

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

Nicholas Wardle  
Imperial College London

A dissertation submitted to Imperial College London  
for the degree of Doctor of Philosophy

## Abstract

I love  $H \rightarrow \gamma\gamma$

## Declaration

My work !

Nicholas Wardle

## Acknowledgements

blah blah...

## Preface

Insert preface here...

# Contents

0.1. Introduction . . . . .	1
<b>1. Theory and Motivations</b>	<b>2</b>
1.1. The Standard Model . . . . .	2
1.2. The SM Higgs . . . . .	2
1.2.1. The Higgs mechanism . . . . .	2
1.2.2. Constraints and previous searches . . . . .	2
1.2.3. Higgs production at the LHC . . . . .	2
<b>2. The LHC and the CMS Detector</b>	<b>3</b>
2.1. The LHC . . . . .	3
2.2. The CMS Detector . . . . .	4
2.2.1. Tracker . . . . .	6
2.2.2. Electromagnetic Calorimeter . . . . .	8
2.2.3. Shower-shape and Isolation . . . . .	14
2.3. Level-1 Trigger . . . . .	15
2.3.1. Jet Energy Calibration . . . . .	15
2.3.2. Calibration Performance . . . . .	17
<b>3. Higgs Decay to Two Photons</b>	<b>21</b>
3.1. Data Samples . . . . .	21
3.2. Object Reconstruction and Identification . . . . .	23
3.2.1. Supercluster Energy Correction . . . . .	23
3.2.2. Vertex Selection . . . . .	26
3.2.3. Photon Identification . . . . .	29
3.3. Event Selection . . . . .	30
3.3.1. Diphoton BDT . . . . .	31
3.3.2. Dijet Tagging . . . . .	35
3.4. Signal Extraction . . . . .	35
3.4.1. Definition of the Signal Region . . . . .	37

3.4.2. Event Categorisation BDT . . . . .	37
3.4.3. Binning of the BDT Output Distribution . . . . .	41
3.4.4. Background Model . . . . .	42
3.4.5. Signal Model . . . . .	50
3.4.6. Statistical Interpretations of the Data . . . . .	56
3.5. Inclusion of 2012 Data . . . . .	69
3.5.1. Updates for the 8 TeV Analysis . . . . .	69
3.5.2. Results from the Combined Datasets . . . . .	71
<b>4. Higgs Combinations and Properties</b>	<b>72</b>
4.1. Combined Higgs Searches . . . . .	72
4.1.1. Diagnostics with Toy Datasets . . . . .	75
4.1.2. Higgs Search Combination . . . . .	78
4.2. Higgs Properties . . . . .	83
4.2.1. Extracting Signal Parameters . . . . .	84
4.2.2. Combined Mass Measurement . . . . .	85
4.2.3. Compatibility with the Standard Model . . . . .	87
<b>5. Conclusions</b>	<b>94</b>
<b>A.</b>	<b>95</b>
A.1. Common Tools . . . . .	95
A.1.1. Isolation Sums . . . . .	95
A.1.2. Boosted Decision Trees . . . . .	95
<b>B.</b>	<b>96</b>
B.1. Energy Scale and Resolution Measurements . . . . .	96
B.2. Binning Algorithm Optimisation . . . . .	99
B.3. Signal Systematics . . . . .	102
<b>Bibliography</b>	<b>105</b>
<b>List of Figures</b>	<b>108</b>
<b>List of Tables</b>	<b>117</b>

*“Un bon mot ne prouve rien.”*

— François-Marie Arouet (Voltaire)

## 0.1. Introduction

Preamble, declaration of work, description etc...

≈ 10 pages (including list of figs/tables)

# **Chapter 1.**

## **Theory and Motivations**

### **1.1. The Standard Model**

Local Gauge theory + SM Lagrangian  $\approx 3$  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**

$\approx 2$  pages

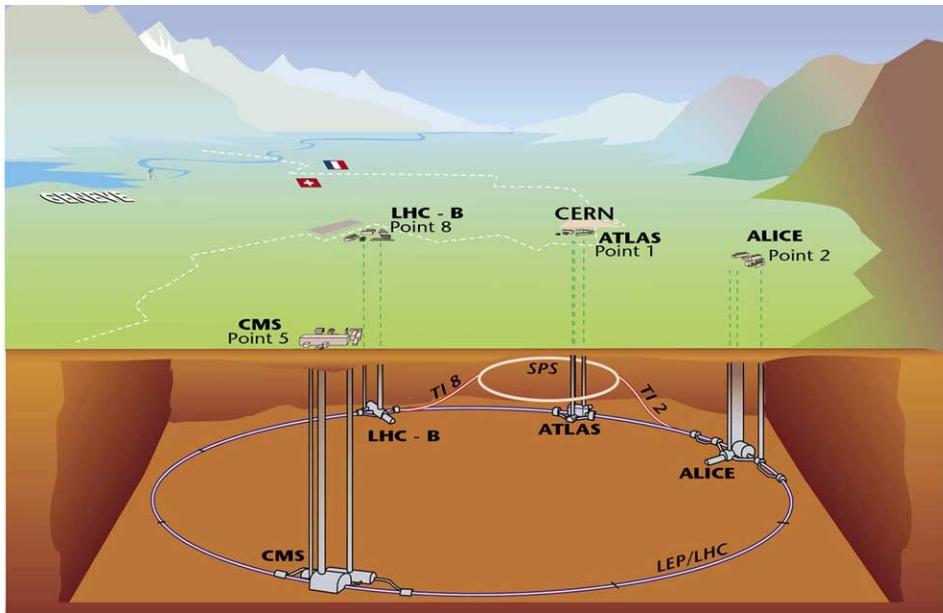
# Chapter 2.

