Observation of a new particle in the search for the Standard Model Higgs boson at the CMS detector

Nicholas Wardle

December 20, 2012

0.1 Introduction

Preamble, declaration of work, description etc…
$\approx$ 10 pages (including list of figs/tables)

Chapter 1

Theory and Motivations

1.1 The Standard Model

Local Gauge theory + SM Lagrangian ≈ 3 pages

1.2 The SM Higgs

1.2.1 The Higgs mechanism

$\approx 2-3$ pages

1.2.2 Constraints and previous searches

$\approx 2-3$ pages Results from LEP and electroweak fits (most recent) Incoude Tevatron searches . . . ?

1.2.3 Higgs production at the LHC

≈ 2 pages

Chapter 2

The CMS Detector

2.1 Detector components

very general description about detector gemetry, componemts ≈ 2 pages

2.1.1 Electromagnetic Calorimeter

≈ 2 pages

2.1.2 L1 Trigger

$\approx 3 - 4$ pages if unclude L1 JEC work (c+p from Internal Note) otherwise remove as a subsection

Chapter 3

Higgs decaying to two photons

The Higgs to two photon channel is one of the most promising decays in the search for the SM higgs at the LHC. Despite having a relatively small branching ratio, the decay $H \to \gamma\gamma$ provides a very clean, fully reconstructable final-state topology, making it one of the most sensitive channels at low mass. The dominant source of background is from real, prompt diphoton events from QCD processes, $pp \to \gamma\gamma$. In addition, there is a contribution from $pp \to \gamma + jet$ and $pp \to jet + jet$ in which jets are mis-identified as photons. As the signal rate in the $H \to \gamma\gamma$ decay is expected to be small compared to the background rates, the search sensitivity is heavily influenced by how well the backgrounds are understood. For this reason, two data-driven background modelling tecnhiques were developed, one in which a fully parametric description of the background from data is devised and one in which a binned model is constructed from sidebands in the $m_{\gamma\gamma}$ spectrum. The latter of these two serves as an independant cross-check of the former building confidence in the understanding of the background. This chapter describes a search for a Higgs boson decaying to two photons which was performed on the full 2011 dataset corresponding to 5.1 fb^{-1}of proton-proton collisions recorded at CMS at a center of mass energy of 7 TeV.

3.1 Data samples

The dataset used for this analysis is the combination of the 2011A and 2011B proton-proton collision runs. The selection for the dataset used for this analysis is based around dedicated diphoton triggers which select events online which satistfy one of two sets of criteria. The first set requires two HLT photon candidates, one with $p_T > 26$ GeV and the other with $p_T > 18$ GeV, which are well isolated in the calorimeter. The second has a lower threshold on the first photon, $p_T > 22$ GeV but requires that both photons have localised showers in the ECAL ($r_9 > 0.8$ in 2011A and $r_9 > 0.9$ in 2011B). Additionally, the invariant mass of the two trigger objects are required to have an invariant mass greater than 60 (70)GeV in the 2011A(B) datasets. Events which would pass the full offline selection but failed to trigger at the HLT lead to an inefficiency, reducing the number of signal events with respect to that expected from an integrted luminosity of 5.1 fb^{-1}. However, the tresholds applied offline are choson to be much tighter than those of the trigger; the trigger efficiency is >99% with respect to the analysis selection.

Signal Monte Carlo (MC) events are generated for a Higgs decaying to two photons via the four main production processes, gluon-gluon fusion, vector boson fusion and assosiated W/Z and $t\bar{t}$ production. The gluon-gluon fusion (ggH) and vector boson fusion (qqH)

Process		**Cross-section** (pb)	**Luminosity** (pb^{-1})
DiPhotonJets		154.7	7400
DiPhoton Box	$\hat{p_T}$ $25 - 250$	12.37	41900
QCD Dijet	$\hat{p_T}$ $30 - 40$	10870	560
	$\hat{p_T}$ $40 - \infty$	43571	920
Gamma+Jet	$\hat{p_T}$ $20 - \infty$	493.44	2400
DrellYan+Jets to ll	$\hat{p_T}$ $50 - \infty$	2475	14000

Table 3.1: Background MC used throughput the analysis with production cross-sections and corresponding equivalent integrated luminsity.

were generated with POWHEG with next-to leading order (NLO) contributions whereas the two associated production processes were generated to leading order (LO) only. The p_T spectrum of the Higgs (p_T^H) from gluon-gluon fusion was calculated at next-to-next-to leading plus next-to leading log resummed order (NNLO+NLL) using the HqT program. The production cross-sections and branching ratios are taken from the LHC Cross-section Working Group.

MC for background processes were generated at LO using POWHEG intefaced with PYTHIA. The QCD dijet and γ + jet samples are filtered by requiring the generated photons, electrons and neutral mesons with $p_T > 15$ GeV have at most one charged particle in a cone, $\Delta R < 0.2$, to increase the production efficiency with respect to the tracker isolation requirements of the full selection. The background samples considered for this analysis are summarized in Table 3.1. A full simulation of the CMS detector is provided in GEANT4 which is used for all signal and background MC samples.

3.2 Object Reconstruction and Identification

The reconstruction of all objects used for this analysis, in both data and MC, is based on the standardized CMS reconstruction software CMSSW_4_2_X. Additional sensitivity can be gained by refining the object selection and reconstruction specifically to the search for $H \to \gamma\gamma$.

3.2.1 Supercluster Energy Correction

As the natural width of Higgs boson is around 100 MeV, the width of a reconstructed mass peak from a $H \to \gamma\gamma$ decay is driven by the experimental energy resolution of the photons. This resolution can be improved dramatically by correcting the raw energy of the supercluster on a per-photon level. These corrections are derived using a multivariate technique in which a regression BDT is trained on prompt photons in the gamma+jet MC sample using the ratio of the generated photon energy to the raw energy of the reconstructed supercluster. As this ratio can vary across different regions of the detector, the input varibles include both the η and ϕ positions of the supercluster. In addition, several variables are included which describe the shower shape: r_9, the energy weighted widths in η and ϕ of the supercluster, the energy weighted crystal width ($\sigma_{i\eta i\eta}$) and the ratio of hadronic energy behind the supercluster to the energy of the supercluster itself (H/E). In the endcap, there is additional information available from the pre-shower.

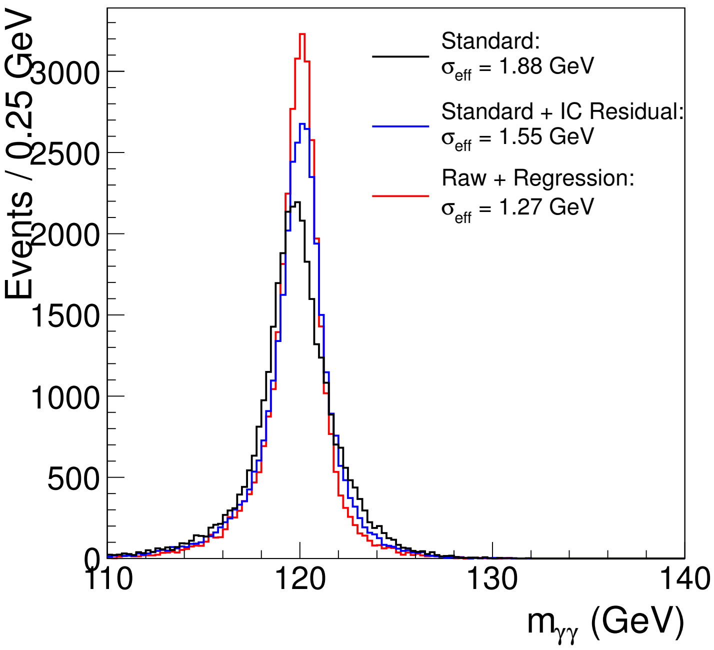


Figure 3.1: Comparison of the diphoton mass peak in MC Higgs with a mass of 120 GeV using different measurements of the photon energy. The black line is from using the raw energy of the supercluster, the blue is from using the analytic fit method and the red from using the regresssion method. The quantity σ_{eff}, the narrowest range in $m_{\gamma\gamma}$ which contains 68% of the distribution, is given for each peak.

The ratio of the energy in the pre-shower to the raw supercluster energy is included for superclusters in the ECAL endcap. Figure **??** shows the improvement in resolution after applying the regression corrections compared to the raw measurement. In addition, a similar set of corrections were derived using by fitting an analytical expression of the residual energy difference between the generated and reconstructed photon energy as a functon of supercluster energy, position and r_9. The regression technique reduces the effective resolution of the Higgs mass peak (σ_{eff}) resolution by around 30% over using the raw supercluster energy compared to the analytic fit which improves the resolution by 15%.

An estimate of the per-photon energy resolution, σ_E, is obtained by training a second regression BDT targetting the the absolute deviation between the correction estimated by the first BDT and the true correction to generator level. This second BDT is trained on an independant set of events to the first. The per-photon resolution is used to calculated an estimate of the per-event mass resolution, $\sigma_{m_{\gamma\gamma}}$, which is used during the event selection (Section 3.3). An additional regression BDT is trained on Zee MC which is used to compare the supercluster energy scale in data and MC.

Energy Scale Measured in Data

Despite correcting the energy of the photons using the regression technique, discrepancies between data and MC are still observed. This is due to additional detector effects which may not be simulated, such as the time dependance of the ECAL crystal transparency.

Further corrections are derived based on $Z \rightarrow e^+e^-$ data which provides an invariant mass peak with almost no background constructed from electromagnetic objects which are reconstructed using a similar procedure to photons. The energy scale of the superclusters is measured by matching the electron invariant mass peak in data to that in MC. This is acheived using an analytic fit to the $Z \rightarrow e^+e^-$ peak in data and MC separately. The natrual peak of the Z is described using a Breit-Wigner distribution whose parameters are fixed to those given by the Particle Data Group (PDG REF), $m_Z = 91.188$, $\Gamma_Z = 2.495$. This is then convoluted with a Crystal Ball (CB) which describes the resolution effects of the calorimeter and energy losses from bremsstrahlung before the ECAL. The CB parameters, Δm ..., are the free parameters of the fit.

The values of these fitted parameters varies with the potition of the supercluster ($|\eta|$). Moreover the variation in data is strongly dependant on the run during which the data were taken. The scale is extracted in six run ranges and four $|\eta|$ regions to account for this effect. The difference between MC and data with time is less dependant on whether the electron showered or not which is characterised by the r_9 of the supercluster. The data-MC difference in each $|\eta|$ region is measured a second time after applying the first set of corrections to the data and obtaining the residual difference for electrons with $r_9 < 0.94$ and $r_9 > 0.94$ separately. The final energy scale correction is then defined as the product of the two corrections. The relative correction,

$$1 - \Delta P = 1 - \frac{\Delta m_{data} - \Delta m_{MC}}{m_Z} \tag{3.1}$$

is applied to the photons in data. The values for the scale in each category are given in Table **??**. The uncertainties on these measurements are primarily due to the difference in the r_9 distribution of electrons and photons. In addition, smaller systemaitcs are included due to the variation of the measurements when changin the electron selection and between using the electron-trained and photon-trained regression corrections. These uncertainties are incorperated into the signal model for the purposes of signal extraction as described in Section 3.4.5.

3.2.2 Vertex Selection

The assignment of the correct vertex to the diphoton pair is an important step in the reconstruction of its invariant mass. Since photons do not leave tracks, computing the angle between the two photons depends strongly on determining the interaction in which they were produced. Figure **??** shows the invariant mass distributions from a SM Higgs for events in which the vertex selected is within 10mm of the generated vertex compared to those in which an incorrect vertex is assigned.

A BDT was trained to rank the standard collection of reconstructed verticies. The input variables are chosen to exploit the correlation between the diphoton system and the recoiling tracks. These are the p_T-balance and p_T-asymmetry calculated as,

$$-\sum_{alltracks} \left(\mathbf{p}_T^{track} \cdot \frac{\mathbf{p}_T^{\gamma\gamma}}{p_T^{\gamma\gamma}} \right) \tag{3.2}$$

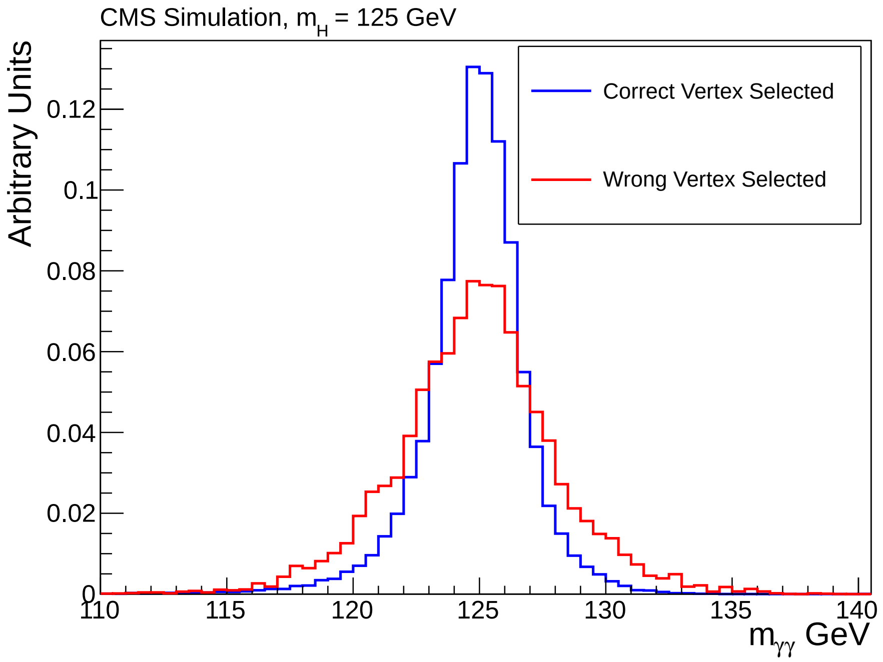


Figure 3.2: Invariant mass peak in $H \to \gamma\gamma$ MC with mass 125 GeV. The blue histogram is from events in which the generated vertex is within 10mm of the vertex assigned to the diphoton pair. The red histogram is from events in which the incorrect vertex is assigned. Both distributions are normalised to unit area for ease of comparison.

and

$$\frac{|\sum_{alltracks} \mathbf{p}_T^{track}| - p_T^{\gamma\gamma}}{|\sum_{alltracks} \mathbf{p}_T^{track}|} \tag{3.3}$$

repsectively. In addition, the sum of the square of the transverse momenta of all the tracks associated to a given vertex is included to preferentially select hard interactions. If at least one of the photons converts to an e^+e^- pair, the difference between the position in z as calcuated using the electron-positron pair and that from the standard vertex, relative to the resolution in z is included as an input variable. The BDT was traind on $H \to \gamma\gamma$ MC with a mass of 120 GeV. Figure 3.2.2 shows the fraction of events in a gluon-gluon MC sample in which the vertex with the highest BDT score is within 10mm of the true vertex as a function of p_T^H.

The fraction of events in which this occurs in data is measured using $Z \to \mu^+\mu^-$ events as a function of the p_T of the Z boson. This is used to correct the Higgs signal MC for the purpose of signal modelling. A second, per-event BDT is trained using the output of the first, to identify under which conditions the correct vertex is selected. The output of this BDT is then used to calculate the probability in a given event that the correct vertex is assigned. The red line in Figure **??** shows a comparison of the per-event vertex probability estimated from the second BDT agaisnt the fraction of the events in which the selected vertex is located within 10mm from the true vertex.

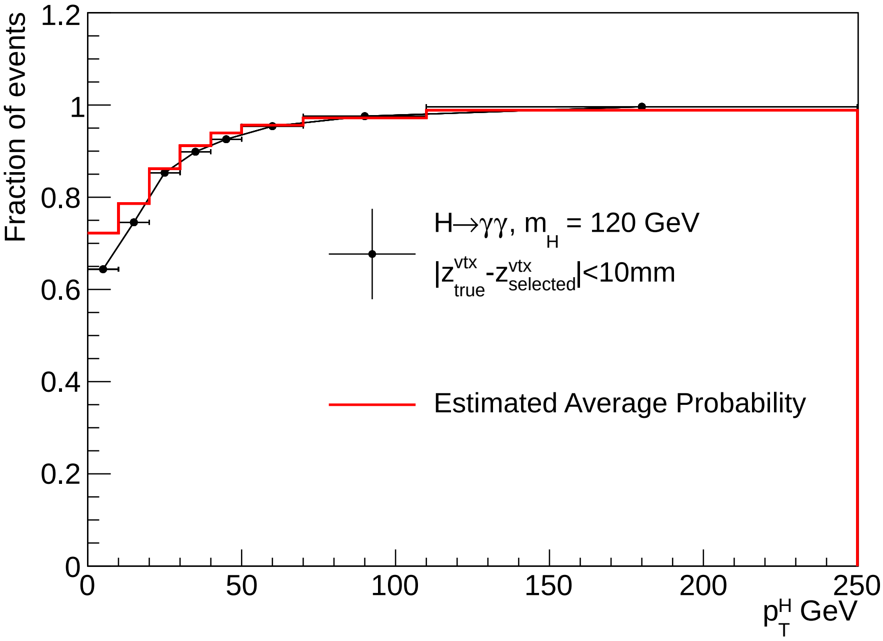


Figure 3.3: Fraction of simulated gluon-gluon fusion events in which the selected vertex z position is within 10mm of the true vertex as a function of Higgs p_T. The red histogram is the average probability to select the correct vertex in each bin estimated from the per-event BDT.

3.2.3 Photon Identification

A large portion of the fake background in the $H \to \gamma\gamma$ search is due to high momentum neutral mesons which decay to two photons where both the photons are combined into the same supercluster. Information from the shower shape of the photon supercluster can be used, as well as the energy isolation within the calorimeter, in order to distinguish these from real photons from the primary interaction point. A BDT was trained on MC samples to combine the relavant information into a single photon identification (ID). The signal used for the training was taken from Higgs decaying to two photons MC with mass 121 GeV while the background was taken from non-prompt photons in the Gamma+Jet sample. Before training, events are required to pass a loose preselection designed to avoid training where the MC is unable to properly describe the data and to match the variables used in the trigger. In addition, photon candidates are removed if there is a reconstructed `GsfElectron` matched to the photon supercluster with no matching conversion reconstruction. This greatly reduces the contribution from $Z \to e^+e^-$ faking photons. The same preselection is applied to all MC and data for extracting the signal. The efficiency of the preselection for signal was measured in $Z \to e^+e^-$ data and MC using a tag-and-probe method (REF). The results are shown in Table 3.2.3.

