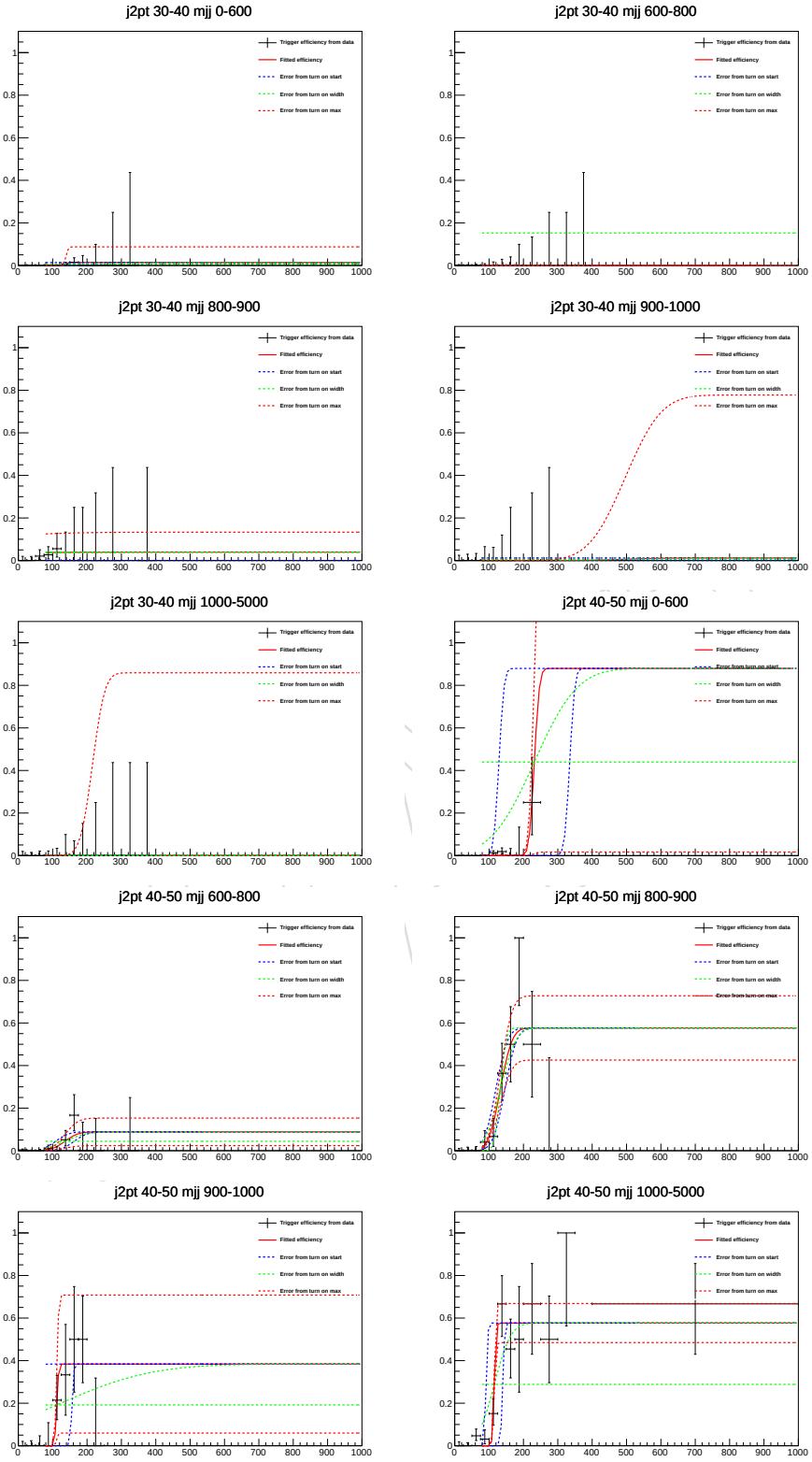


Figure 30: The measured efficiency of the trigger used in run A as a function of MET in bins of dijet mass and sub-leading jet p_T