## The LHC and the CMS Detector

### 2.1. The LHC

The Large Hadron Collider (LHC) at CERN is currently the leading collider experiment designed to study physics at the TeV scale. The collider is an octagonal ring, 27 km in circumference, hosted in the former LEP tunnel in France/Switzerland. Both proton-proton (p-p) and heavy ion (Pb-Pb) collisions are studied as part of the LHC physics programme however, the former are used for direct searches for new physics. Proton beams are formed inside the proton synchrotron (PS) from bunches of protons 50 ns apart with an energy of 26 GeV. The protons are then accelerated in the super proton synchrotron (SPS) to 450 GeV before being injected into the LHC. Around 1200 superconducting dipole magnets maintain two beams of protons accelerating around the ring in opposite directions before being collided at one of the four major experiments; Atlas, CMS, LHCb and ALICE. Figure 2.1 is a cartoon of the accelerator indicating the sites of the four experiments.

The first major physics runs began in May 2010 with a centre of mass energy  $\sqrt{s} = 7 \text{ TeV}$  and continued until November providing a  $44\text{pb}^{-1}$  integrated luminosity of data. The LHC resumed collisions in April 2011 delivering a further  $6\text{fb}^{-1}$  by the end of October. The centre of mass energy was increased to  $\sqrt{s} = 8 \text{ TeV}$  for the 2012 p-p collision run, improving the sensitivity of searches for new physics. A total of  $6\text{fb}^{-1}$  of 8 TeV data were taken by July 2012 which were combined with earlier data resulting in the discovery of a new boson reported by CMS and Atlas at the ICHEP conference that year.



**Figure 2.1.:** LHC accelerator ring. The relative locations of the four main experiments are indicated along with their points of access to the beam.

## 2.2. The CMS Detector

The Compact Muon Solenoid (CMS) detector is one of two general purpose detectors at the LHC designed to search for new physics. Among the wide range of physics programs at CMS, the search for the SM Higgs boson has a high priority. The decay rates of the SM Higgs boson in different channels vary dramatically as a function of its mass ( $m_H$ ). A key feature of the experiment's design was therefore the necessity to maintain a high sensitivity to the SM Higgs for a wide range of masses in as many decay channels as possible. To achieve this, several detector components are layered around the beam axis reconstruct almost any known particle produced at the interaction point. Each component consists of a cylindrical barrel section and two endcaps to provide an almost hermetic coverage of the outgoing particle flux.

The tracker, providing measurements of the momentum of charged particles and the location of primary and secondary vertices (from decays of heavy flavour mesons), is the first layer of detection. This is followed by the electromagnetic calorimeter (ECAL) which is used to measure the energy deposited as electromagnetic showers from interacting electrons and photons. The hadronic calorimeter (HCAL) complements this by providing energy measurements of sprays of hadrons, known as jets, which deposit energy through nuclear interactions. The HCAL is a sampling calorimeter in that the

active material (plastic scintillators) are sandwiched between dense absorbing material to increase the depth of the calorimeter to around 11 radiation lengths. The addition of the forward calorimeter (HF) extends the HCAL coverage to roughly  $|\eta| = 5$ . The tracker and calorimeters are situated within a 4T axial magnetic field provided by the superconducting magnet surrounding them. The magnetic flux return is implemented within the muon detector systems which lie outside the superconducting coil and form the outermost detection layers. Muons deposit very little energy throughout the detector and can carry on into the surrounding cavern. The barrel muon system is constructed from layers of drift-tubes (DT) interleaved with resistive plate chambers. The combination of the two provides high resolution timing and hit positions which are used to determine the trajectory of muons both from p-p collisions and cosmic sources for calibration. For the endcaps, the DTs are replaced with cathode strip chambers as the higher flux of particles along the beam line requires the use of radiation hard components.