The input variables are chosen to be insensitive to the kinematics of the diphoton system itself including the diphoton invariant mass. The first set of variables describe the shower shape of the supercluster: H/E, $\sigma_{i\eta i\eta}$, r_9 and the energy weighted widths of the supercluster in η and ϕ (σ_η, σ_ϕ). The η of the supercluster is included as the shower shape is dependant on the position within the calorimeter. The second set of

Category	Data	MC	Data/MC
EB $r_9 > 0.9$	0.9267 ± 0.0012	0.9275 ± 0.0006	0.999 ± 0.0013
EB $r_9 < 0.9$	0.8882 ± 0.0023	0.9025 ± 0.0010	0.984 ± 0.0025
EE $r_9 > 0.9$	0.9442 ± 0.0010	0.9387 ± 0.0009	1.006 ± 0.0014
EE $r_9 < 0.9$	0.8639 ± 0.0010	0.8517 ± 0.0011	1.014 ± 0.0015

Table 3.2: Signal efficiency for the preselection measured in data and MC using tag-and-probe in $Z \to e^+e^-$ events. The ratio Data/MC are applied as corrections to the signal MC for the purposes of signal modelling. The uncertainties listed here are statistical only.

input variables describe the isolation of the photon in the calorimeter and tracker scaled to account for the additional expected energy density due to pileup, ρ. These are the sum of the track isolation, calculated relative to the chosen vertex and the vertex giving the maximum track isolation, ECAL isolation and HCAL isolation in a cone with $\Delta R < 0.3$ minus ρ times the effective area 0.17 and the absolute ECAL and HCAL isolations within cones of $\Delta R < 0.3$ and $\Delta R < 0.4$ respectively. In addition, the number of reconstructed vertices in the bunch crossing is included to reduce the pileup dependance of the isolation variables.

A separate BDT is trained for application in the ECAL barrel and endcaps as the shower shape and isolation variables are rather distinct between the two components. A cut is made on the photon ID BDT output to select events used for the signal extraction which keeps practically all (> 99%) of the signal while removing around 22% of background events. The cut is choson to be loose as the output of the photon ID will be used as input for the event selection (diphoton BDT) desribed in Section 3.3.1,

3.3 Event Selection

In addition to passing the preselection, the two photons are required to pass mass-dependant transverse momenta cuts, $p_T/m_{\gamma\gamma} > 1/3, 1/4$ for the leading and subleading p_T photon respectively. Where more than one diphoton pair satifies these criteria in an event, the pair which has the largest sum of photon transverse momenta is selected as the Higgs candidate. The final selection of diphoton candidates used for the signal extraction is based on using as much information in the event as possible to distinguish likely signal candidate events from the background. Although the photon ID BDT is successful at rejecting fake backgrounds, a large portion of the background is due to real prompt photons from QCD processes. In order to distinguish these from a Higgs signal, the specific kinematics and toplogy of the event are exploited.

3.3.1 Diphoton BDT

A BDT was trained to utilise the kinematics of the selected diphoton pair to discrimiate prompt photons from QCD background from those produced by the decay $H \to \gamma\gamma$. The BDT was trained using the QCD Dijet, Gamma+Jet, DiphotonJets and Diphoton Born samples for background and Higgs MC with a mass of 123 GeV. As the mass of the Higgs boson is unkown, the search is performed under different mass hypotheses. In order to allow for the application of the same selection to the data under any mass hypothesis, the input variables to the BDT are chosen to be mass-independant. In addition, this philosophy allows for a fully data-driven estimation of the background shape as described in Section **??**. These input variables which describe the kinematics are: the relative transverse momenta of the photons, p_T^1, p_T^2, their pseudo-rapidites, η^1, η^2 and the cosine of the angle between the two photons in the transverse plane $cos(\Delta\phi) = cos(\phi^1 - \phi^2)$. In addition, information regarding the quality of the objects, the two photons and the selected vertex, is included in the form of the output of the photon ID and the vertex probability. The per-photon resolution estimate, σ_E is combined for each photon to produce a per-event mass resolution estimate $\sigma_{m_{\gamma\gamma}}$ under the assumption that the correct vertex,

$$\sigma_{m_{\gamma\gamma}}(right - vtx) = \frac{m_{\gamma\gamma}}{2}\sqrt{\left(\frac{\sigma_E^1}{E^1}\right)^2 + \left(\frac{\sigma_E^2}{E^2}\right)^2} \tag{3.4}$$

where E^1, E^2 are the energies of the two photons.

Since the correct vertex is not always chosen, the mass resolution assuming the incorrect vertex is chosen is calculated using the average beamspot length in data, $\sigma_Z = 5.8cm$. In this case, the distance between the selected and true vertex will be distributed as a Gaussian with width $\sqrt{2}\sigma_Z$. The contribution to the resolution, $\sigma_{m_{\gamma\gamma}}^{vtx}$ can be calculated analystically given the positions of the two photons. The mass resolution estimator under the assumption that the incorrect vertex is chosen is given by the sum in quadrature of $\sigma_{m_{\gamma\gamma}}^{vtx}$ with the mass resolution assuming the correct vertex is chosen. Both estimators for the mass resolution relative to the invariant mass, $\sigma_{m_{\gamma\gamma}}/m_{\gamma\gamma}\ right/wrong - vtx$, are included as inputs to the diphoton BDT. Figure 3.4 shows the diphoton BDT distribution in data and MC. In addition to further separating the contribution to the background from fakes, the diphoton BDT can discriminate between prompt diphotons in QCD and those from a $H \rightarrow \gamma\gamma$ decay. The final events used for the signal extraction are selected as those with a diphoton BDT output greater than 0.05. This cut is chosen following an optimization study to minimize the expected exclusion limit in the absence of signal. Events below this cut value were found to provide negligible improvement in the expected limit.

Figures 3.3.1 and 3.3.1 show the input variables from the final set of selected diphoton candidates in data and MC. The expectation in each plot from a SM Higgs with a mass of 125 GeV, scaled by 10, is shown in red.

Figure 3.7 is the invariant mass distribution in data and MC for events passing the full selection. The expected peak from a Higgs with mass 125 GeV, scaled by 10, is shown in red. After the application of the full selection, the total background contains 76% prompt diphoton events.

Diphoton BDT Validation with $Z \rightarrow e^+e^-$ Data

By using a BDT for the full event selection, subtle correlations between the input variables are acounted for which improves the separation between the signal and background. Unlike the background model, the signal model will be taken from corrected MC simulation. It is important therefore to ensure that the BDT will respond in the same way in data as for the signal MC used for the signal extraction. The MC can be validated using $Z \rightarrow e^+e^-$ data-MC comparisons by inverting the electron veto and treating the electrons as though they were photons. This is done by using the supercluster associated to the electron for the electron's energy measurement and ignoring the track information. In this way, the reconstruction of the electrons is the same as that of the photons allowing

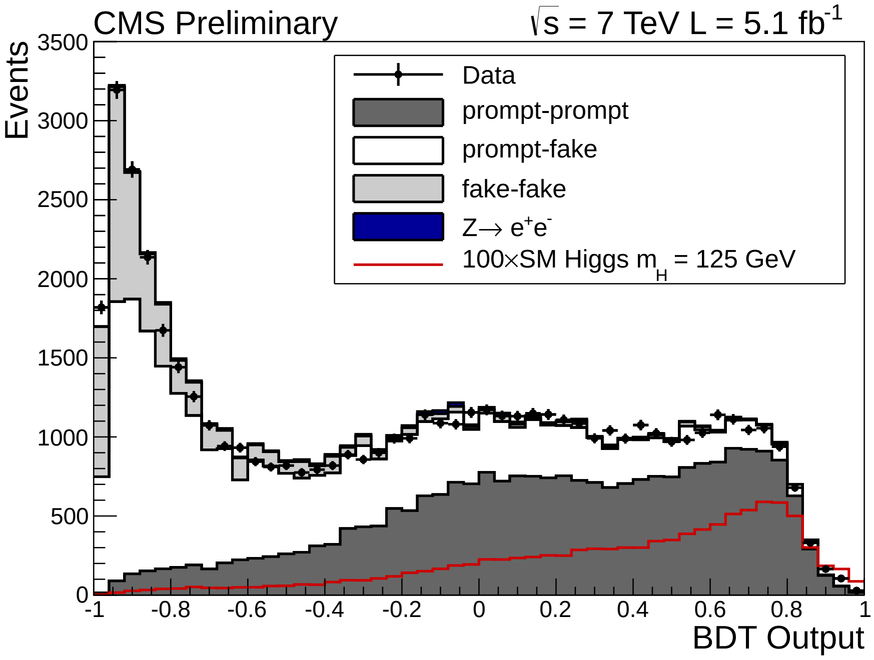


Figure 3.4: Diphoton BDT distribution in data and MC. The contribution expected from a SM Higgs with mass 125 GeV, scaled by 100, is shown in red.

for validation of the BDT's response to real photons from a resonance decay. Figure **??** shows the diphoton BDT distribution in $Z \to e^+e^-$ MC and data after applying the full selection using this technique.

Both the photon ID and regression BDT's rely on a correct simulation of the shower simulation in MC to correctly describe the data. Due to impoerfections of this simulation, systematic uncertainties are included in the signal model to cover the residual difference observed between MC and data in a high p_T photons as descibed in Sections **??** and **??**. These uncertainties are validated using $Z \to e^+e^-$ in the same way as the diphoton BDT. Figures 3.3.1 and **??** show the distributions of the per photon energy resolution estimator σ_E relative to the photon energy and the output of the photon ID MVA (BDT output) in $Z \to e^+e^-$ MC and data treating the electrons as photons. The red lines show the $\pm 1\sigma$ error envelopeattributed to the systematic uncertainty on the shower simulation. These uncertainties are propagated through the diphoton BDT and included in the signal model as described in Section 3.4.5.

3.3.2 Dijet Tagging

The contribution to Higgs production from vector boson fusion (qqH) is around a factor ten smaller than that of gluon-gluon fusion. However, additional information from the two jets associated with the qqH production allows for further reduction of the diphoton background. Events which pass the full selection and in addition satisfy a series of criteria designed to target the specific topoloy of the dijet system are tagged as likely to have originated from qqH production. Figure 3.3.2 shows the separation in η between the two jets. Signal events from vector boson fusion production are more likely to have a large separation than those from background processes. The full set of criteria is given

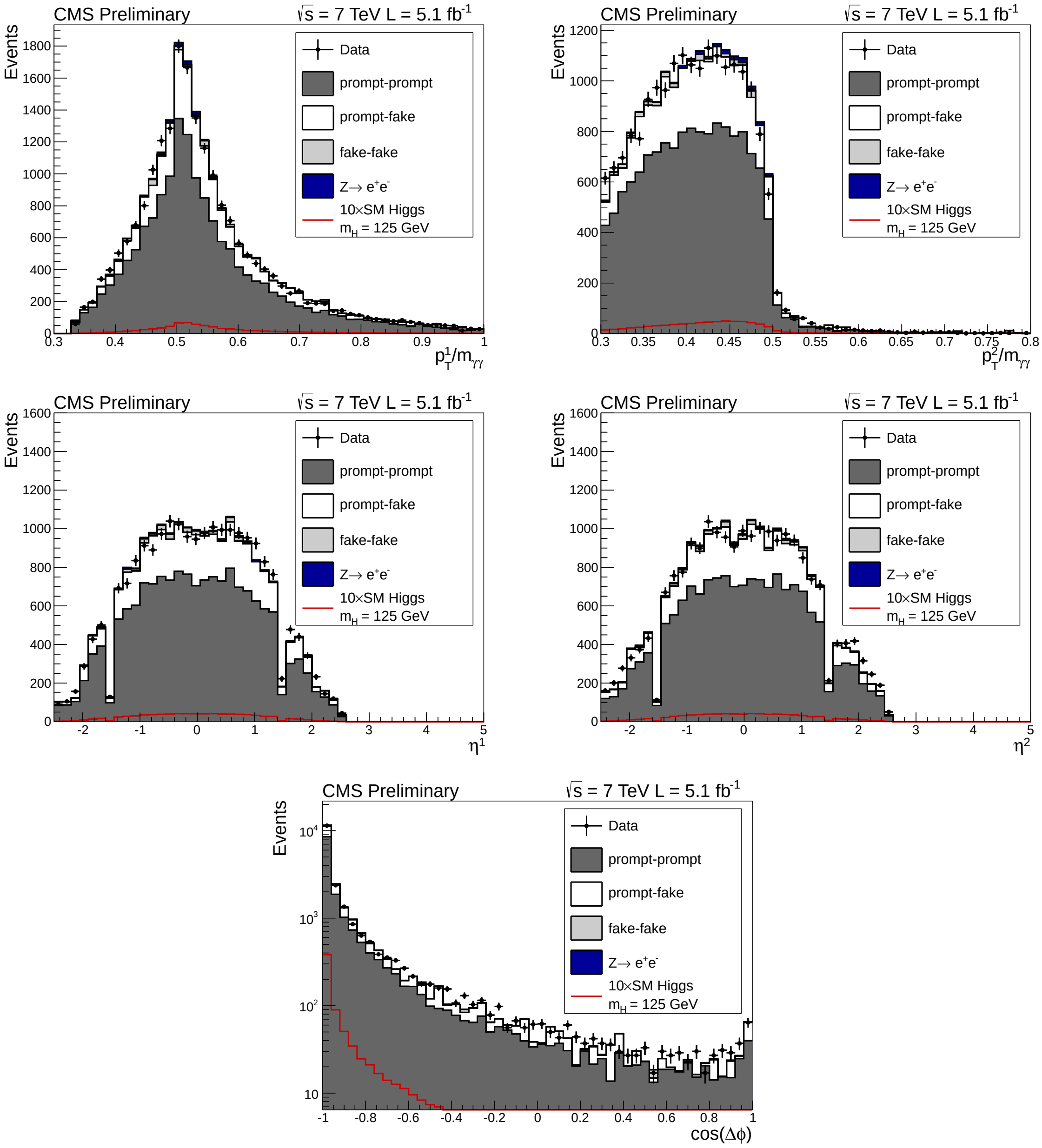


Figure 3.5: Kinematic diphoton BDT input variable distributions in data and MC. The distributions are for events which pass the full selection including a cut on the diphoton BDT output of 0.05. The expectation from a SM Higgs with 125 GeV is shown in red.

in Table **??**.

The dijet tagged events are categorized separately to the remaining events, thereby exploiting their high signal to background ratio for the purpose of signal extraction.

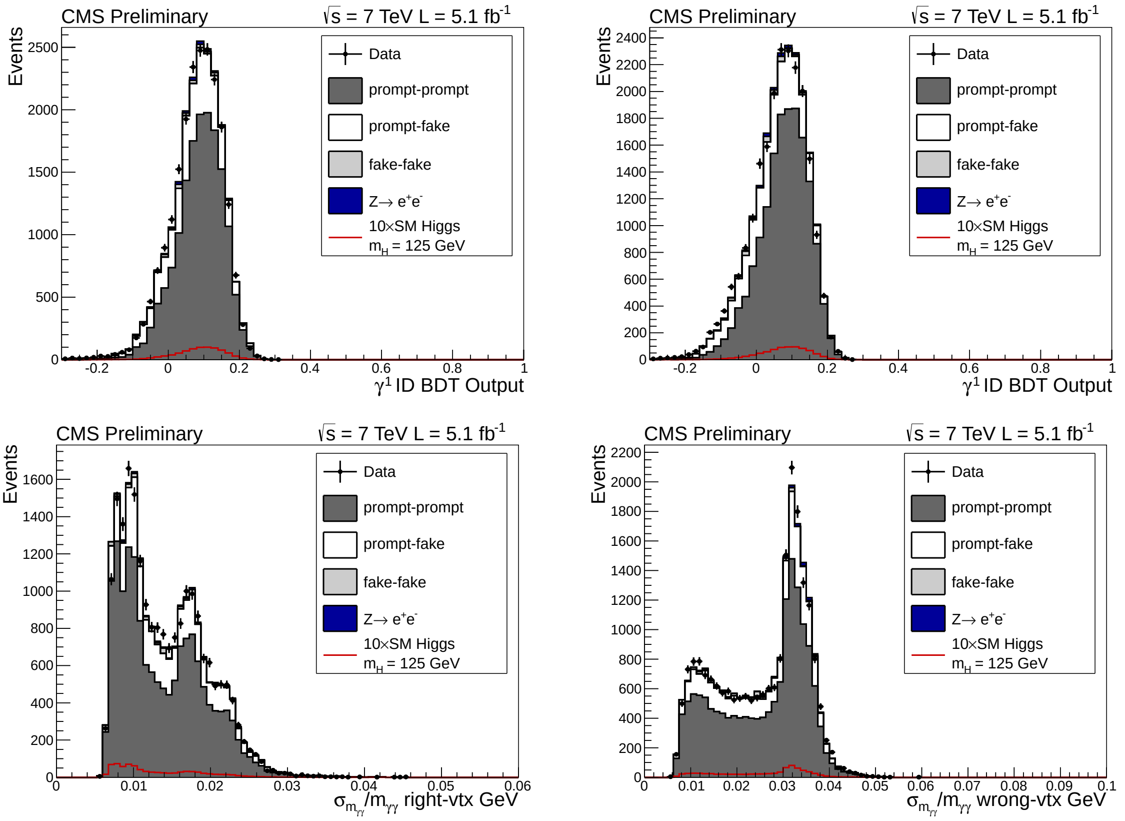


Figure 3.6: Additional diphoton BDT input variable distributions in data and MC. The distributions are for events which pass the full selection including a cut on the diphoton BDT output of 0.05. The expectation from a SM Higgs with 125 GeV is shown in red.