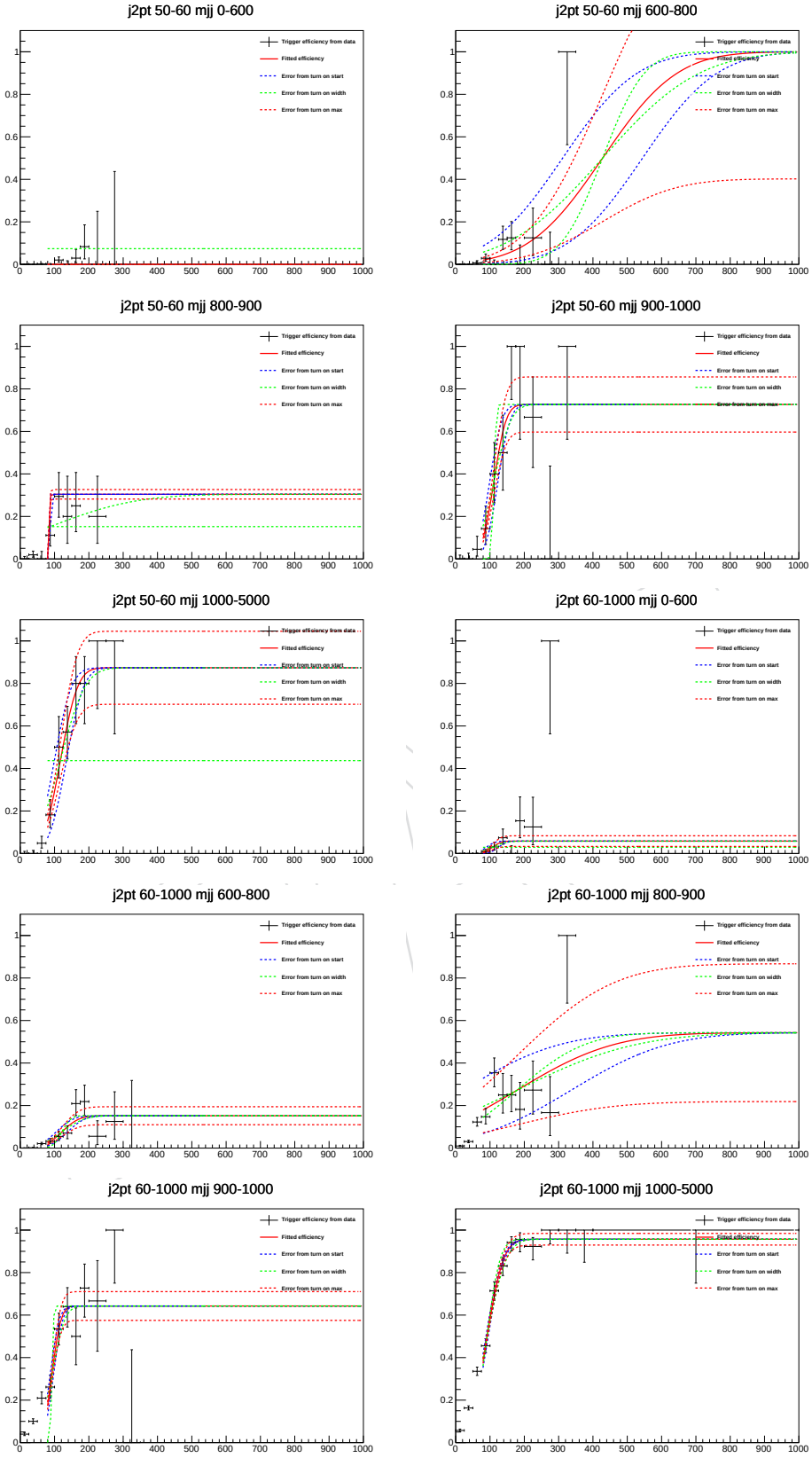


Figure 31: The measured efficiency of the trigger used in run A as a function of MET in bins of dijet mass and sub-leading jet p_T