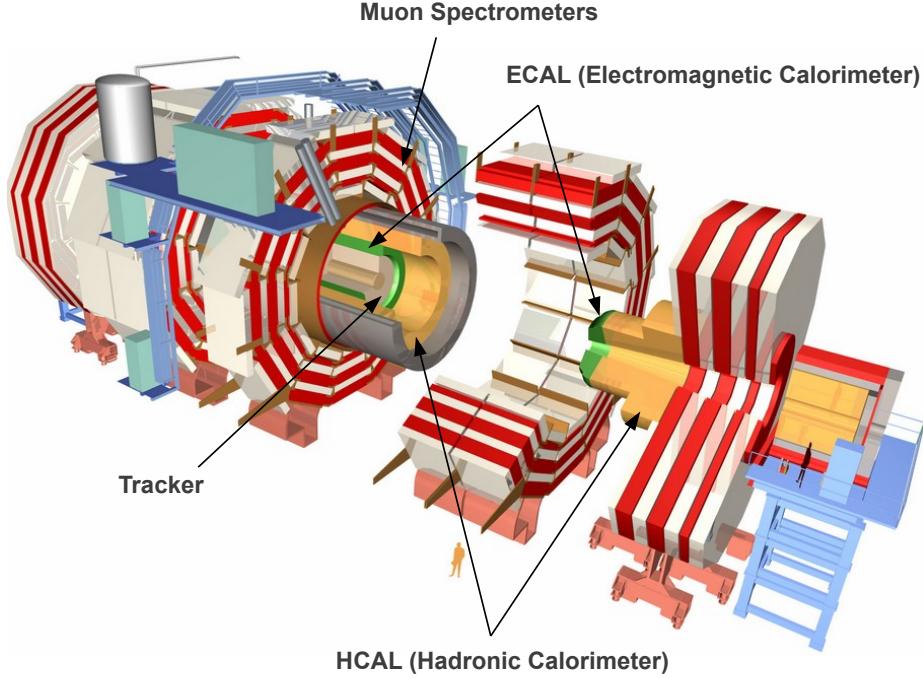
CMS uses a Cartesian coordinate system with the origin at the interaction point and the  $z$ -axis pointing along the beam axis. The  $x$ -axis points towards the centre of the LHC ring and the  $y$ -axis points vertically upwards. The azimuthal angle,  $\phi \in [-\pi, \pi]$ , is defined with respect to the  $x$ -axis in the transverse ( $x - y$ ) plane. The polar angle  $\theta$  is measured from the  $z$ -axis. Commonly, the direction of an outgoing particle is defined by  $\phi$  and its pseudo-rapidity  $\eta$  defined as

$$\eta = -\log \tan \left( \frac{\theta}{2} \right). \quad (2.1)$$

As hard collisions produce high momentum particles travelling perpendicular to the beam line, particles are often characterised by the magnitude of the projection of their momenta onto the transverse plane,  $p_T = \sqrt{p_x^2 + p_y^2}$ . Similarly, the transverse energy is defined as  $E_T = E \sin \theta$ . Figure 2.2 shows the geometry of the CMS detector and its major components.

### 2.2.1. Tracker

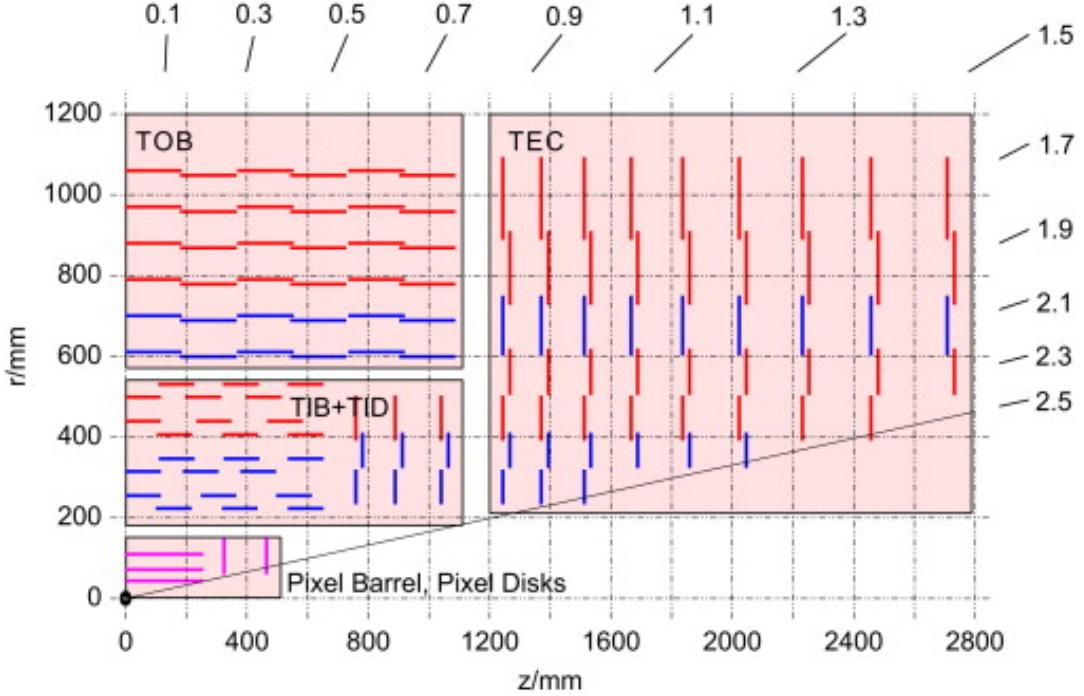
The CMS tracker is designed to reconstruct charged particle tracks which make up a large portion of the complex topology of p-p collisions. The tracker provides precise measurements of observables such as the momentum of charged particles and the location of the vertex at which they are produced. In addition to the high level of granularity required to make such measurements, the high rate of interaction at LHC requires a fast



**Figure 2.2.:** Diagram of the CMS Detector. The arrows indicate the main detector elements. The figure has been altered from its original source [21]

response from the tracking elements. The tracker is formed of a pixel detector component encased by layers of silicon strip detectors. The pixel detector is the closest tracking element to the interaction point. It is a composite of 66 million individual silicon pixels,  $100\mu\text{m} \times 150\mu\text{m}$  in size, forming three cylindrical layers around the beam line and two forward disks. Outside the pixel detector, ten cylindrical layers of silicon strip detectors (TIB/TOB) and twelve discs (TID/TEC) extend the tracking system out to a radius of 120cm from the beam line. The tracker geometry, as shown in Figure 2.3, covers a pseudo-rapidity range  $|\eta| < 2.5$ .