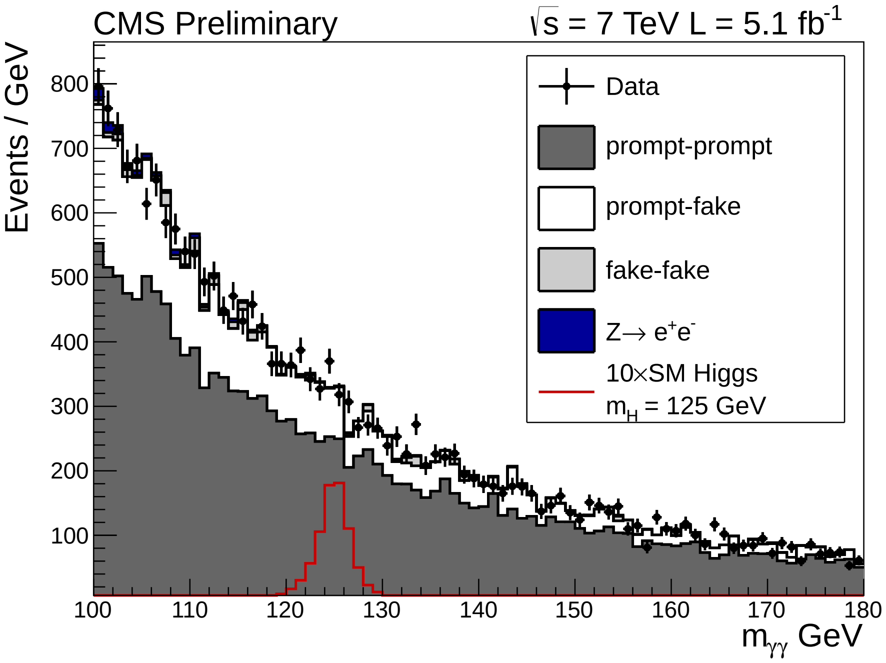


Figure 3.7: Invariant mass distrobution in data and MC after applying the full event selection in the range 100 to 180 GeV. The contribution expected from a SM Higgs with mass 125 GeV, scaled by 10, is shown in red.

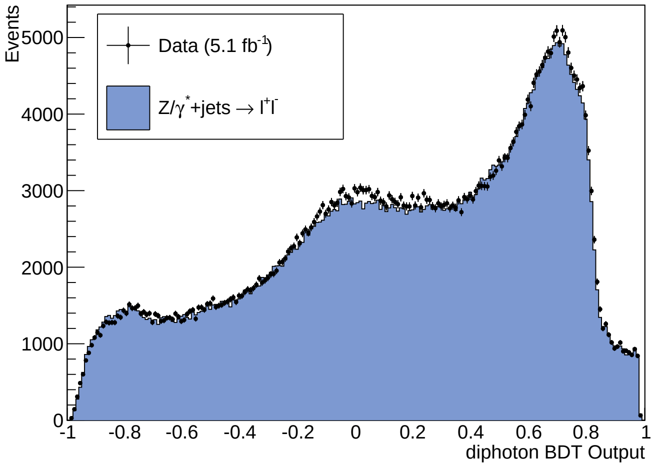


Figure 3.8: Diphoton BDT output distribution in $Z \to e^+e^-$ MC and data after the full selection treating the electrons as photons for the purposes of energy reconstruction. The electron veto is inverted to preferentially select electrons.

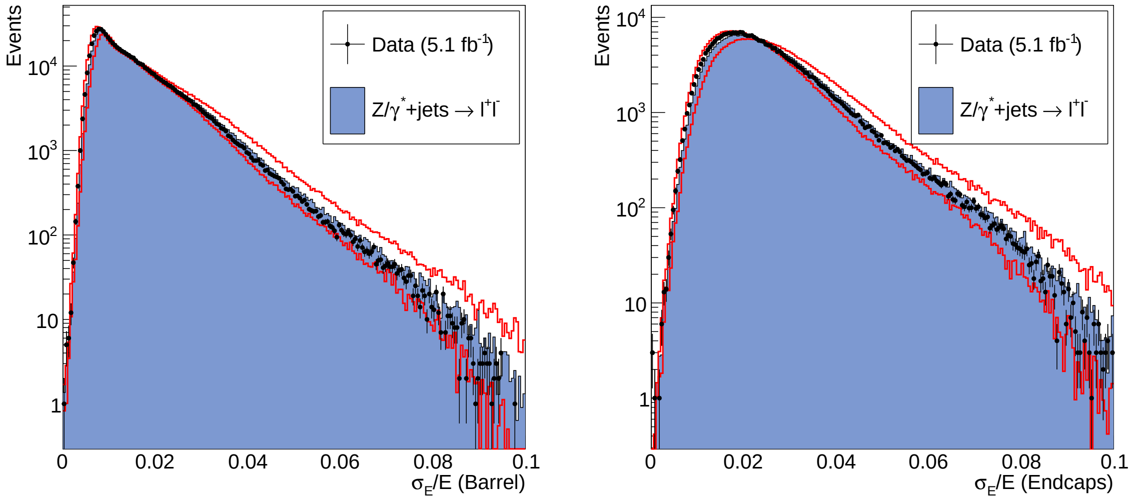


Figure 3.9: Upper: Per-photon resolution estimator, σ_E relative to the measured energy in $Z \to e^+e^-$ MC and data treating the electrons as photons in the barrel (left) and endcaps (right). The red lines show the $\pm 1\sigma$ systematic error envelope obtained by scaling the value of σ_E by $\pm 10\%$.

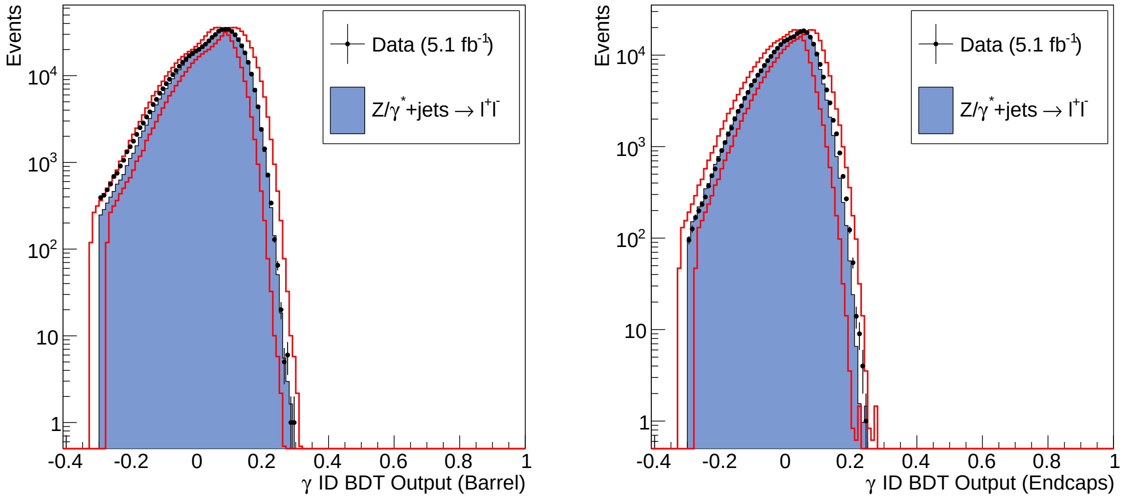


Figure 3.10: Photon ID BDT output in $Z \to e^+e^-$ MC and data treating the electrons as photons in the barrel (left) and endcaps (right). The red lines show the $\pm 1\sigma$ systematic error envelope obtained by shifting the output value by σ_E by $\pm 0.025\%$.

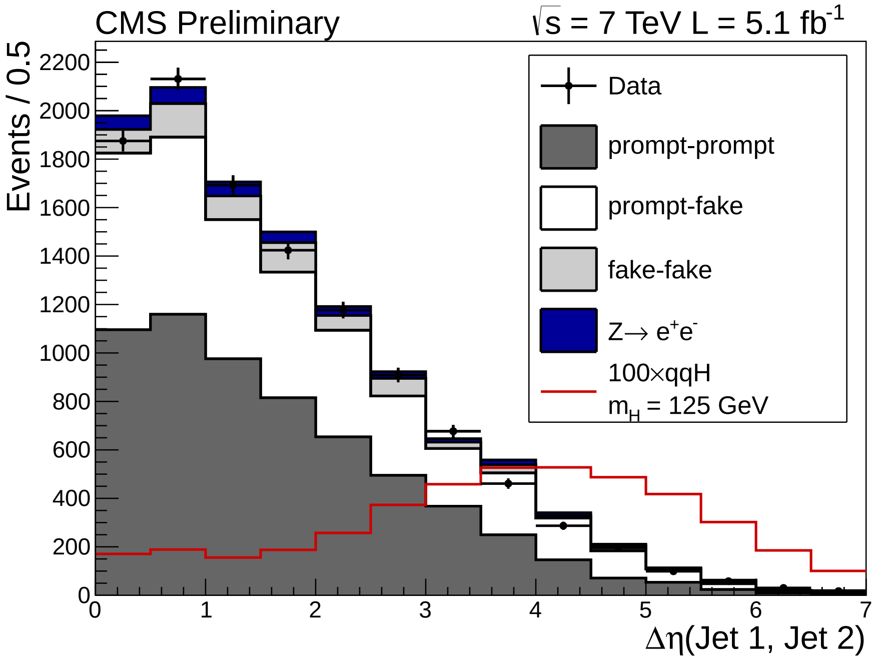


Figure 3.11: Separation in η between two identified jets in data and MC. The expectation from a SM Higgs produced via vector boson fusion (qqH), scaled by 100, is shown in red. All cuts other than the one on $\Delta\eta(Jet1, Jet2)$ are applied to these distributions.

3.4 Signal Extraction

The signature for the decay $H \to \gamma\gamma$ is the presence of a narrow peak on a smoothly falling background in the invariant mass spectrum. The signal to background ratio can be dramatically increased by focusing on events falling in a window around the mass of the Higgs boson, m_H. Since this mass is unconstrained in the Standard Model, the search is performed for a range of mass hypotheses effectively sliding the signal window across the diphoton invariant mass spectrum, $m_{\gamma\gamma}$.

As the signal yield for a SM Higgs decaying to two photons is expected to be small, additional event information from the detector and the kinematics of the diphoton system can be used to increase the sensitivity of the search.

This section describes a MVA based approach to extracting the signal, categorizing events within a sliding signal region window based on a single event discriminator (categorisation BDT). The approach allows for use of data in sidebands to determine expected event yields within the signal region, making little assumption about the specific composition and kinematics of the background.

3.4.1 Definition of the Signal region

Once the expected resolution of the Higgs peak is determined, the choice of signal window can be optimized to reduce the uncertainty on the background while selecting as many signal events as possible. The size of the signal window is chosen using a simplified analysis in which the number of signal events from a SM Higgs with hypothesised mass m_H expected within the range $|\Delta M/M_H| = |(m_{\gamma\gamma} - m_H)/m_H| < w$ is compared to the uncertainty on the total number of events (from background and signal) in that range. The figure of merit, $N_S/\sigma = N_S/\sqrt{\sigma_S^2 + \sigma_B^2}$, is calculated as a function of signal region cut value, w, for a range of mass hypotheses as shown in Figure 3.12 The error on the background is calculated using the procedure defined in whereas the error on the signal is purely statistical.

For this analysis, $w = 0.02$ was chosen as the optimal signal region cut value.

3.4.2 BDT Event Descriminator

The inputs to the diphoton BDT contain information from the event kinematics and the quality of the photons and vertex location in the form of the photon ID MVA output and event resolution estimators. The output of the diphoton BDT combined with the invariant mass of the diphoton system therefore provides the necessary information to separate signal from background.

Figure shows the variation in the signal to background ratio across different regions in the two-dimensional plane defined by the output of the diphoton MVA and $\Delta m/m_H$.

The two variables are combined to produce a single event discriminator by training a BDT using the diphoton BDT output and $\Delta m/m_H$ as inputs. The BDT is trained with Higgs signal MC with $m_H = 123$ GeV including all four production processes and background MC including prompt-prompt, prompt-fake and fake-fake events. The performance of several different training methodologies were compared to find which gave the optimum separation of signal and background. Two different choices of boosting were studied. The first, known as adaptive boosting, reforms decision trees by reweighting events in which the incorrect decision is made initially. The second, known as gradient

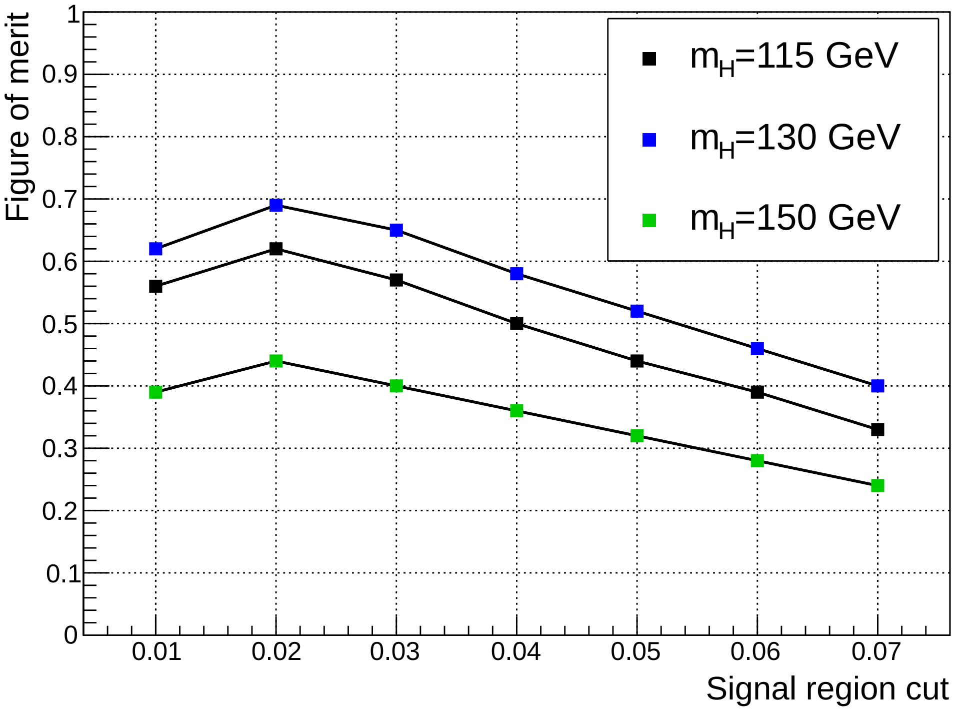


Figure 3.12: Figure of merit for selection of the signal region cut value, w. Each color shows the evaluation under different Higgs mass hypotheses.

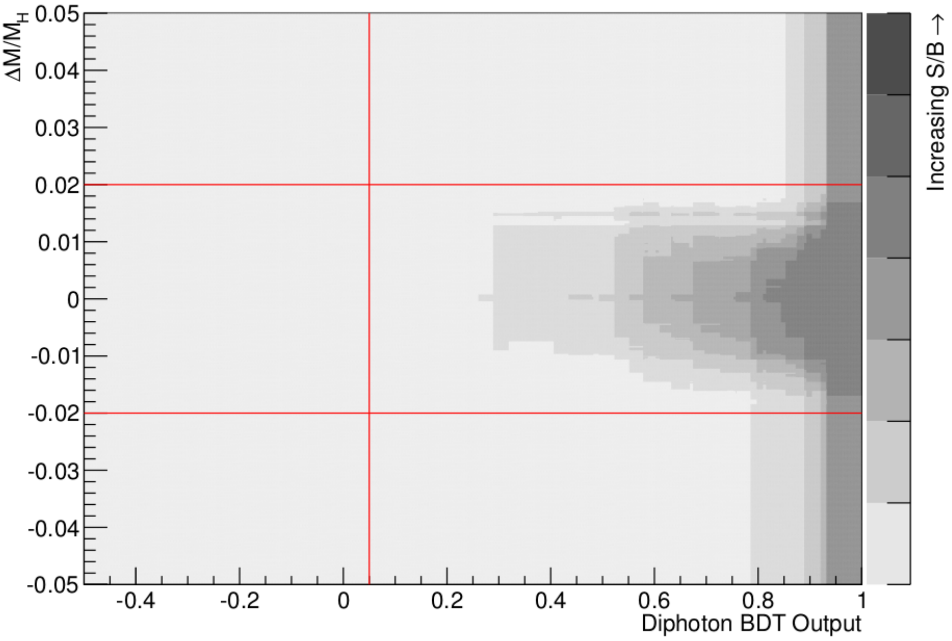


Figure 3.13: Signal to background ratio as a function of diphoton BDT output and $\Delta m/m_H$. The red lines indicate the cuts applied before the training and for applying the event selection.

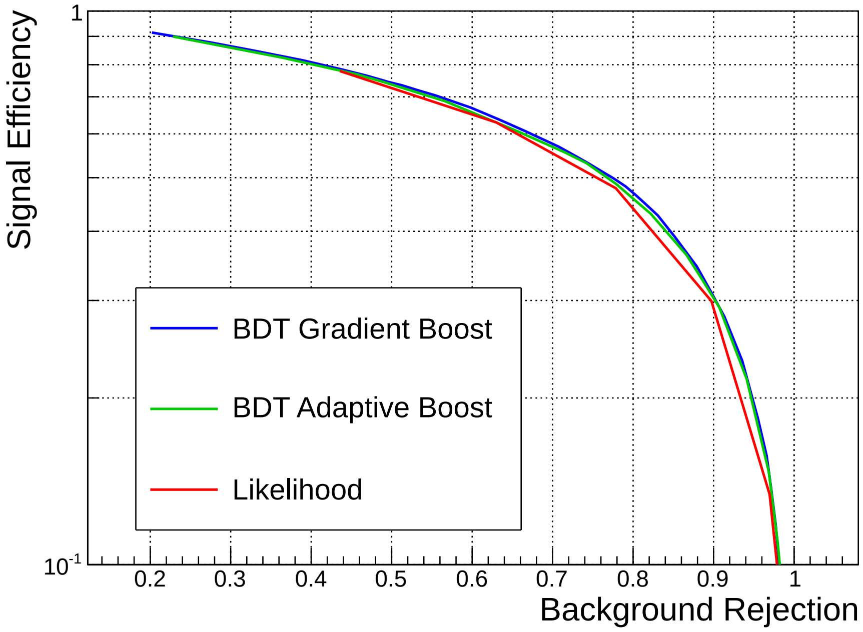


Figure 3.14: Signal efficiency vs background rejection curves for three different MVA techniques used to train the signal-background event discriminator. The curves give the (in)efficiencies for signal (background) after applying sequentially tighter cuts on the discriminator output.

boosting, involves weighting the set of decision trees so as to minimize a log-likelihood. In addition, these were compared to a simple likelihood which does not account for correlations between the diphoton BDT and $\Delta m/m_H$ as shown in Figure 3.14. The gradient boosting method was found to give the best performance although the variation between methodologies is small.