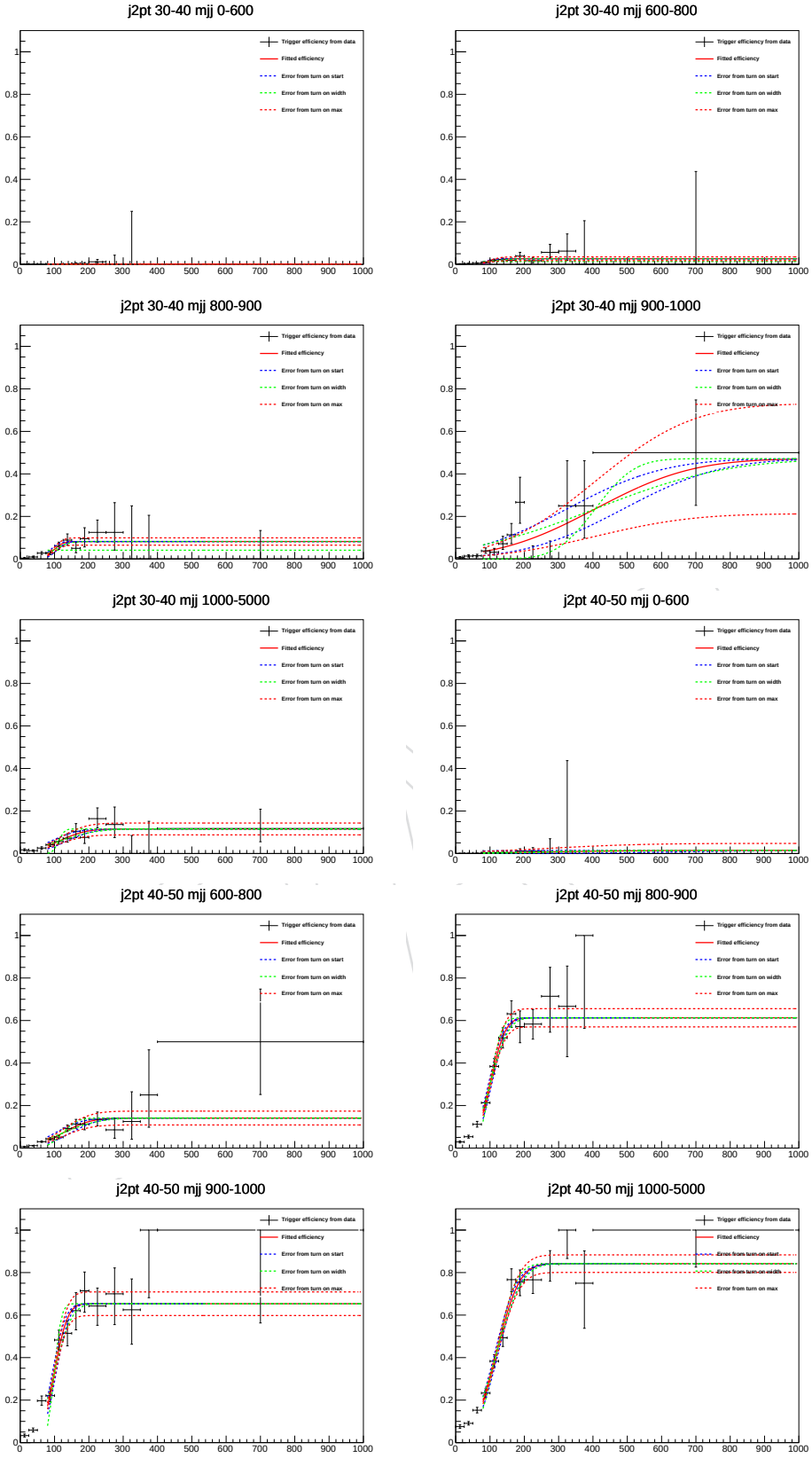


Figure 32: The measured efficiency of the trigger used in runs B and C as a function of MET in bins of dijet mass and sub-leading jet p_T