By making multiple precise measurements throughout the tracker system, the trajectories (tracks) of charged particles can be reconstructed. Tracks are associated to a common point of origin (primary vertex) by grouping those which are separated by less than 1cm in the  $z$  coordinate of the point of closest approach to the beam line. The vertex resolution is dependant both on the number of tracks associated to the vertex and their average transverse momenta ( $\bar{p}_T$ ). The resolution was measured in early data from 2010 by splitting tracks associated to a vertex randomly into two groups with equal kinematic distributions. The difference between the vertex locations calculated from the two groups was used to provide an estimate of the resolution [13]. Figure 2.4 shows the



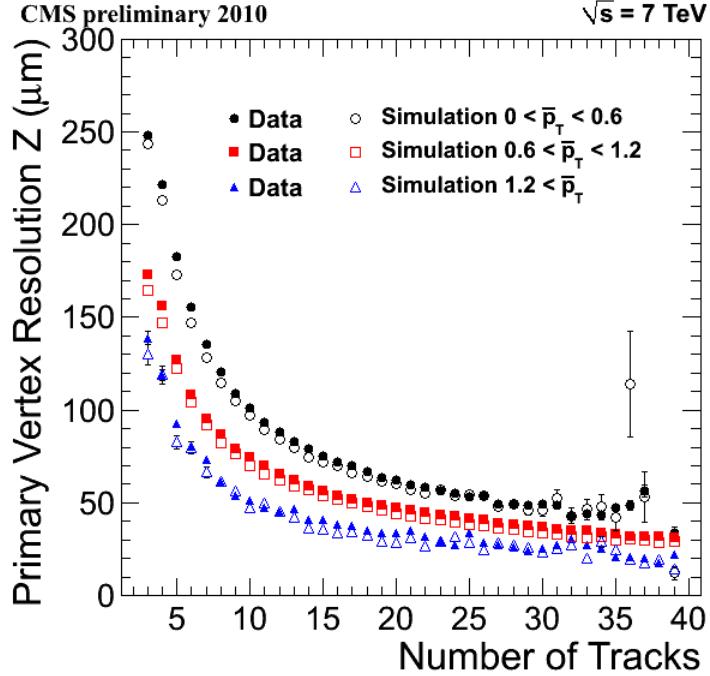
**Figure 2.3.:** Cross-section of the pixel and silicon strip detector components of the CMS tracker [1].

resolution in  $z$  as a function of the track multiplicity measured in data and simulation. The simulation provides a good description of both the trend with number of associated tracks and the improvement in resolution with  $\bar{p}_T$  in the data.

### 2.2.2. Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is used to reconstruct the energies of electromagnetically interacting particles such as electrons and photons. It is constructed from high density lead tungstate ( $\text{PbWO}_4$ ) crystals which form a barrel section (EB) and two endcaps (EE) outside the tracker. Two lead plates in front of a fine grained silicon strip detector are situated just before the endcaps forming the ECAL pre-shower (PS). Photons travelling at high  $\eta$  will convert in the lead and the resulting electron-positron pair will produce tracks which can be used to pinpoint the position of the incoming photon. This additional resolution in spatial separation can be used to distinguish prompt photons from those produced in neutral pion decays.

The ECAL is designed to cover a pseudo-rapidity range of  $|\eta| < 3$ . The crystals are arranged to form modules which surround the beam line in a ‘non-projective geometry’:

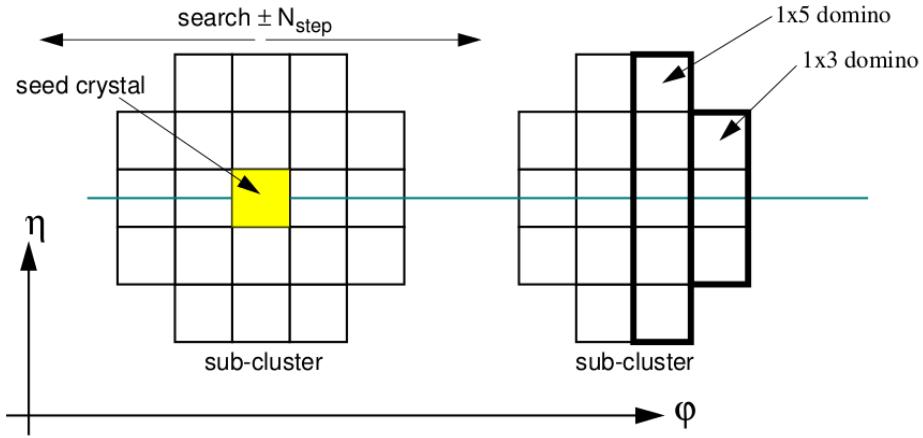


**Figure 2.4.:** Resolution of vertex  $z$ -position as a function of the number of tracks associated to the vertex measured in simulation and 2010 data. The resolution is given for three different average track momenta.