With finite statistics, a BDT can be over-trained by allowing the training to emphasise statistical fluctuations which are not physical and will not necessarily be representative of the data. To test for this, the MC samples are split into two equal samples, the first of which is used to train the BDT. The distribution in the output values of the BDT from the second set are compared to that of the training sample as shown in Figure 3.15. The comparison is shown using both an arbitrary binning scheme and in the final set of bins derived in Section 3.4.3. A χ^2 test was performed on the distributions in the final bins giving p-values of 0.06 for the background and 0.95 for the signal indicating that over-training has not occurred.

In this analysis, the background is estimated entirely from data. This means that any disagreement between data and MC will only effect the performance of the BDT and not the validity of the final results. The agreement between the data and MC is shown in Figure 3.16 for a mass hypothesis, $m_H = 145$ GeV. The level of agreement is sufficient so as not to require in-depth study of the BDT output distributions of the background MC.

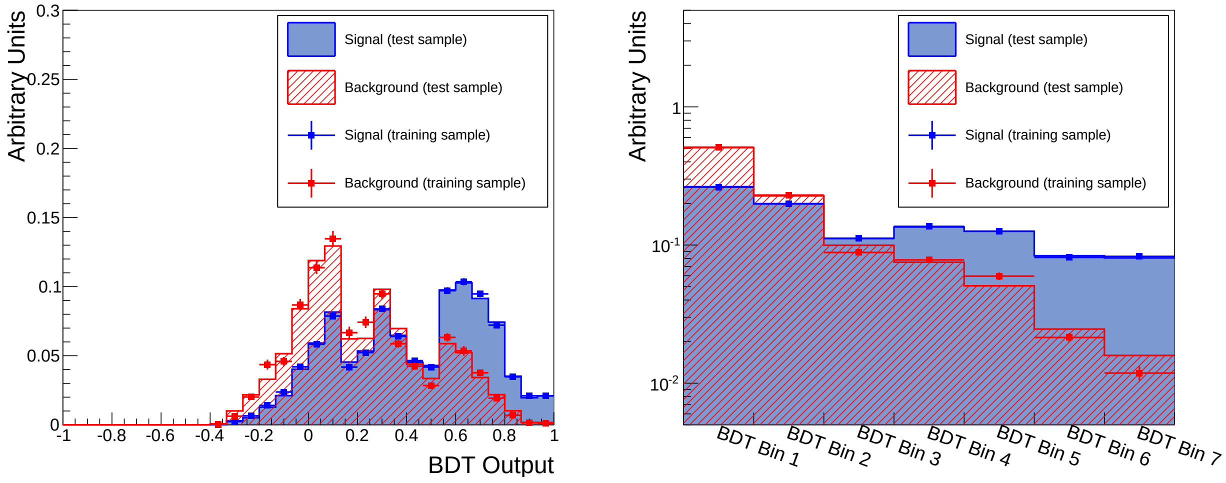


Figure 3.15: Signal and background BDT output distribution with the training sample (points) and testing sample (solid area) superimposed. The comparison is shown using an arbitrary uniform binning (left) and in the bins used for extracting the signal (right).

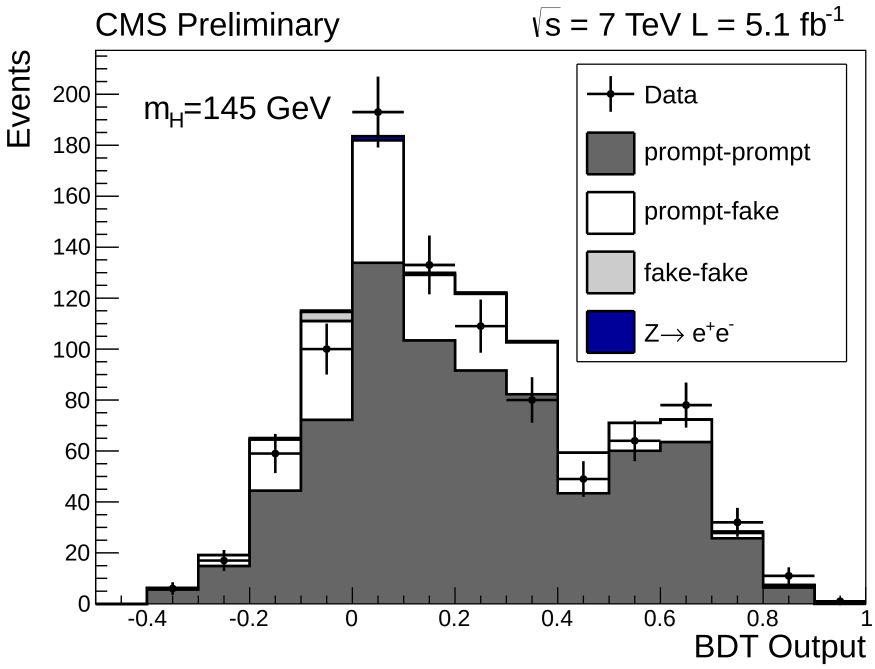


Figure 3.16: Comparison of the distributions of BDT output at $m_H = 145$ for data and background MC. The distributions are arbitrarily binned for the purposes of comparison only.

3.4.3 Binning of the BDT Output Distribution

The BDT provides a single variable with which to classify events based on their signal to background ratio, S/B, which will have a discrete number of response values based on the number of trees used. The boosting procedure provides a pseudo-continuous distribution which is used to model the signal and background. However, the resulting distribution will still be only pseudo-continuous. In addition, the BDT response does not directly correspond to a physical distribution and it is therefore difficult to motivate any parameterisation of either the signal or background distributions. To overcome these issues, a binning procedure is defined to construct templates which are used as models for the signal and background expectation as a function of BDT response range (BDT bin). This procedure is designed firstly to ensure that no bin has zero background expectation and secondly that as few bins as possible are used without reducing the sensitivity of the BDT.

A scan is performed in which the definitions of the bin boundaries are varied in order to find the maximum expected significance in the presence of a SM Higgs signal. For N bins ($N-1$ boundaries) with background and signal expectation yields b_i and s_i respectively, the expected significance is given in Equation 3.5.

$$\frac{\sigma_{exp}}{\sqrt{2}} = \sum_{i=1}^{N}(s_i+b_i)\ln\left(\frac{s_i}{b_i}+1\right) - s_i \tag{3.5}$$

The binning procedure is defined as follows:

1. The distribution of background MC is binned very finely to provide an almost discrete dataset (5000 equally spaced bins are used). The background is re-binned such that there are 20 expected events per bin at a luminosity of 5.1 fb^{-1}. The procedure starts from the highest BDT value bin since the final step bin may have less than 20 events. If that is the case, the last and penultimate bins are combined.

2. Smoothed versions of the signal (at each 5 GeV step mass) and background MC tem- plates are produced in order to obtain a stable model of S/B as a function of BDT bin. The smoothing procedure is done via binning a fit (of a 9th order polynomial) to the signal distribution. Other smoothing techniques were found to give less stable performance.

3. N bin edges (boundaries), b_i, are defined on the remaining bins such that $N+1$ bins are formed with $b_1 < b_2 < \ldots < b_N$. The first bin is defined as $[-1, b_1)$ and the last is defined as $[b_N, 1]$. The N dimensional scan is performed varying these bin edges to find the maximum expected significance in the presence of a SM Higgs signal.

4. An extra boundary is added and the scan is repeated and the maximum expected significance is found for $N+1$ boundaries. If the maximum expected significance is increased by more than 0.1% compared to that of step 3, the new boundary is kept and step 4 is repeated, if not, the procedure terminates.

The scan in step 3 is split into two parts, first using a large step size to find the region where the maximum lies followed by a fine scan in small steps within that region. The ratio of small to large step size is chosen to be that which minimizes the total number

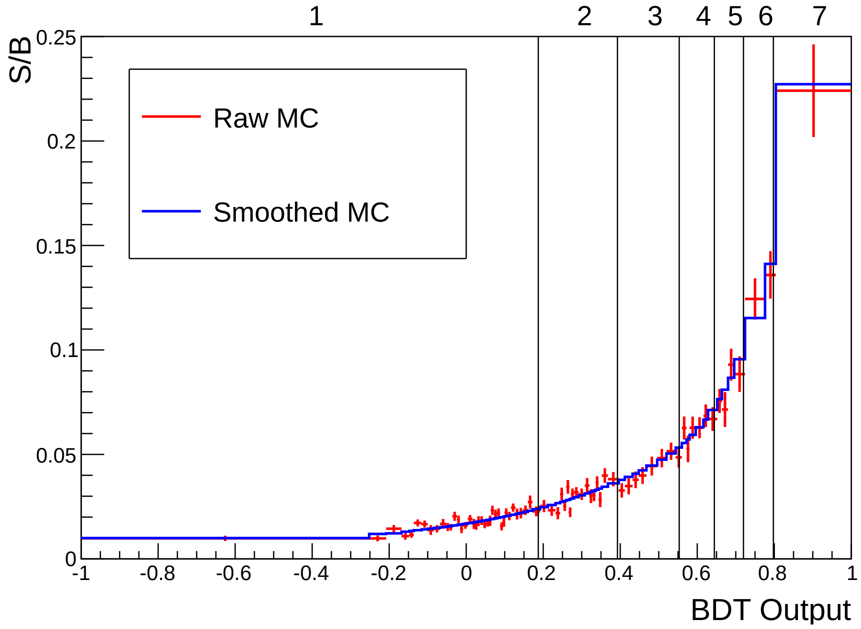


Figure 3.17: Signal to background ratio as a function of BDT output bin. The red and blue histograms show the distribution after applying step 1 of the binning procedure before and after smoothing respectively. The black vertical lines indicate the boundaries of the final binning choice from the full procedure.

of iterations in the scan to reduce the time taken for the procedure. An example of the binning procedure is shown in Figure 3.17. The red histogram is the S/B distribution after step 1, the blue after step 2 and the black vertical lines show the final set of 7 bins chosen for this analysis. Dijet tagged events are treated in the same way as the rest of the events in the analysis by introducing an eighth bin containing events from any BDT output bin inside the range $\Delta m/m_H < w$ which pass the dijet tag.

3.4.4 Background Model

The diphoton background is expected to have a smoothly varying invariant mass spectrum. However, detector effects such as selection and trigger efficiencies and energy resolution shape this distribution in ways which are imperfectly modelled in MC simulation. Moreover, the background contains fakes whose sizeable contribution vary as a function of $m_{\gamma\gamma}$. This means the exact composition of the background is needed to model the shape with MC. In order to remove the impact of systematic uncertainties associated to this, an entirely data-driven approach to the background model is needed.

For a given mass hypothesis, the shape and normalization of the background model are obtained separately. The shape, meaning the fraction of events in each BDT output bin, is extracted from the BDT output distributions in mass-sidebands, while the overall normalization is obtained from a parametric fit to the mass distribution for all selected events excluding the signal region.

Figure 3.18 shows the invariant mass distribution after event selection in the range $100 < m_{\gamma\gamma} < 180$ GeV for the full 2011 dataset. The red band indicates the signal region for $m_H = 124$, while the six blue bands indicate the corresponding sidebands used to

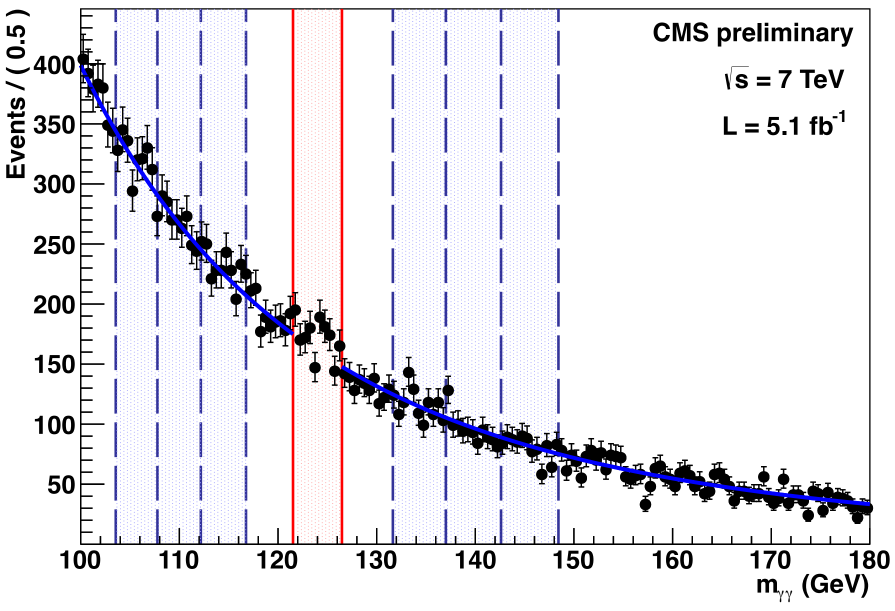


Figure 3.18: Invariant mass distribution of the full 2011 dataset after selection over the mass range used in the analysis (100 to 180 GeV). The ±2% signal region for $m_H = 124$ is indicated in red, while the six corresponding sidebands are indicated as blue bands. The blue line is the double power law fit to the data for the background normalisation for this mass hypothesis.

determine the shape of the background model (Section 3.4.4). The blue line indicates the fit of a double power law used to determine the normalisation of the background in the signal region as described in Section 3.4.4.

Obtaining the normalisation of the background

The normalisation of the background model is estimated using an un-binned maximum likelihood fit of a parametric function to the diphoton invariant mass distribution in the range $100 < m_{\gamma\gamma} < 180$ GeV. The normalisation of the background model is given by the integral of the function over the ±2% signal region for each mass hypothesis. The signal region is excluded from the fit to avoid potential bias in the presence of a signal.

The parameterization is chosen following a study of different parametric forms which also provide a good fit to the data. Since the actual functional form is unknown, the choice of parameterization is taken to be that which minimises the total uncertainty when comparing to the other functional forms. Twelve different functional forms were considered, which can be grouped into four general classes; exponentials, power laws, real Laurent polynomials and standard polynomials. Within each of these classes, three functions were used. For the exponentials and power law cases, these were sums of one, two or three exponential or power law ($m_{\gamma\gamma}^{-r}$) terms, while only first, third and fifth order standard polynomials were used. For the Laurent polynomials, the functions were sums of two, four or six terms, specifically

$$\begin{aligned} m_{\gamma\gamma}^{-4} + am_{\gamma\gamma}^{-5}, \\ m_{\gamma\gamma}^{-4} + am_{\gamma\gamma}^{-5} + bm_{\gamma\gamma}^{-3} + cm_{\gamma\gamma}^{-6}, \\ m_{\gamma\gamma}^{-4} + am_{\gamma\gamma}^{-5} + bm_{\gamma\gamma}^{-3} + cm_{\gamma\gamma}^{-6} + dm_{\gamma\gamma}^{-2} + fm_{\gamma\gamma}^{-7} \end{aligned}$$

For each class therefore, the three functions have one, three or five parameters for the shape.

To asses the bias introduced through choosing one particular parameterisation, pseudo-experiments are generated from each functional form and the invariant mass of those experiments are fit with the other functional forms. The parameters for generation of the pseudo-experiments are fixed by fitting each functional form to the data in the full mass range. In each pseudo-experiment, the integral of a particular fitting function, A, over the signal region is compared to that from a generating function, B. The distribution of the difference between the two values across all of the pseudo-experiments are used to determine the bias introduced from choosing function A when B was the true function. The distributions are then weighted according to the probability of the initial fit and combined so that the total uncertainty from choosing a particular function is computed as the RMS from zero of the weighted summed distributions for all generating functions. Since one of the generating functions can also be the fitting function, the error includes both the statistical uncertainty from the limited data sample and the systematic uncertainty due to an incorrect choice of parameterisation. The total error for all twelve fitting functions is given in Figure 3.19. This study is repeated at 5 GeV intervals in m_H as the overall uncertainty varies as a function of mass hypothesis. Figure REF shows total error determined for each of the twelve functions for each value of m_H. The double power law was found to give a low total uncertainty while also demonstrating good fit stability in the pseudo-experiments. The total error on the background normalisation is included as a single systematic uncertainty for the purpose of signal extraction (Section on signal extraction).

Obtaining the shape of the background

Both inputs to the BDT are designed to be insensitive to the invariant mass of the diphoton system therefore, the BDT output distribution should be the same for any region of the $m_{\gamma\gamma}$ spectrum. Since the background composition remains relatively constant across the range 100 to 180 GeV, data in sidebands of $m_{\gamma\gamma}$, away from the signal, can be defined to determine the distribution of the background inside the signal region. For a particular m_H, a contiguous set of lower/upper sidebands are defined to be the ranges $|(m_{\gamma\gamma} - m_{H,i})/m_{H,i}| < w$ centered on $m_{H,i}$ as given in Equation 3.6 where $w = 0.02$.

$$m_{H,i} = m_H \left(\frac{1-w}{1+w}\right)^i \tag{3.6}$$

The two sidebands adjacent to the signal window (corresponding to $i = \pm 1$ in Equation 3.6) are not used in order to avoid signal contamination. Dijet tagged events are treated in the same way as the rest of the events by introducing an eighth bin containing dijet tagged events inside the range $\Delta m/m_H < w$. The distributions for the two input

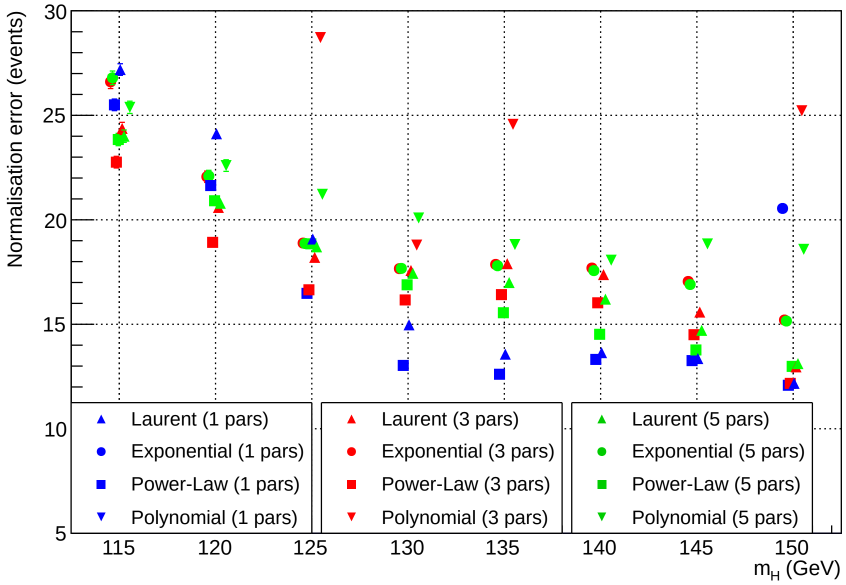


Figure 3.19: Total error on background normalisation as a function of m_H from different choices of the background shape parameterisation of $m_{\gamma\gamma}$. The total error for the one-parameter exponential and polynomial functions are off the scale of this plot.