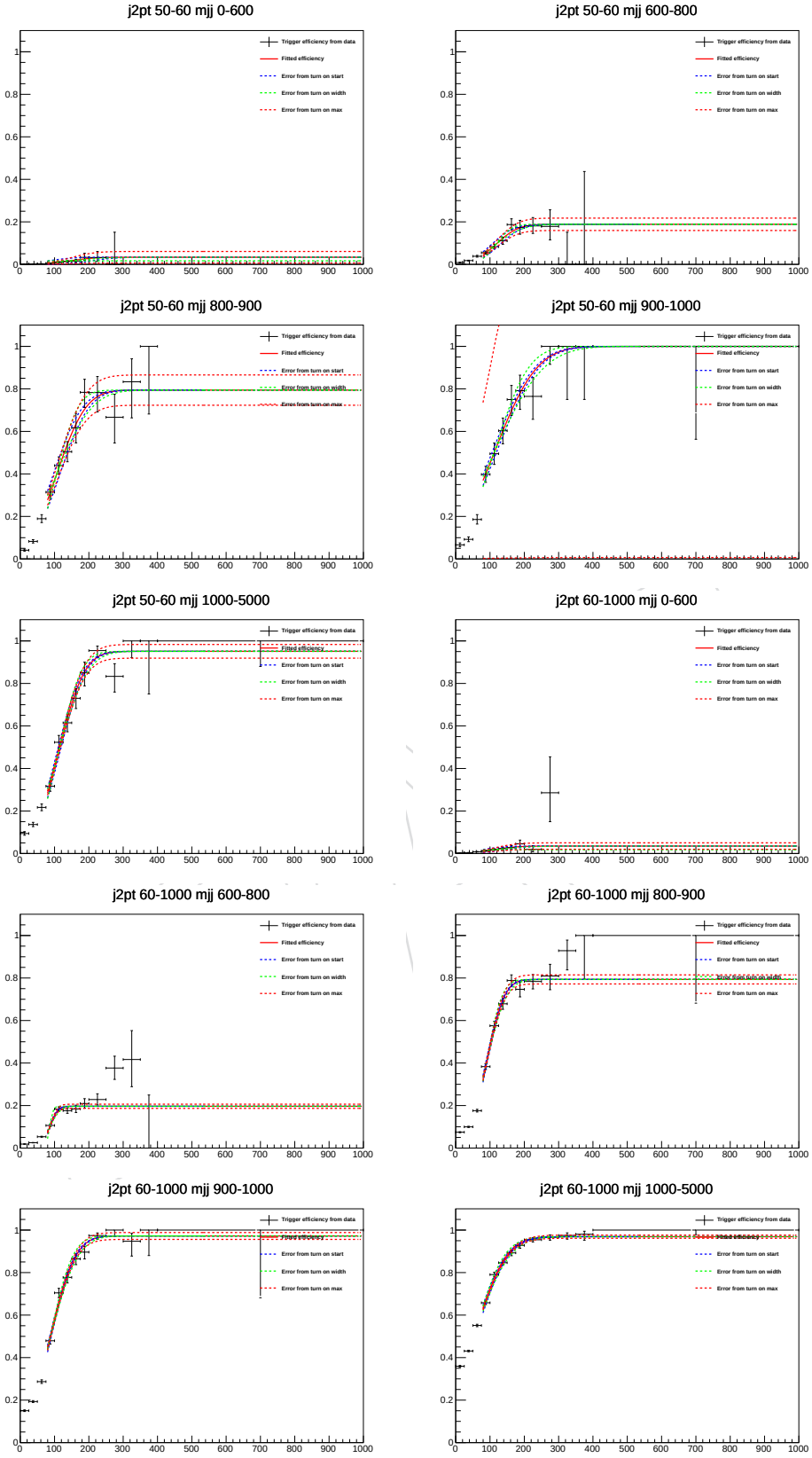


Figure 33: The measured efficiency of the trigger used in run B and C as a function of MET in bins of dijet mass and sub-leading jet p_T