the gaps between crystal modules are offset by  $3^\circ$  with respect to particle trajectories originating from the interaction point. Electrons and photons deposit most of their energy within the crystals as the depth of the crystals is equivalent to 25.8 radiation lengths [15]. Electromagnetic showers produced by the interaction of electrons and photons in the ECAL crystals produce scintillation light which is collected to measure the energy of the particle. The scintillation output of the crystals is, however, low and temperature dependant ( $\sim 2.1\%/\text{K}$  at the ECAL operating temperature of 291 K). To overcome this low yield, avalanche photo-diodes (APD's) and vacuum photo-triodes (VPT's) are used to collect the scintillation light and amplify the signal in the calorimeter barrel and endcaps respectively. Around 4.5 photo-electrons per MeV are produced in both APD's and VPT's.

### Electron and Photon Reconstruction

Electron and photon candidates are formed by clustering deposits of energy caused by electromagnetic showers in the ECAL. For unconverted photons, these clusters will likely be well localised in  $\eta$  and  $\phi$  around the incident photon. However, for photons which



**Figure 2.5.:** Sub-cluster construction of the Hybrid algorithm used to reconstruct photons and electrons in the ECAL barrel.

convert in the tracker, the resulting electron-positron pair will deposit energy across several regions of the calorimeter. In the presence of the axial magnetic field, electrons radiate bremsstrahlung photons causing deposits which are spread over a wide range in  $\phi$  while being fairly narrow in  $\eta$ . This characteristic is exploited by the “Hybrid” clustering algorithm used to reconstruct high energy electrons and photons in the ECAL barrel [34]. The algorithm proceeds as follows;

- Step 1: A seed crystal is determined to be a single crystal in the barrel with the highest  $E_T$  satisfying  $E_T > 1 \text{ GeV}$ .
- Step 2:  $1 \times 3 (\phi \times \eta)$  crystal dominoes are formed with their central crystal aligned with the seed crystal in  $\eta$ . If the energy contained in the  $1 \times 3$  domino is larger than 1 GeV, the domino is extended by two crystals in  $\eta$ . A maximum of 10 dominoes are added in each direction in  $\phi$  starting from the seed crystal forming a sub-cluster.
- Step 3: Dominoes containing less than 100 MeV are removed and the remaining dominoes are grouped into sub-clusters providing each seeding domino for a sub-cluster contains more than 350 MeV. The final group of sub-clusters form a supercluster for the electromagnetic object.

The values of the particular thresholds used for seeding clusters and dominoes were tuned providing an efficiency for electrons with  $p_T > 7 \text{ GeV}$  greater than 99% [39]. Figure 2.5 is an illustration of the Hybrid clustering algorithm.

In the ECAL endcaps, superclusters are built using the “Multi5 $\times$ 5” algorithm which connects overlapping  $5 \times 5$  grids of crystals whose positions lie within 0.3 radians in  $\phi$  [42].

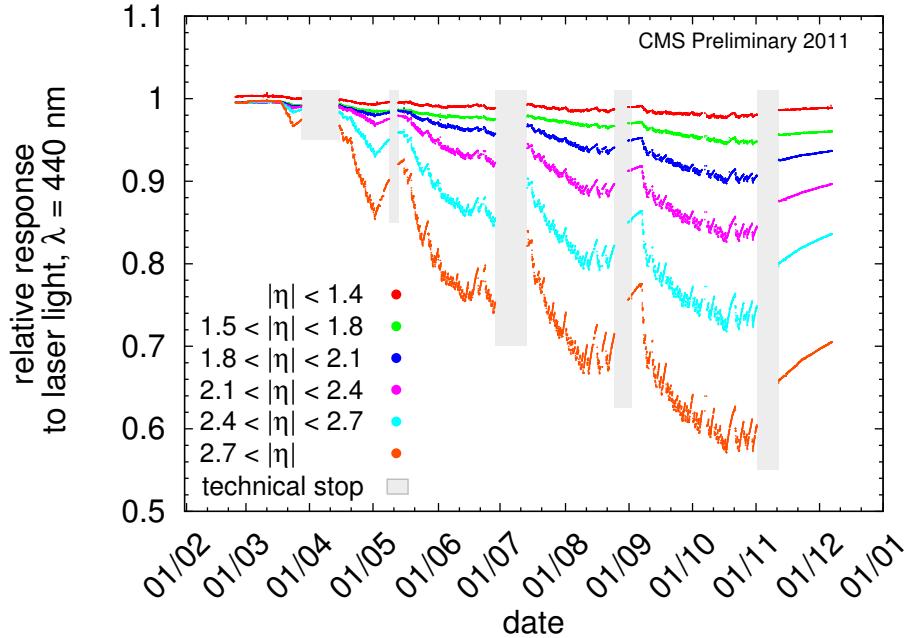
Additional information is used from the PS to enhance the energy reconstruction in the endcaps.