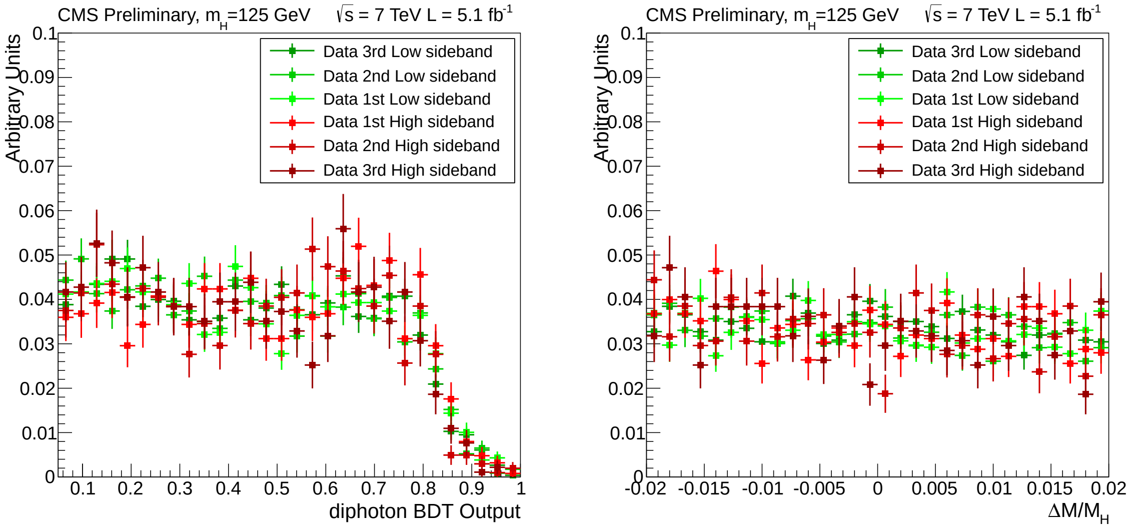


Figure 3.20: Distribution in data from the six sidebands corresponding to m_H = 125 GeV of the two BDT input variables, diphoton BDT (left) and $\Delta m/m_H$ (right).

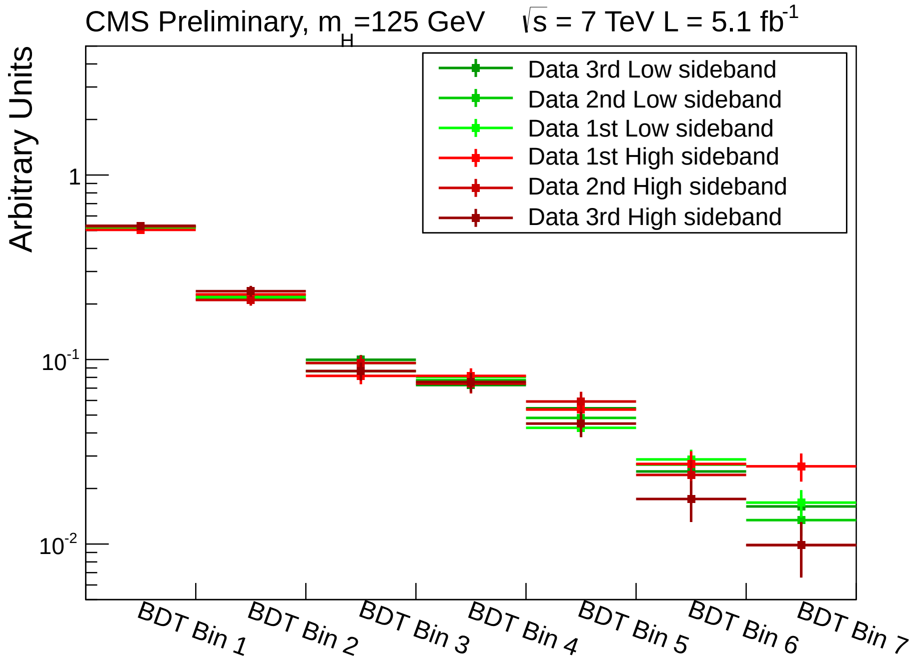


Figure 3.21: Distribution in data from the six sidebands corresponding to $m_H = 125$ GeV of the BDT output binned in the 7 BDT output bins used for signal extraction.

variables, diphoton BDT output and $\Delta m/m_H$, for each of the six sidebands corresponding to $m_H = 125$ are shown in Figure **??**. Each distribution is normalised to unit area. The resulting BDT output distributions are shown in Figure 3.21.

The residual variation in BDT output is due to the small variation in background composition with mass. This is mostly due to the photon ID MVA distribution being sensitive to the fake component which varies with mass. In order to account for this variation, the background model is constructed using a simultaneous linear fit to the BDT output shape in the data sidebands. The expected fraction of events in each bin, f_j, for a given mass hypothesis, $m_{H,i}$, is given by Equation 3.7, where $j\epsilon\{1,8\}$ and $i\epsilon\{\cdots,-4,-3,-2,2,3,4\cdots\}$.

$$f_j = p_{0,j} + p_{1,j}(m_{H,i} - m_H) \tag{3.7}$$

Since the normalisation for the background model is determined independently, the sum over all bins is constrained to be one. The expectation value for the background in each bin, j, is then determined as Nf_j where N is the normalisation estimated in section 3.4.4. This constraint is imposed for all $m_{H,i}$ by fixing

$$p_{0,1} = 1 - \sum_{i=2}^{8} p_{0,j} \qquad p_{1,1} = -\sum_{j=2}^{8} p_{1,j}$$

The coefficients $p_{0,j}, p_{1,j}$ of Equation 3.7 are determined by performing a binned maximum likelihood fit to the observed fractions in the data assuming the contents of each bin in each sideband are Poisson distributed. The results of the fit for $m_H = 124$ are shown

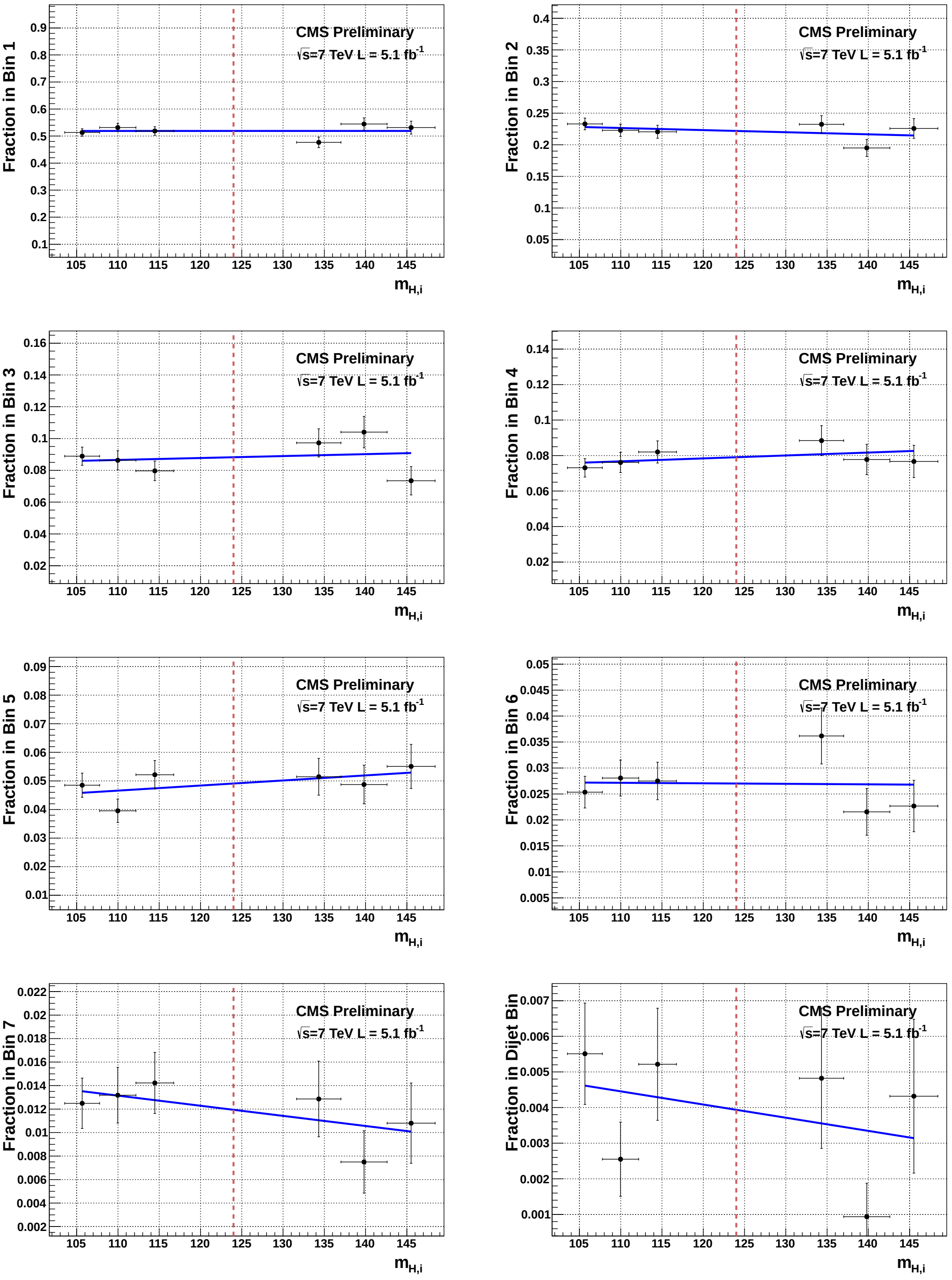
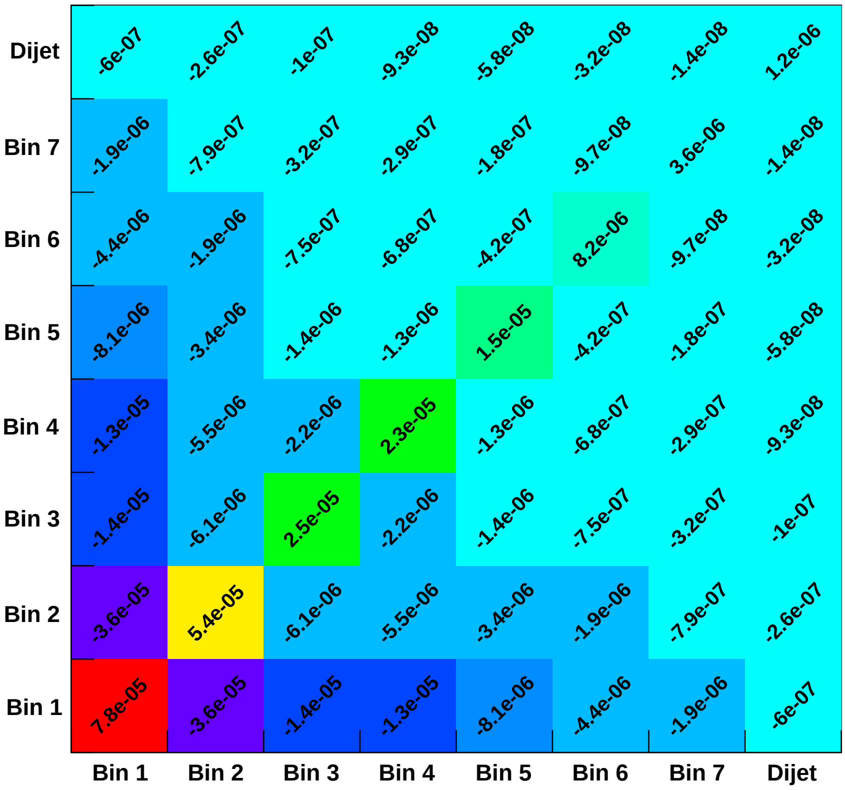


Figure 3.22: Simultaneous fits to the six sidebands in data to determine the background shape for $m_H = 124$. There are eight panels showing the result in each of the seven BDT bins plus one for the dijet tagged bin. The six black points in each panel are the are fractional populations of the data in each sideband. The blue line represents the linear fits used to determine the fraction of background in each bin at $m_H = 124$.



	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Dijet
Dijet	-6e-07	-2.6e-07	-1e-07	-9.3e-08	-5.8e-08	-3.2e-08	-1.4e-08	1.2e-06
Bin 7	-1.9e-06	-7.9e-07	-3.2e-07	-2.9e-07	-1.8e-07	-9.7e-08	3.6e-06	-1.4e-08
Bin 6	-4.4e-06	-1.9e-06	-7.5e-07	-6.8e-07	-4.2e-07	8.2e-06	-9.7e-08	-3.2e-08
Bin 5	-8.1e-06	-3.4e-06	-1.4e-06	-1.3e-06	1.5e-05	-4.2e-07	-1.8e-07	-5.8e-08
Bin 4	-1.3e-05	-5.5e-06	-2.2e-06	2.3e-05	-1.3e-06	-6.8e-07	-2.9e-07	-9.3e-08
Bin 3	-1.4e-05	-6.1e-06	2.5e-05	-2.2e-06	-1.4e-06	-7.5e-07	-3.2e-07	-1e-07
Bin 2	-3.6e-05	5.4e-05	-6.1e-06	-5.5e-06	-3.4e-06	-1.9e-06	-7.9e-07	-2.6e-07
Bin 1	7.8e-05	-3.6e-05	-1.4e-05	-1.3e-05	-8.1e-06	-4.4e-06	-1.9e-06	-6e-07

Figure 3.23: Covariance matrix from the sideband fit to determine the background shape at $m_H = 124$. The covariance matrix includes the additional 20% systematic attributed to possible second order variations in the BDT output background distribution with mass.

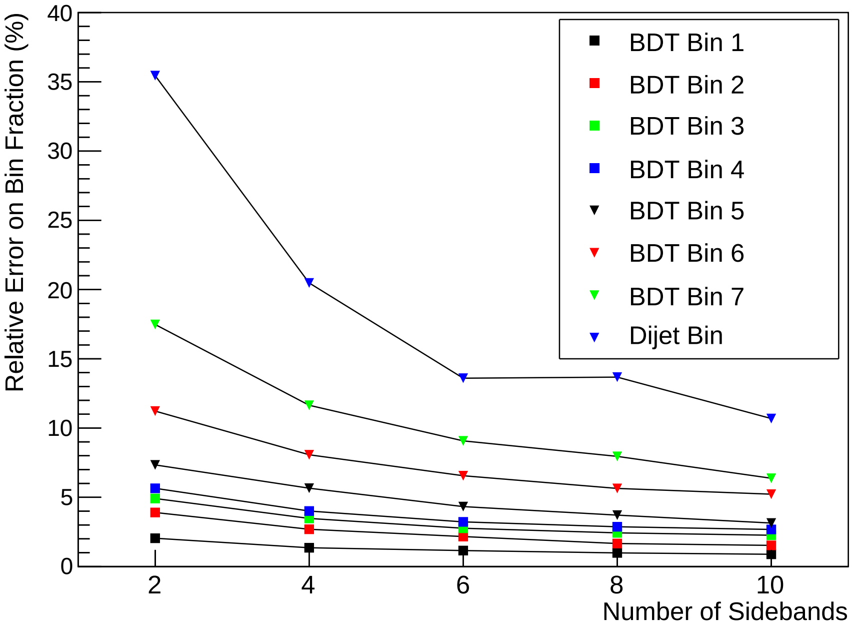


Figure 3.24: Relative total fit uncertainty on the background model in each bin at $m_H =$ 130 as a function of the number of sidebands used in the fit to determine the shape of the background.

in Figure 3.22 and the resulting covariance matrix obtained is shown in Figure 3.23. The fit was performed using `TMinuit` under `ROOT 5.2.0`.

There are seven degrees of freedom (eight bins minus one constraint) which are correlated. In order to account for the statistical uncertainty from the fit, a set of seven uncorrelated variables are determined from the covariance matrix using eigenvector-decomposition. These variables provide are treated as seven independent sources of systematic uncertainty on the background shape for purpose of signal extraction (Section **??**). Figure **??** shows the total relative fit error for each bin, at $m_H = 130GeV$, as the number of sidebands, is varied. Increasing the number of sidebands beyond six, three on each side of the signal region, provides negligible reduction in the statisitcal uncertainty. In order to avoid Drell-Yan contamination at the lower mass hypotheses any lower sideband whose lower boundary is less that 100 GeV is removed and an additional higher sideband is introduced. Consequently mass hypotheses in the range $111 \leq m_H < 115.5$ have two lower and four upper sidebands and mass hypotheses in the range $110 \leq m_H < 111$ have one lower and five upper sidebands.

At most linear variations with mass are considered for the background BDT output distribution. This corresponds to evaluating the first term in a Taylor series for the true shape of the distribution about m_H. Higher terms can be introduced but the statistical precision of the fit will be reduced in doing so. To check for potential significant deviations in the data from linearity, pseudo-experiments were generated in which the expected fractions, f_i are assumed to follow Equation 3.8.

$$f_j = p_{0,j} + p_{1,j}(m_{H,i} - m_H) + \frac{1}{2}p_{2,j}(m_{H,i} - m_H)^2 \tag{3.8}$$

The parameter values, $p_{0,j}, p_{1,j}$ and $p_{2,j}$ and their uncertainties were determined by fitting over a larger number of sidebands for a particular mass hypothesis. This is done by extending the range of j to allow any sideband which is contained inside the range $100 < m_{\gamma\gamma} < 180$ GeV. For most mass hypotheses, this corresponds to fifteen sidebands in total. For each pseudo-experiment, the parameters were varied within their uncertainties (accounting for correlations) thereby systematically altering the expectation value for the number of events in each bin before generating a Poisson toy for the observed number of events per bin in each sideband. The usual linear fit is then performed and the fraction of events in each bin for the signal region is extracted and compared to the true generating fraction. The difference of these two values can be used to determine the total error under the assumption that a second term in the Taylor expansion is present in the data. This error is taken as the RMS around zero of the difference between the true and fitted values for f_i in 10,000 pseudo-experiments. When compared to the error from the linear fits, it was found that the total uncertainty was covered by inflating the errors systematically by 20%. The value of 20% is a conservative choice being the largest value found when repeating the study over a range of mass hypotheses.