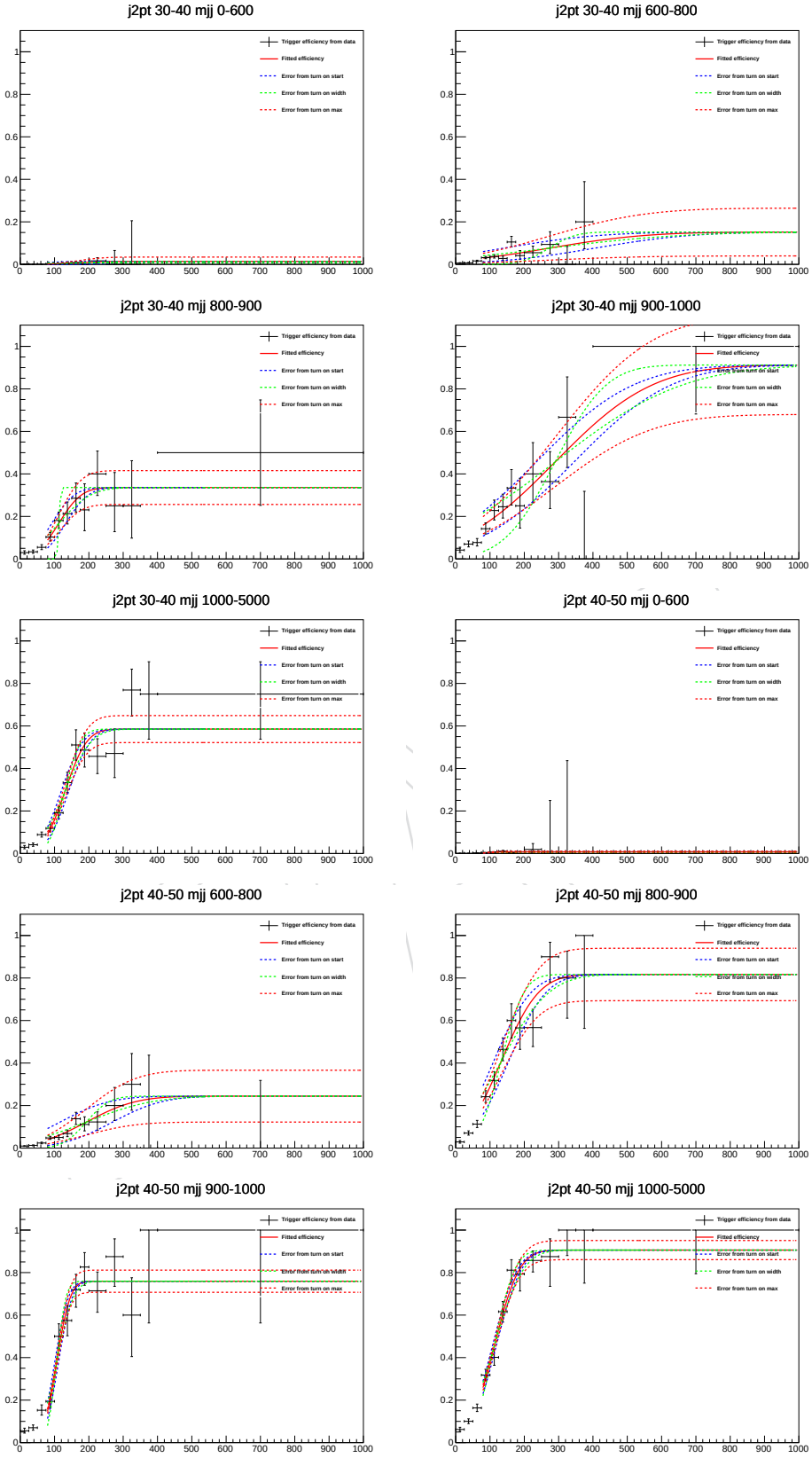


Figure 34: The measured efficiency of the trigger used in run D as a function of MET in bins of dijet mass and sub-leading jet p_T