Superclusters are associated to electron candidates (`GsfElectrons`) where a track can be reconstructed from compatible hits in the tracker using a Gaussian sum filter algorithm [16]. This provides an additional measure of the electron's momentum which is used to improve the resolution of the electron energy. Aside from this, the reconstruction of photons and electrons is identical which is an important feature allowing for data driven calibrations and validations of photons using electrons such as those described in Chapter 3.

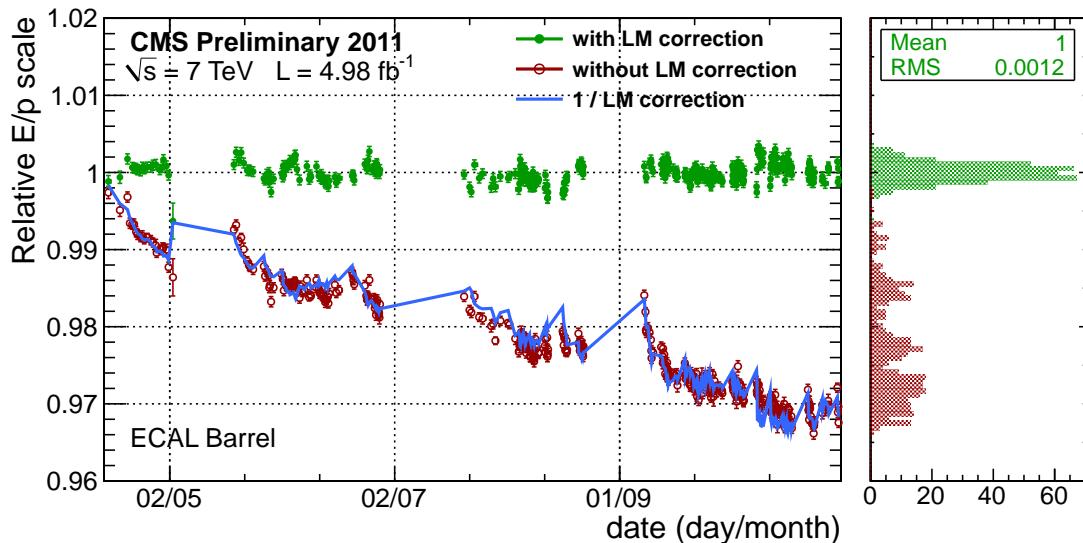
## Laser Calibration

ECAL crystals suffer from loss of optical transmission when irradiated through the formation of crystal-lattice defaults which absorb some of the scintillation light. Annealing acts to balance the damage from radiation which results in an equilibrium optical transmission which is dose-dependant [15]. At the LHC, the dose varies during each run. This requires that the time varying optical transmission of the ECAL crystals be monitored to asses the impact on energy measurements. The crystal transparency is monitored by comparing the relative transmission in blue laser light (440 nm), which is close to the scintillation emission peak, to infra-red (796 nm), which is far from the peak and relatively unaffected by the radiation damage. Figure 2.6 shows the relative response to the blue laser of the monitoring system averaged over all the crystals in bins of  $|\eta|$  throughout the 2011 data taking runs [2]. The time dependence of the response has larger variation for higher values of  $|\eta|$  due to the larger flux of particles along the beam axis.

The response of the crystals measured using the laser monitoring system is used to calibrate the energy reconstruction of the ECAL. These calibrations are validated in  $W \rightarrow e\nu$  data events by comparing the electron energy ( $E$ ) measured as measured by the ECAL to the momentum ( $p$ ) of the electron measured in the tracker [2]. Figure 2.7 shows the relative variation in the ratio  $E/p$  as a function of time throughout 2011. A stable scale is achieved through application of the laser calibrations.



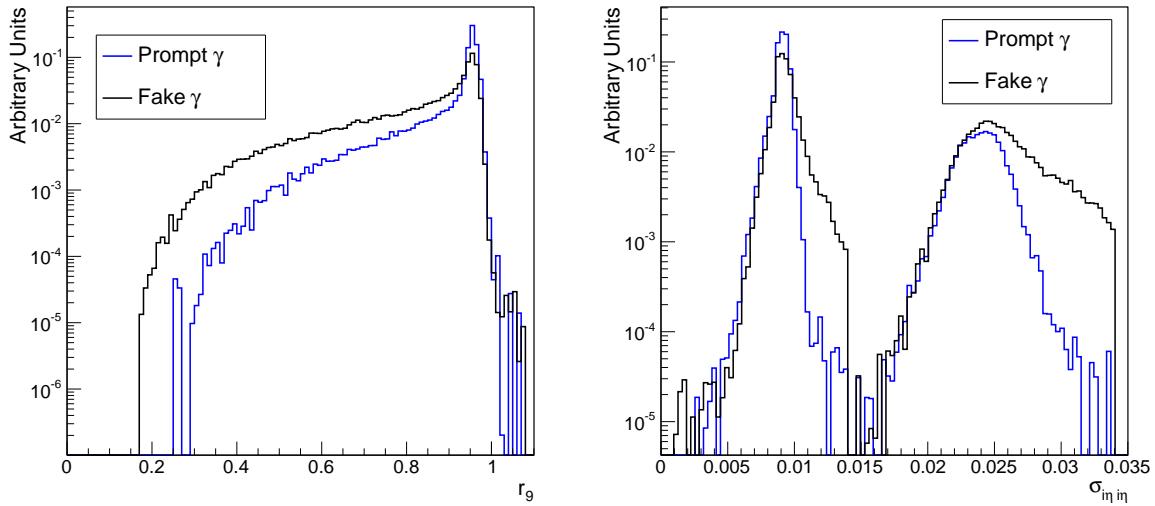
**Figure 2.6.:** Relative ECAL crystal response to blue laser light (440 nm) in bins of pseudorapidity, for the 2011 data taking period runs. The grey bands indicate periods during which there was no beam.