3.4.5 Signal model

The signal model for the Higgs at a given mass is constructed by binning the BDT response from MC simulation of the four production processes, ggH, qqH, wzH and ttH. The simulation is corrected using auxiliary measurements from $Z \rightarrow e^{+}e^{-}$ events in data to account for imperfect modeling of the detector. These corrections are applied to the Monte Carlo event by event and can be categorized into photon level and di-photon level corrections.

Photon level corrections

The energy resolution of the calorimeter is measured in data using $Z \rightarrow e^{+}e^{-}$ events in photon-level categories separated by the ECAL module boundaries and depending on how likely the photon is to have converted in the tracker material. Photons in the central region of the detector with $r_9 > 0.94$ are further divided into those whose supercluster seed lies close to a module boundary and those who do not. The additional resolution smearing required for the Monte Carlo in each category is determined by smearing $Z \rightarrow e^{+}e^{-}$ MC until the $e^{+}e$ invariant mass distribution matches that of the data. This additional resolution is included in the Higgs MC by scaling the energy of each photon by $G(1, \sigma_{cat})$ where G is a Gaussian distributed random variable centered at 1, and σ_{cat} is the additional resolution required to match the data in a particular category. The exact definitions of the photon-level categories and the additional resolution measured in each category are given in Table **??**.

The efficiency for a photon to pass the pre-selection is measured in $Z \rightarrow e^{+}e^{-}$ data four photon-level categories separated by the ECAL barrel-endcap boundaries and the value of r_9 for the photon being greater than or less than 0.94. This is then applied to signal MC as a reweight of each event given by the product of the efficiencies for each photon of the selected diphoton.

In addition to these corrections, the value of σ_E and the photon ID MVA for each photon is shifted in each signal event to account for imperfections in detector simulations as described in Section

Diphoton level corrections

The efficiency to select the correct vertex in the event is measured using $Z \to \mu^+\mu^-$ events as a function of the boson p_T. Signal MC events are categorized by whether or not the selected vertex is within 10mm of the generated vertex. Each event is then re-weighted according to whether or not the selected vertex is the true vertex according to the $Z \to \mu^+\mu^-$ measurement. The L1/HLT efficiency is measured in four di-photon categories depending on the maximum super-cluster η and minimum r_9 value of the two photons. The simulated events are re-weighted according to that efficiency as measured using a tag and probe method in $Z \to e^+e^-$ data.

Systematic uncertainties

For each correction applied to the MC, the accuracy to which that correction is measured provides an estimate of the uncertainty present in the signal model. In the case of the energy scale measurement, no correction is applied to the MC although the uncertainty in that measurement is treated as a systematic on the per-photon energy in signal MC events. The systematic uncertainties which effect the shape of the signal are treated as correlated, migrations across the BDT output bins. The effect of each systematic in each bin is derived by shifting the relevant quantity in the signal MC and recalculating the BDT output for each event. The difference between the signal yield after applying the shift in each bin from their nominal values gives quantifies the variaton due to that uncertainty. In practise, these quantities are derived by applying shifts to the MC corresponding to 3σ variation of each uncertainty and interpolating the difference from the nominal values back to the 1σ level. This is done so that the evaluation of the variation in each bin is more robust for sytematics which have a small effect on the BDT output and in signal processes with fewer available MC statistics. Figure 3.25 shows the effect of the energy scale and resolution uncertainties on the BDT output of signal from gluon-gluon fusion production.

Imperfections in the simulation of the shower shape variables can cause discrepancies in the photon ID and σ_E distributions obtained from the respective BDT's between data and MC. To account for this, systematic uncertainties are included corresponding to shifting or scaling the output of the photon ID BDT and regression BDT respectively and recalculating the BDT output for each event in signal MC. The size of the uncertainty is chosen to be that which voers the maximal difference in the rato of each distribution in high p_T photons between data and MC. This is then validated using $Z \to e^+e^-$ in which the electrons are reconstructed as photons.

Due to the large variations observed when using different underlying event parton showering (UEPS) model for the two dominating production processes, systematics of 70% and 10% are included for the uncertainty in the fraction of gluon-gluon fusion and vector boson fusion respectively which pass the dijet tag.

In addition to the shape systematics, theoretical errors on the standard model Higgs cross-section are included due to uncertainties on the QCD scale and pdf variations of the various production modes as detailed in CITE. A 2.2% luminosity error is also included as an uncertainty on the overall signal yield. A complete table of the systematics included in the signal model is given in Table 3.3.

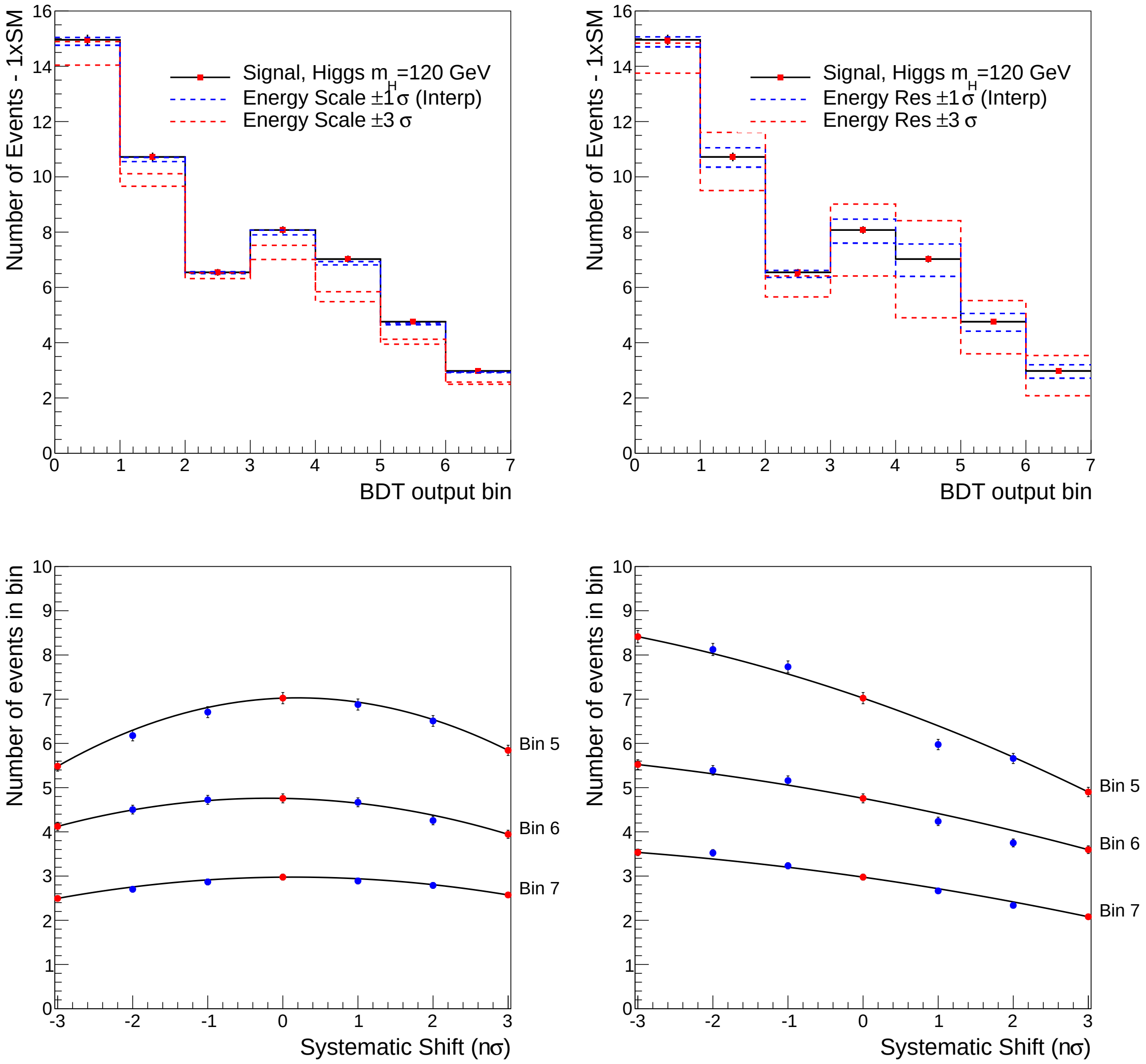


Figure 3.25: Top: Energy scale (left) and resolution (right) uncertainties in the ggH signal model. The effect of $\pm 3\sigma$ variations derived in MC are shown in red dashed lines while the interpolated $\pm 3\sigma$ are shown in blue. Bottom: Variation in bin content at different quantiles (number of standard deviations from the nominal) for the three highest S/B BDT bins. The blue and red markers indicate the yields extracted directly from MC while the black line indicates the quadratic interpolation function used to derive the $\pm 1\sigma$ variations for the signal model.

Source of systematic uncertainty		**Uncertainty**	
Per photon		Barrel	Endcap
Photon identification efficiency		1.0%	2.6%
Energy resolution	$r_9 > 0.94$ (low η, high η)	0.22%, 0.61%	0.91%, 0.34%
$(\Delta\sigma/E_{MC})$	$r_9 < 0.94$ (low η, high η)	0.24%, 0.59%	0.30%, 0.53%
Energy scale	$r_9 > 0.94$ (low η, high η)	0.19%, 0.71%	0.88%, 0.19%
$(E_{data} - E_{MC})/E_{MC})$	$r_9 < 0.94$ (low η, high η)	0.13%, 0.51%	0.18%, 0.28%
Photon identification MVA		±0.025 (output shift)	
Photon energy resolution MVA		10% (output scaling)	
Per Event			
Integrated luminosity		4.5%	
Vertex finding efficiency		$p_T^{\gamma\gamma}$-differential	
Trigger efficiency	either photon, $r_9 < 0.94$ in endcap	0.4%	
	Other events	0.1%	
Dijet-tagging efficiency	Vector boson fusion process	10%	
Dijet-tagging efficiency	Gluon-gluon fusion process	70%	
Production cross sections		Scale	PDF
Gluon-gluon fusion		+12.5% -8.2%	+7.9% -7.7%
Vector boson fusion		+0.5% -0.3%	+2.7% -2.1%
Associated production with W/Z		1.8%	4.2%
Associated production with $t\bar{t}$		+3.6% -9.5%	8.5%
Scale and PDF uncertainties		p_T-differential	

Table 3.3: Sources of systematic uncertainties included in the signal model Where a magnitude of the uncertainty from each source is given, the value represents a $\pm 1\sigma$ variation which is applied to the signal model.

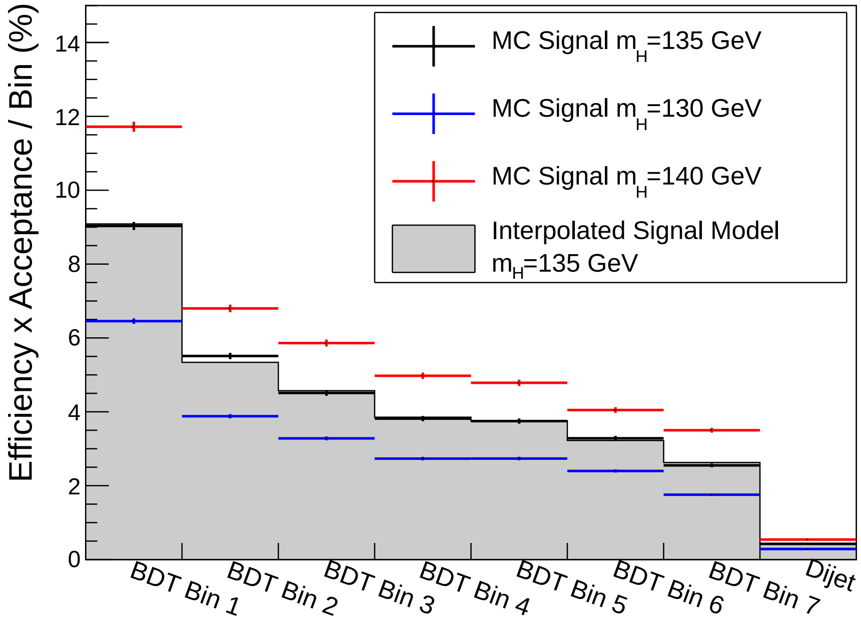


Figure 3.26: Closure test for signal interpolation to intermediate mass points. The solid grey histogram is the result of a linear intepolation between the efficiency×acceptance in each bin of the blue ($m_H = 130$) and red ($m_H = 140$) histograms. The efficiency×acceptance from ggH MC generated with mass 135 GeV is shown in black for comparison.

Interpolation to intermediate mass points

Signal Monte Carlo is available in m_H steps of 5 GeV in the range of 110 to 150 GeV. Due to the high resolution of the signal peak in the $H \to \gamma\gamma$ channel, it is necessary to interpolate between these generated mass points in order to construct the signal model at intermediate masses in finer steps. As a result of selecting BDT input variables that do not scale with mass, the BDT output distribution in signal varies slowly and smoothly with m_H. This allows for construction the BDT output signal distribution at an intermediate mass point by performing a bin by bin vertical interpolation between the distributions from MC at neighboring mass hypotheses. The interpolation is performed separately for each signal production mode. The normalization at intermediate points is defined as the cross section times branching ratio, which is known for any m_H, for the intermediate mass multiplied by a linear interpolation of the acceptance times efficiency. A closure test on the interpolation procedure was performed by comparing the efficiency times acceptance per bin at $m_H = 135$ with one derived from gluon-gluon fusion MC generated with $m_H = 130$ and $m_H = 140$ GeV (Figure 3.26). The closure test shows good agreement between the distributions; residual differences are negligible compared with the other systematics included in the signal model.

Validation with $Z \to e^+e^-$ data

As with the other MVA discriminators in the $H \to \gamma\gamma$ analysis, the signal model is validated by running the BDT in both Zee MC and data with the electron veto inverted. A

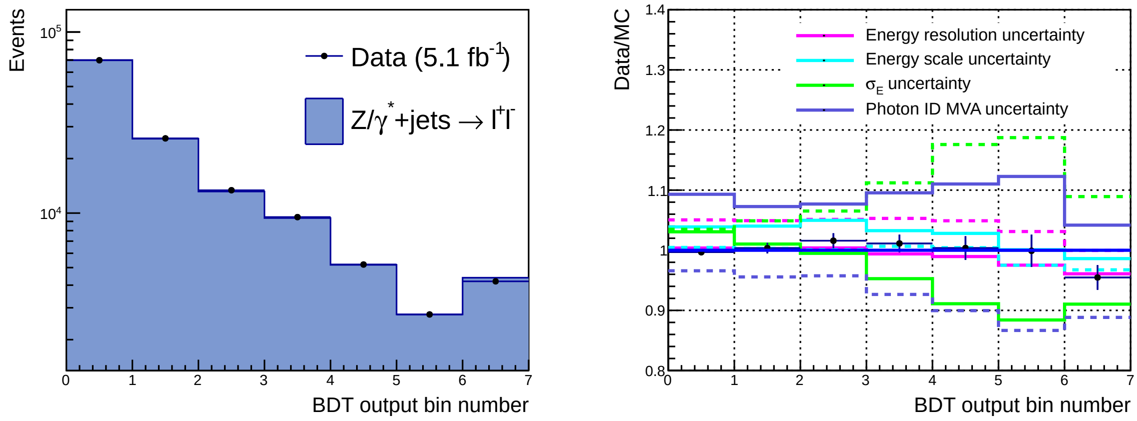


Figure 3.27: BDT output distribution for $Z \to e^+e^-$ events in data and MC (left). Data/MC ratio for the BDT output distribution (right). The variation in MC due to the largest systematic uncertainties included in the signal model are shown for comparison.

comparison of the data and MC is shown in Figure 3.27. Although the BDT output shape is not expected to be the same for $Z \to e^+e^-$ events as for $H \to \gamma\gamma$ events, the agreement seen between data and MC for $Z \to e^+e^-$ events indicates that the reconstruction and kinematics of a potential signal in data will be well modelled in the signal MC.

3.4.6 7 TeV results

The $H \to \gamma\gamma$ analysis was performed on the full 2011 dataset collected at CMS corresponding to 5.1 fb^{-1}of proton-proton collision data at a centre of mass energy of 7 TeV. Figure 3.28 show the observed number of events in data in each BDT output bin and from the dijet tagged events in the ±2% signal region centered on 124 GeV. The background model described in section 3.4.4 is shown in blue with the maximal uncertainty represented by the coloured bands. The expected contribution from a SM Higgs with a mass of 124 GeV is shown in red.

Statistical Interpretations of the Data

For the purposes of signal extraction, the analysis can be expressed in the form of a simple combination of counting experiments. The likelihood function (Equation 3.4.6) parameterises the relative compatibility of the data with the signal and background models as a function signal strength μ, where $\boldsymbol{\theta} = (\boldsymbol{\theta}^s, \boldsymbol{\theta}^b)$ are the nuisance parameters and ρ is a product of unit width Gaussian distributions centered at $\boldsymbol{\theta}_0$.

$$\mathcal{L}(data|\mu, \boldsymbol{\theta}) = \rho(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \cdot \prod_{j=1}^{8} Poisson\left(d_j|\mu \sum_p s_j^p(\boldsymbol{\theta}) + b_j(\boldsymbol{\theta})\right) \tag{3.9}$$

Before fitting to the data inside the signal region $\boldsymbol{\theta}_0 = \mathbf{0}$. The observed number of events in each bin, d_j, and expected contributions from each signal production process and background, s_j^p ($p = (ggH, qqH, wzH, ttH)$) and b_j, correspond to one mass hypothesis although the general form is applicable to all values of m_H.