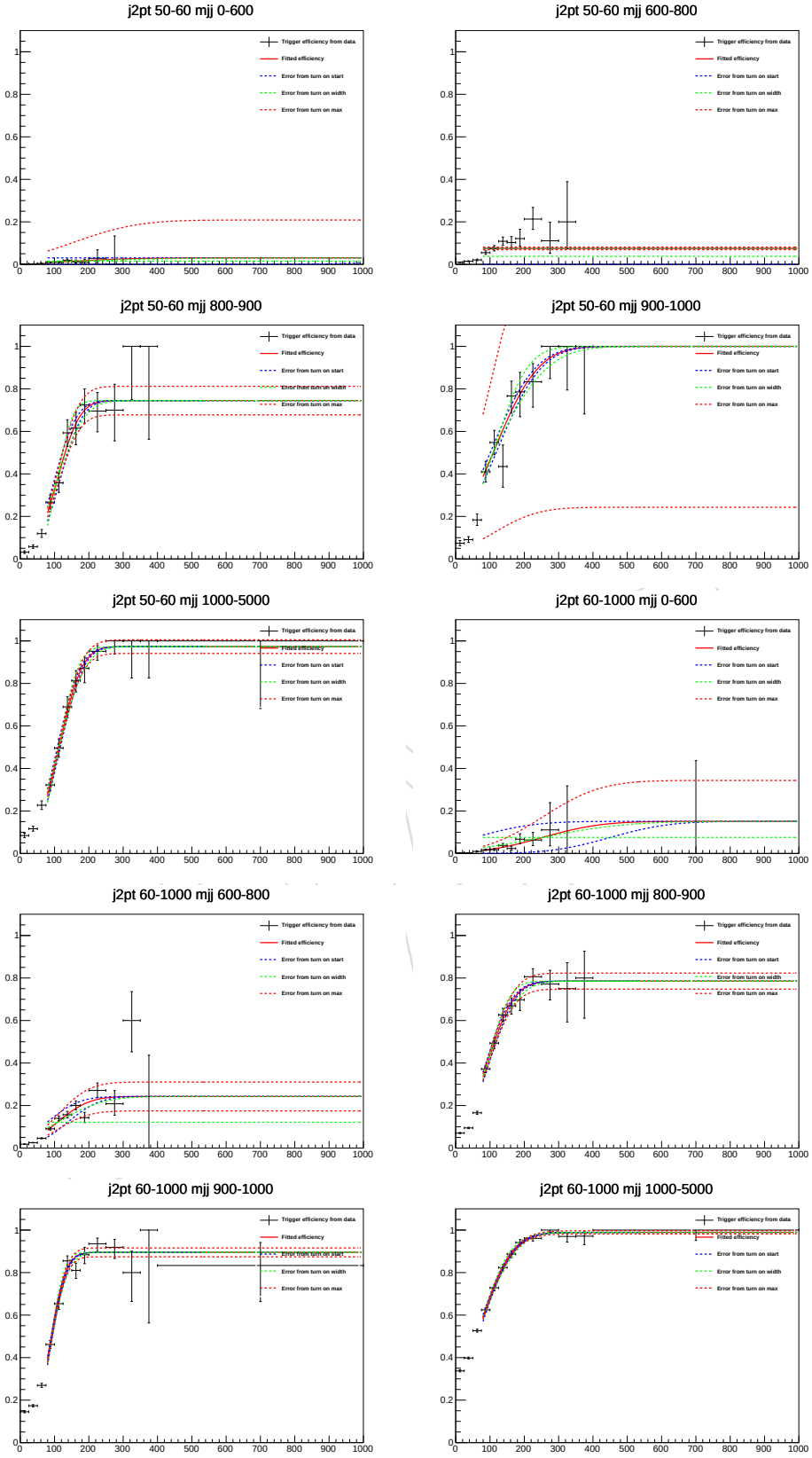


Figure 35: The measured efficiency of the trigger used in run D as a function of MET in bins of dijet mass and sub-leading jet p_T