**Figure 2.7.:** Ratio  $E/p$  in electrons reconstructed in the ECAL Barrel from  $W \rightarrow e\nu$  events in 2011 data as a function of time before and after applying transparency corrections from the laser monitoring (LM) system. The blue line indicates the correction applied per point averaged over all crystals used in the electron energy measurement.

### 2.2.3. Shower-shape and Isolation

In addition to providing a measurement of the energy of incoming electromagnetic particles, the ECAL’s fine granularity provides additional information which can be used to characterise the supercluster and distinguish prompt electrons and photons from fakes. The shape of the electromagnetic shower can be described by the ratio of the energy contained in the central  $3 \times 3$  cluster surrounding the seed crystal to the total energy of the supercluster ( $r_9$ ). Superclusters associated to real unconverted photons will typically have larger value of  $r_9$  than those which are in reality due narrow  $\pi^0$  decays. Another common variable used for identification is the energy weighted crystal width of the sub-cluster used to seed the supercluster,  $\sigma_{in\eta\eta}$ . Prompt photons will tend to have a more localised cluster leading to lower values of  $\sigma_{in\eta\eta}$ . The distributions of  $r_9$  and  $\sigma_{in\eta\eta}$  are shown for a simulated sample of superclusters identified as photons from real and fake sources in Figure 2.8. The two distinct peaks in the  $\sigma_{in\eta\eta}$  distribution are due to the different superclustering algorithms used in the barrel and endcaps.



**Figure 2.8.:** Shower shape variable  $r_9$  (left) and  $\sigma_{in\eta\eta}$  (right) distributions for superclusters associated to simulated real and fake photons. The real photon is taken from simulated  $H \rightarrow \gamma\gamma$  events while the fake photon is taken from a  $\gamma + jet$  sample where the photon candidate is matched to a generated quark leg.

Hard interaction processes tend to produce electromagnetic particles which are well isolated in the detector. A cone is defined around the candidate with radius  $\Delta R$  defined

as,

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} \quad (2.2)$$

where  $\Delta\phi$  and  $\Delta\eta$  are positions relative the particle trajectory. The sum of  $E_T$  for each crystal inside the cone, after removing those associated to the supercluster itself, quantifies the isolation of the electron or photon candidate. Similarly, such isolation variables exist for the HCAL and tracker detector elements, summing over the  $E_T$  and  $p_T$  of HCAL deposits and tracks respectively, are defined in an analogous way to the ECAL isolation.

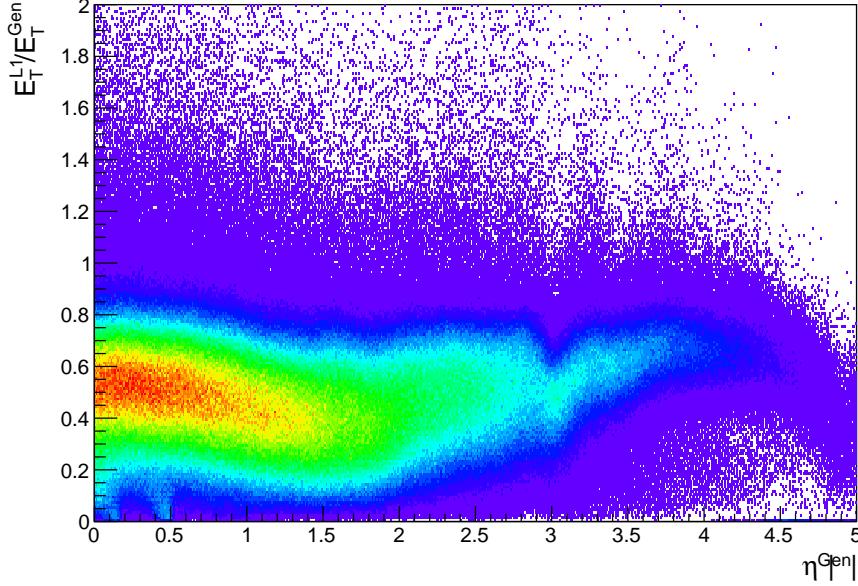
## 2.3. Level-1 Trigger

In order to cope with the high data output, a two-tier trigger system is implemented at CMS. The trigger is able to use limited information from each event to decide whether or not to record the event. This allows for a large reduction of the rate of data-taking while maintaining a high efficiency to select events producing interesting physics objects. The first level, the Level-1 (L1) trigger, uses custom-built electronics in order to reduce the data output of CMS from 40 MHz to 100 kHz [3]. Events which satisfy some relatively loose set of criteria are passed to the second level ,the higher-level trigger (HLT), where more sophisticated algorithms, much closer to those used in the offline reconstruction, are used to decide whether or not to store an event [4].