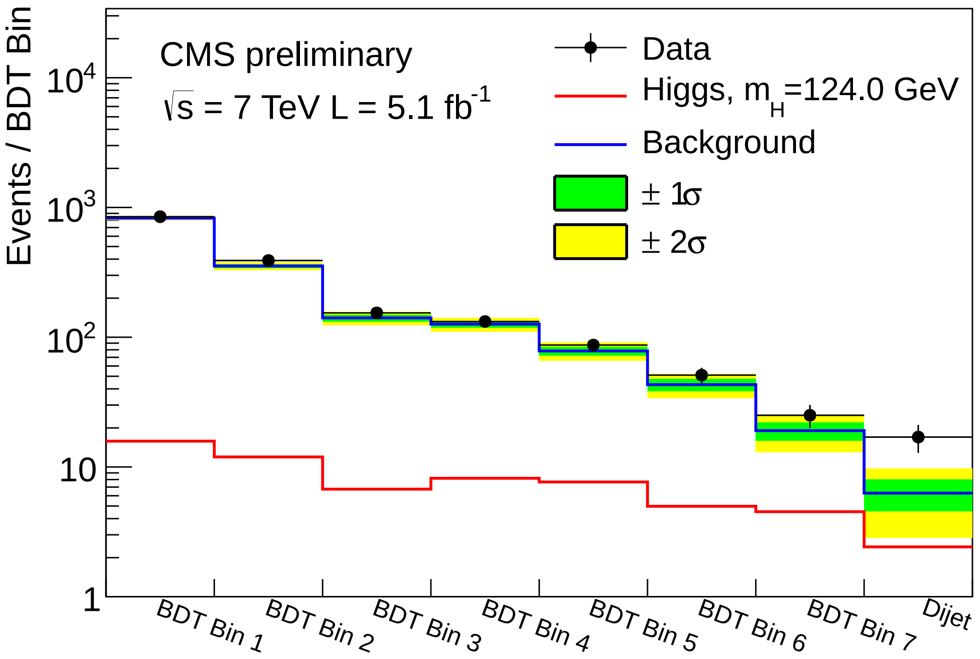


Figure 3.28: Observed number of events in data for each of the seven BDT bins and dijet bin at $m_H = 124$. The background model is shown in blue along with the maximal $\pm 1/2\sigma$ variations. The expected contribution from a SM Higgs is shown in red.

In order to avoid cases in which expectations for the contents of each bin become negative, the effect of each systematic on the signal or background is modelled using log-normal distributions. In this analysis, each systematic affects either the signal model or the background model. The functions $s_i(\boldsymbol{\theta}^s)$ and $b_i(\boldsymbol{\theta}^b)$ are given by Equations 3.4.6 and 3.4.6 respectively where $\boldsymbol{\theta}^s$ represents the nuisance parameters of the signal model and $\boldsymbol{\theta}^b = \left(\theta_N, \theta_1^b \ldots \theta_7^b\right)$ represent the eight independent nuisances of the background model.

$$s_j(\theta^s) = s_j^{p,mc} \cdot \prod_k \left(1 + \frac{\sigma_k^{s,p}}{s_j^{p,mc}}\right)^{\theta_k^s} \tag{3.10}$$

$$b_j(\theta^b) = N\left(1 + \frac{\sigma_N}{N}\right)^{\theta_N^b} \cdot f_j \prod_{k=1}^{7} \left(1 + \frac{\sqrt{\lambda_k} V_{kj}}{f_j}\right)^{\theta_k^b} \tag{3.11}$$

The values $s_j^{p,mc}$ in Equation 3.4.6 are the expectation values for the signal from each of the four Higgs production processes (ggH,qqH,wzH,ttH) derived from the signal MC taking all MC to data corrections into account. The values of $\sigma_k^{s,p}$ are the correlated bin uncertainties of the signal model due to each independent source of uncertainty calculated using the quadratic interpolation described in Section 3.4.5. In practise, $\sigma_k^{s,p}$ has two values, one corresponding to positive values of θ_k^s and one for negative values. This is to account for asymmetric variations caused by uncertainties in the signal model such as that due to the energy scale. The values V_{kj} and λ_k in Equation 3.4.6 are the eigenvectors and corresponding eigenvalues of the covariance matrix determined in Section 3.4.4. Finally, σ_N is the uncertainty on the background normalisation.

Exclusion limits on Higgs decay to two photons

To compare the compatibility of the data with the hypotheses that a Higgs signal is present, the test statistic, q_μ, is constructed as the ratio of two values of the likelihood given in Equation 3.4.6,

$$q_\mu = -2\ln\frac{\mathcal{L}(data|\mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(data|\hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{3.12}$$

where $\hat{\mu}, \hat{\boldsymbol{\theta}}$ denote the values for μ and $\boldsymbol{\theta}$ at which the likelihood attains its maximum and $\hat{\boldsymbol{\theta}}_\mu$ is the value at wich the likelihood is maximal under the condition that μ is fixed. An upper limit on $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$ can be calculated as an upper limit on μ by comparing the compatibility of the data against different hypotheses for μ. The background only hypothesis can be obtained by setting $\mu = 0$. For computing upper limits, the condition $0 \leq \hat{\mu} \leq \mu$ is imposed.

The compatibility of the data with a given value of μ is expressed using the CL_s procedure which is known to give conservative limits in the case of downward fluctuations of the background. This procedure involves computing two p-values (tail probabilities) under two hypothesis, $\mu = 0$ and $\mu \neq 0$ given by,

$$\begin{aligned} CL_{s+b} &= \int_{q_\mu^{obs}}^{\infty} f(q_\mu|\mu, \boldsymbol{\theta} = \boldsymbol{\theta}_\mu^{obs})dq_\mu \\ CL_b &= \int_{q_\mu^{obs}}^{\infty} f(q_\mu|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs})dq_\mu \end{aligned}$$

where q_μ^{obs}. The value of μ for which the ratio $CL_s = \dfrac{CL_{s+b}}{CL_b} = 0.05$ is the 95% confidence upper limit on $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$. When the upper limit on μ is less than one, the production of a SM Higgs which decays to two photons is ruled out at the 95% confidence level.

The distribution of the test statistic under the two hypothesis are generated by throwing pseudo-experiments using the signal and background models derived in Section 3.4. First, the values of $\boldsymbol{\theta}_\mu^{obs}$ and $\boldsymbol{\theta}_0^{obs}$ are set by fitting the likelihood to the observed data fixing μ and setting $\mu = 0$ respectively. Pseudo data, d_j, for each bin are generated according to a Poisson distribution with expectation value $\mu s_j(\theta_\mu^{obs}) + b_j(\theta_\mu^{obs})$. The nuisance parameter expectation values, $\boldsymbol{\theta}_0$, are then randomized according to their Gaussian pdfs before evaluating the test statistic q_μ in order to model the effect of sytematic uncertainties. Examples of the normalised distributions of q_μ for $\mu = 0.6$ and $\mu = 0$ are shown in Figure 3.29.

The 95% confidence upper limit on $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$ was determined using the full 2011 dataset for different values of m_H in the range to which the channel $H \rightarrow \gamma\gamma$ is most sensitive. Since the resolution of the signal peak in the $H \rightarrow \gamma\gamma$ channel is of the order 1 GeV, the limit is calculated in 100 MeV steps in the range $110 < m_H < 150 GeV$. Figure 3.30 shows the expected and observed upper limit on the ratio $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$ in that range. Where the observed line falls below the red line at one, a SM Higgs decaying to two photons, with mass m_H, is excluded at the

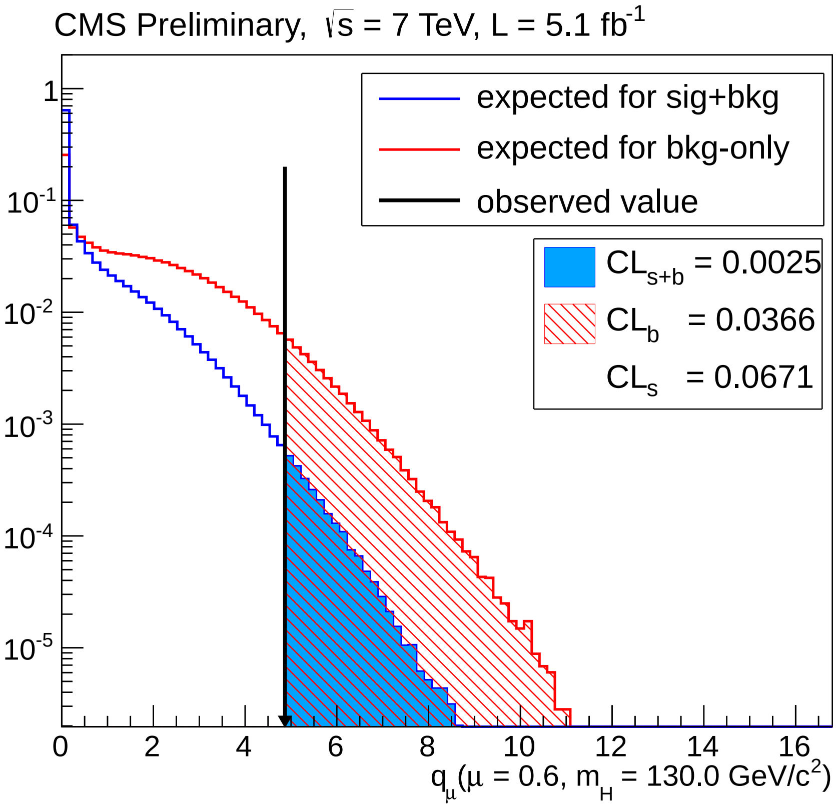


Figure 3.29: Distributions of the test statistic q_μ under a background only hypothesis ($\mu = 0$) and signal plus background hypothesis ($\mu = 0.6$) for a Higgs of mass 130 GeV. The distributions are normalised to unit area. The observed value of the test statistic from data is indicated by the black arrow.

	Toys	**Asymptotic**
$m_H = 120$		
2.5%	0.519	0.540
16%	0.327	0.347
median	1.174	1.174
84%	0.516	0.517
97.5%	1.199	1.226
$m_H = 130$		
2.5%	0.506	0.528
16%	0.321	0.344
median	1.148	1.145
84%	0.507	0.503
97.5%	1.178	1.200
$m_H = 140$		
2.5%	0.633	0.631
16%	0.399	0.399
median	1.361	1.345
84%	0.606	0.603
97.5%	1.406	1.436

Table 3.4: Comparison of expected median upper limit and quantiles obtained using the asymptotic calculation of CL_s and toys. The comparison is made at five mass hypotheses in the range 120 to 140 GeV.

95% confidence level. The limits were calculated using an asymptotic approximation for the distribution of q_μ thereby removing the need for generation of pseudo-experiments. The procedure involving the generation of toys was however conducted for several mass hypotheses and found to agree with the asymptotic calculation. Table 3.4.6 show this comparison for the median expected, 68% and 95% quantile ranges at different values of m_H.

Quatifying excesses in the observed data

Excesses above the background can be caused by fluctuations of the background itself or due to the presence of a signal. The significance of such excesses can be expressed as the probability to observe a signal like background fluctuation at least as unlikely as the one observed in data. This is the same as the probability one would attribute such an excess to a signal when no such signal is present.

The test statistic which quantifies the relative compatibilty of the data with the background only hypothesis and the presence of a signal, with any signal strength, is q_0. This is obtained by setting $\mu = 0$ in Equation 3.12 and removing the upper bound on $\hat{\mu}$. Again, there is an implicit assumption that the test statistic is defined only given a particular value of m_H. The test statistic desinged this way means that only excesses which are compatible in shape with that of a $H \to \gamma\gamma$ signal at some m_H are considered significant. As the mass peak of $H \to \gamma\gamma$ is narrow, this results in only localised excesses in $m_{\gamma\gamma}$ being significant. The probability that the background can fluctuate to produce a localised excess (local p-value) p_0 is given in Equation 3.13 where q_0^{obs} is the value of

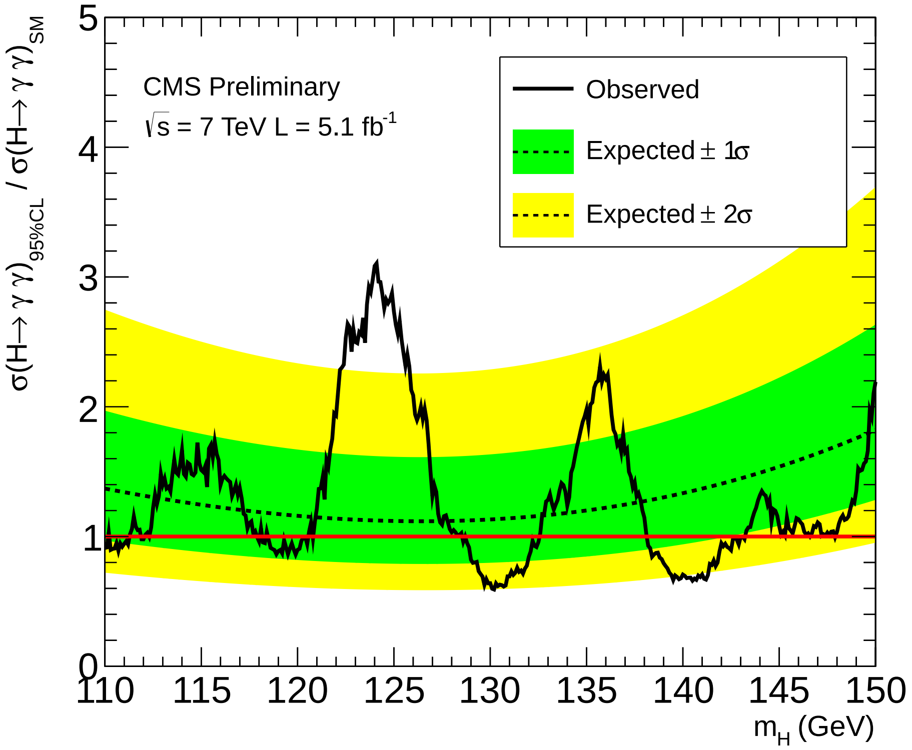


Figure 3.30: Exclusion limits on SM higgs production and subsequent decay to two photons in the range $110 < m_H < 150$ GeV. The black dahsed line indicates the median expected value for the upper limit on μ given the size of the dataset while the green and yellow bands indicate the 68% and 95% quantile ranges respectively. The black solid line shows the observed upper limit extracted from the data at steps in m_H of 100 MeV. Where this line falls below the red line at 1, a SM Higgs at that mass is excluded at the 95% confidence level.

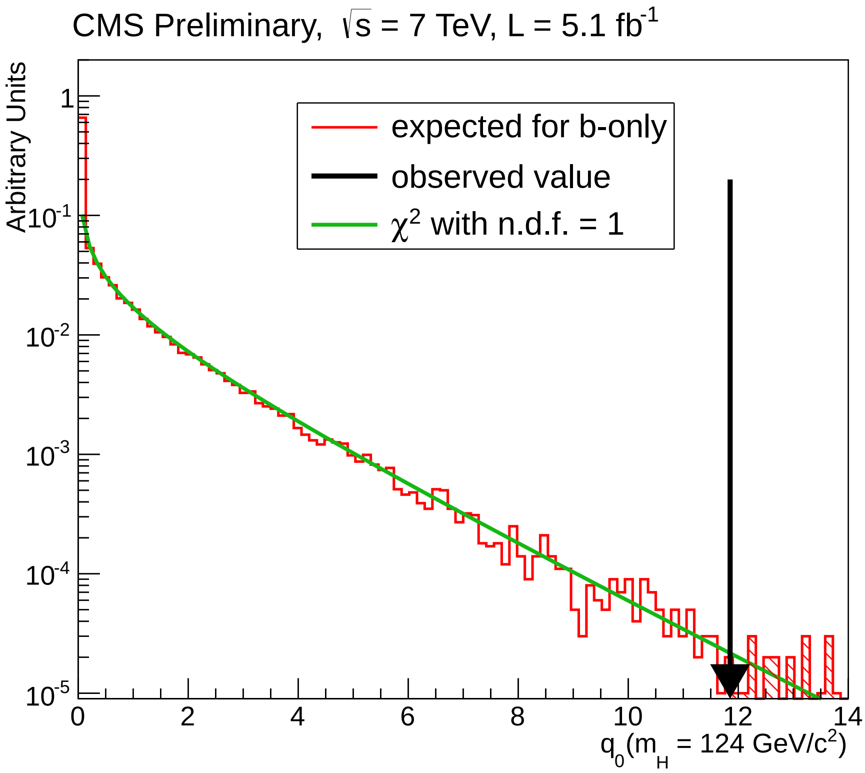


Figure 3.31: Normalised distribution of q_0 at $m_H = 124$ GeV under the background only hypothesis generated from toys (red histogram) and from the analytic form (green line). The observed value, q_0^{obs}, obtained from the data is indicated by the black arrow.

the test statisitc obtained in data.

$$p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs}) dq_0 \tag{3.13}$$

Analagous to calculating limits, the distribution $f(q_0|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs})$ can be obtained either through generating toys or using an analytic form. Figure 3.31 shows the normalised distribution of q_0 under the background only hypothesis generated from pseudo-experiments compared with the analytic form, in this case a χ^2 distibution with a single degree of freedom, at $m_H = 124$ GeV. The local p-value from the data is determined in steps of 100 MeV in the range $100 < m_H < 150$ GeV using the analytic expression $p_0 = sqrtq_0^{obs}$ as shown in Figure 3.32. The expectation in the presence of a SM Higgs at each m_H tested is shown in blue while the expectation from a SM Higgs with mass 125 GeV is shown in red. The largest excess in the range occurs near $m_H = 124$ GeV corresponding to a local significance of 3.4σ. The excess is larger than expected in the presence of a SM Higgs signal near that mass. This is reflected in Figure 3.33 which shows the value of μ at which the likelihood attains its maximum, $\hat{\mu}$, as a function of m_H. The excess observed at 124 GeV correpsonds to $\hat{\mu} = 1.93^{+0.67}_{-0.60}$, that is nearly twice the expectation from a SM Higgs.

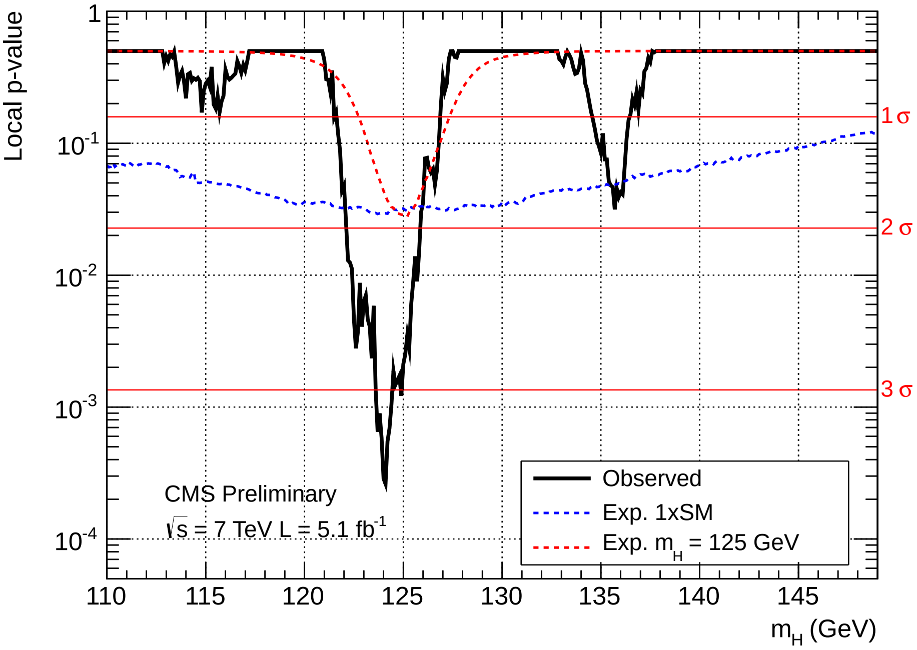


Figure 3.32: Local p-value (p_0) calculated in steps of 100 MeV in the range $110 < m_H < 150$. The observed p_0 obtained from the data is shown in black while the expected value in the presence of a SM Higgs is given by the dashed blue line. The expectation from a Higgs with mass 124 GeV is shown as a red dashed line. The right hand scale shows the significance in standard deviations at each m_H.

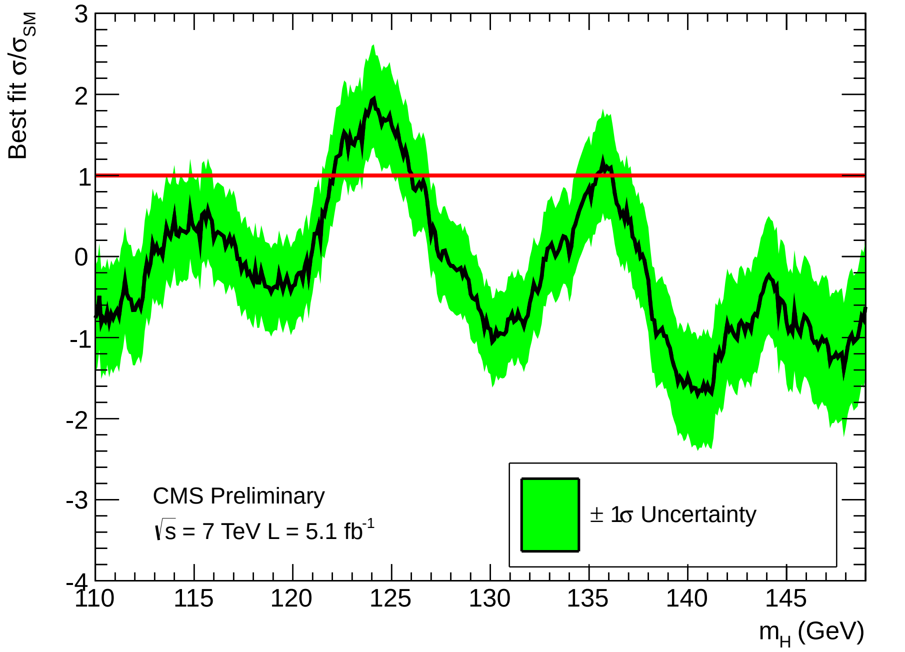


Figure 3.33: Best fit for the signal strength, $\hat{\mu}$, in steps of 100 MeV in the range $110 < m_H < 150$. The green bands indicate the 68% uncertainty on $\hat{\mu}$ for a fixed m_H. The red line at 1 represents the expectation for a SM Higgs.

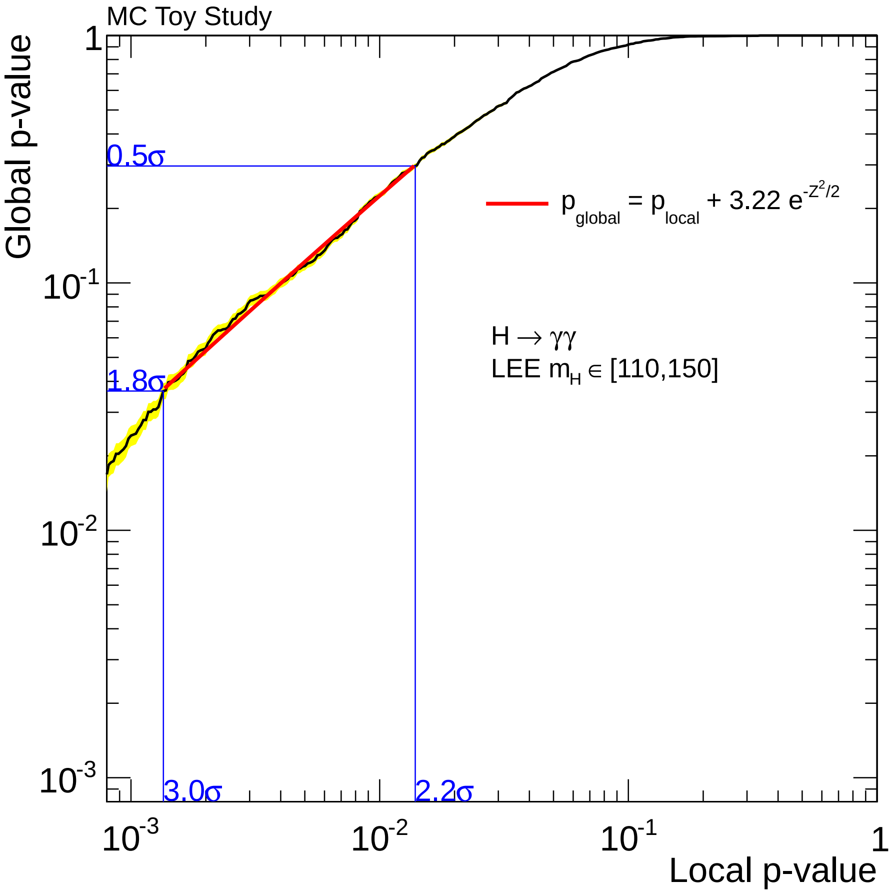


Figure 3.34: Relationship between local and global p-values to determine the look-elsewhere effect in the $H \to \gamma\gamma$ search for the range 110 to 150 GeV. The yellow band indicates the statistical precision of the relationship due to the limited number of toys produced. The red line indicated a fit of an analytic relation between the two and is used to calculate the global p-value for larger local significances.

The look-elsewhere effect

As the signal for the decay $H \to \gamma\gamma$ is a narrow mass peak, the probability to observe a local excess anywhere in the search range is much larger than the probability to find one at any particular m_H. This is an example of the look-elsewhere effect. Due to this, the local p-value of must be modified so as to express the probability to find an excess at least as significant as the one seen in data for all values of m_H. This is done by throwing background only pseudo-experiments and finding the minumum p_0 across all values of m_H. The fraction of pseudo-experiments with a minimum p_0 less than the one observed in data is then global p-value. Figure 3.34 shows the relationship between local and global p-values. The red line shows a fit of the function $p_{global} = p_{local} + Ce^{\frac{-Z^2}{2}}$ where Z is the local significance and C is a free parameter. This function is then used to determine the look-elsewhere effect for larger significances. The excess observed at 124 GeV corresponds to a 2.4σ global significance.

In order to generate suitable background only toys, pseudo-data are generated in two variabkes, $m_{\gamma\gamma}$ and the diphoton BDT output. The value of $m_{\gamma\gamma}$ for each event in the pesudo-data is generated from a double power law fit to the full $m_{\gamma\gamma}$ spectrum in data in the range $100 < m_{\gamma\gamma} < 180$ GeV. The value of the diphoton BDT is generated by fitting a `RooKeysPdf` to the distribution in data. The value of $\Delta m/m_H$ is then calculated for each pseudo-event at every m_H and the pseudo-dataset is analysed using the usual likelihood of Equation 3.4.6. This approach is necessary to maintain the correlations in the likelihood between neighbouring mass-hypotheses.

3.5 8 TeV results

The search described in the previous sections was repeated on data collected at CMS during the 2012 proton-proton run of the LHC at a center of mass energy of 8 TeV. The additional data was combined with the 7 TeV dataset as separate categories. The following section contains the results from the combined datasets corresponding to a total integrated luminosity of 10.4 fb^{-1}.

3.5.1 Updates for the 8 TeV Analysis

The majority of the analysis remains unmodified between the two data taking periods. Due to increased pileup conditions in the 2012 data, the regression BDTs and vertex BDTs were re-trained usig MC weighted to a higher average number of pileup vertices. As a result of this, both the diphoton and event categorisation BDTs were re-trained to incorperate the changes. In addition, the slight variation in kinematics between centre of mass energies 7 and 8 TeV are accounted for in the retraining. Both the energy scale and resolution were re-measured for the 2012 dataset and the corrections applied to data and MC as appropritate. Finally, the dijet category was further subdivided by separating events with a large reconstructed dijet mass, $m_{jj} > 350$ GeV, to improve the sensitivity of the search. Figure 3.35 shows the observed number of events from the 2012 dataset in each of the BDT output and the two dijet categories for $m_H = 124$ GeV. The background model is rederived using the same procedure described in Section 3.4.4 from the additional data. The contribution expected from a SM Higgs is shown in red.

3.5.2 Results from the Combined Datasets

The 2011 and 2012 datasets were combined statistically by extending the likelihood in Equation **??** to include a new set of categories which correspond to the updated analysis for the 2012 dataset. By including the additional data as separate categories, Exclusion limits and p-values are calcualated as described in Section **??**. Figure 3.36 shows the expected and observed 95% upper limits on $\sigma(H \to \gamma\gamma)/\sigma(H \to \gamma\gamma)_{SM}$ calculated in steps in m_H of 500 MeV from the combined datasets.

The observed local p-value, p_0, is determined for the 7 TeV, 8 TeV and combined datasets as function of m_H (Figure 3.37). The largest excess is observed at $m_H = 124$ GeV corresponding to a local significance of 4.8σ. This is reduced to a global significane of 3.9σ when considering the look-elsewhere effect in the range 110 to 150 GeV.

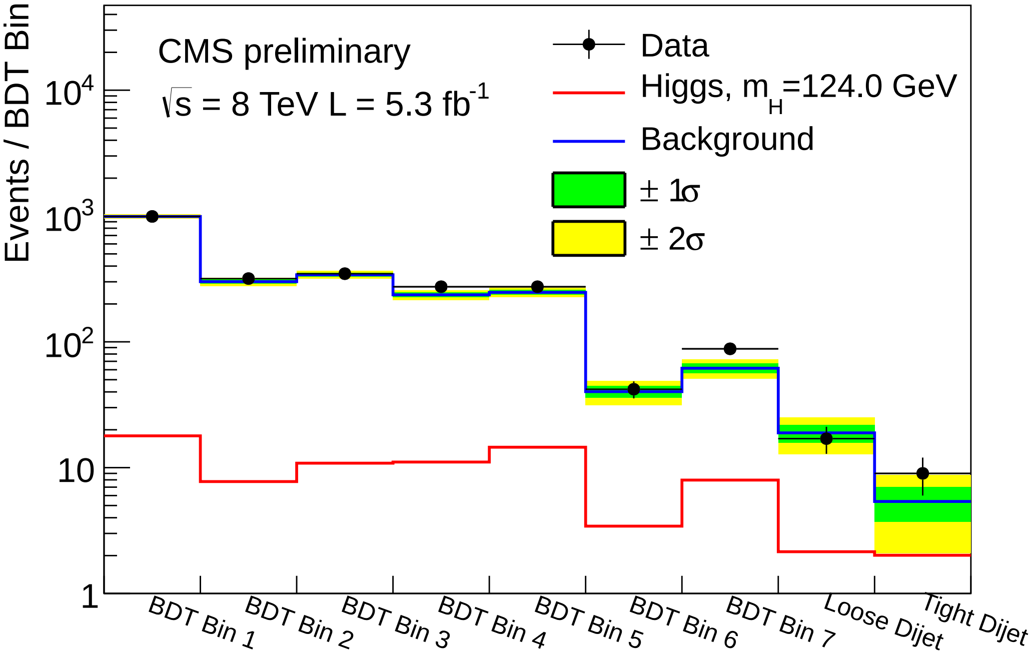


Figure 3.35: Observed number of events in the 2012 dataset for each of the seven BDT bins and tight/loose dijet bins for $m_H = 124$. The background model is shown in blue along with the maximal $\pm 1/2\sigma$ variations. The expected contribution from a SM Higgs is shown in red.

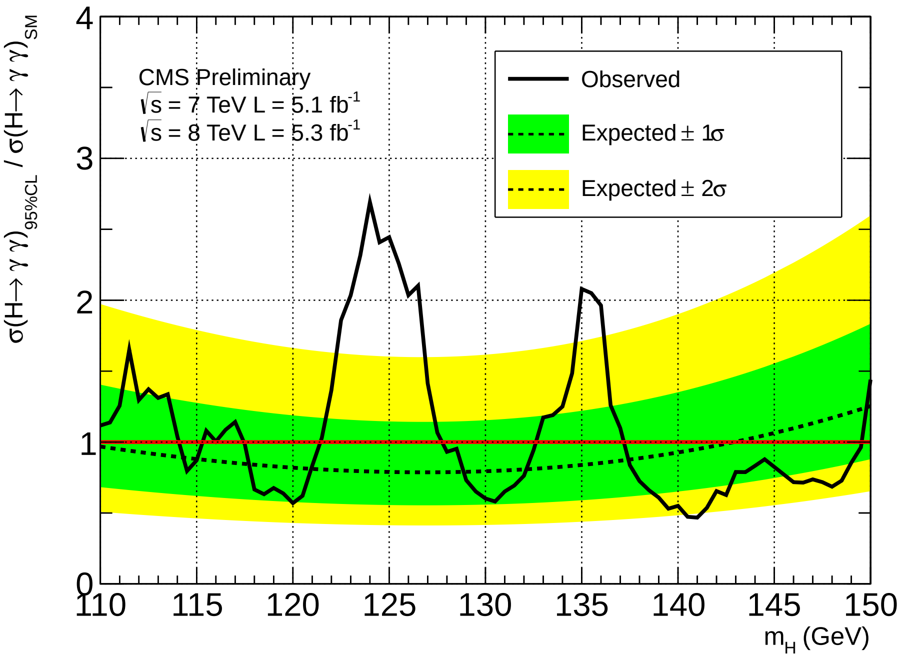


Figure 3.36: Exclusion limits on SM higgs production and subsequent decay to two photons in the range $110 < m_H < 150$ GeV from the combined 2011 (7 TeV) and 2012 (8 TeV) datasets. The black dassed line indicates the median expected value for the upper limit on μ given the size of the dataset while the green and yellow bands indicate the 68% and 95% quantile ranges respectively. The black solid line shows the observed upper limit extracted from the data at steps in m_H of 500 MeV. Where this line falls below the red line at 1, a SM Higgs at that mass is excluded at the 95% confidence level.

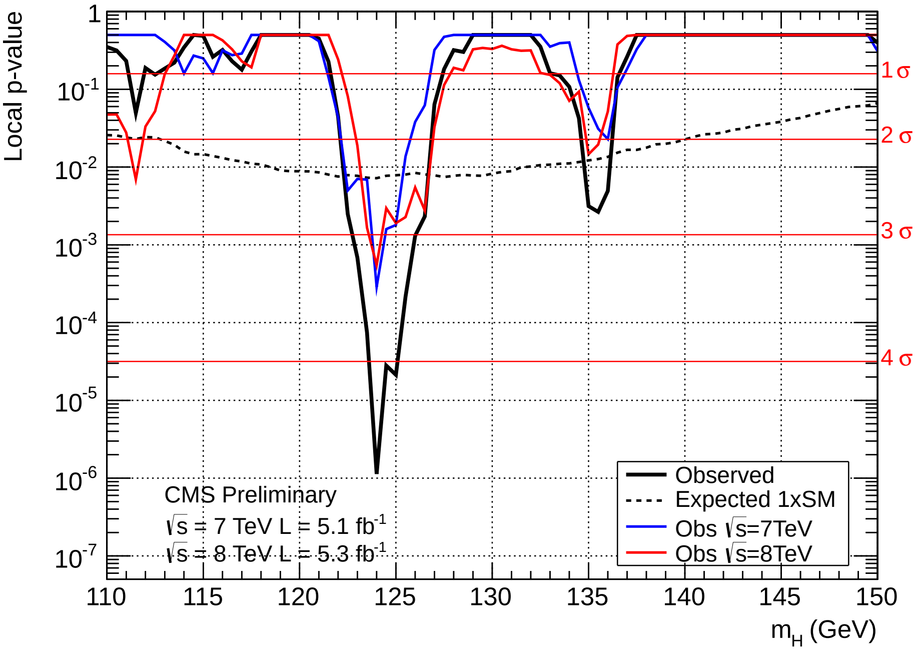


Figure 3.37: Local p-value (p_0) calculated in steps of 100 MeV in the range $110 < m_H < 150$. The observed p_0 obtained from the combined datasets is shown in black while the expected value in the presence of a SM Higgs is given by the dashed line. The observed p_0 from the 2011 (7 TeV) and 2012 (8 TeV) datasets individually are shown in the blue and red solid lines respectively. The right hand scale shows the significance in standard deviations at each m_H.

Chapter 4

Higgs Combinations and Properties

4.1 Statistical interpretation of data

≈ 2-3 pages explaining frequentist confidence intervals, CLs, p-values etc…

4.1.1 Diagnostics

≈ 2-5 pages depending on how much just gets lifted from the note.

4.1.2 Combined Higgs search results

≈ 3 pages results from sub-combiations, full combination, discovery

4.2 Properties

≈ 10 pages … unknown yet!

Chapter 5

Conclusions

≈ 1 page