The L1 calorimeter trigger is able to use coarse measurements of the energy deposited in the ECAL and HCAL to form physics object candidates such as electrons, photons, tau leptons decaying hadronically and hadronic jets. With the exception of electrons and photons, all of the L1 algorithms run in the Global Calorimeter Trigger (GCT). The following section is a description of a set of calibrations for the GCT designed to improve the resolution of the L1 jets.

### 2.3.1. Jet Energy Calibration

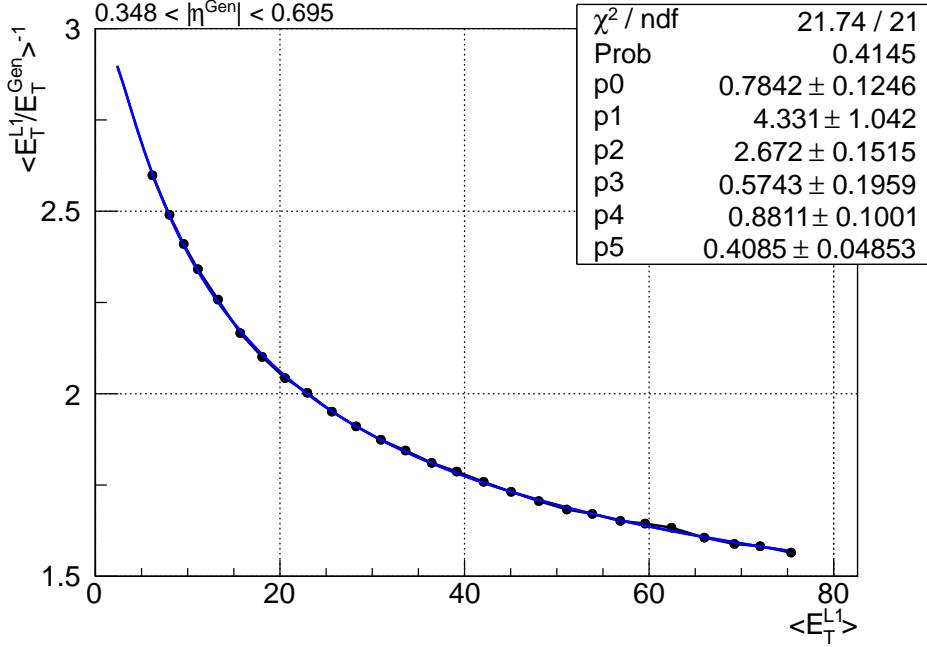
The response to hadronic energy of the CMS calorimeters varies considerably across its barrel, endcap and forward sections. The energies of jets are corrected offline to account for these effects, however, if left uncalibrated at L1, this can lead to inefficiencies in the



**Figure 2.9.:** Response measured from matched generator-L1 jet pairs in MC simulation as a function of the generator jet pseudo-rapidity  $|\eta^{Gen}|$ .

trigger system. The response is measured in QCD Monte Carlo (MC) simulation by comparing the  $p_T$  of L1 jet candidates to generated jets (defined using and anti- $k_T$  jet finding algorithm). L1 jets are matched to generator jets by determining the minimum separation,  $\Delta R$ , between each generator jet and any L1 jet candidate and requiring it be less than 0.7. This is much looser than typical matching requirements applied offline due to the coarser spatial resolution of the L1 jets. The generator and the closest of these L1 jet is defined as a matched pair and the response is calculated as  $E_T^{L1}/E_T^{Gen}$  for that pair. Figure 2.9 shows the response as a function of the pseudo-rapidity of the generated jet  $|\eta^{Gen}|$ .

The response is measured in 11  $|\eta|$  bins which correspond to the 11 GCT regions. Corrections for each region are derived as a function of  $E_T^{L1}$  by determining the average response,  $\langle E_T^{L1}/E_T^{Gen} \rangle$ , and  $\langle E_T^{L1} \rangle$  in 4 GeV bins of  $E_T^{Gen}$  between 14 GeV and 200 GeV. Below 14 GeV, the resolution in  $E_T^{L1}$  restricts a proper measurement of the response while above 200 GeV, the response approaches unity. The average response is taken from the mean of a Gaussian fit to the distribution of  $E_T^{L1}/E_T^{Gen}$  while  $\langle E_T^{L1} \rangle$  is taken as the mean average of the  $E_T^{L1}$  distribution. For low values of  $E_T^{Gen}$ , the response becomes very non-Gaussian due to the limited resolution of the L1 trigger so in this case, the average response is taken as the mean of the  $E_T^{L1}/E_T^{Gen}$  distribution. The response is inverted to provide a corrective scale factor in each region as a function of  $E_T^{L1}$ . This



**Figure 2.10.:** Correction function for the  $0.348 < |\eta^{Gen}| < 0.695$ . The points represent the average quantities as measured in MC simulation. The blue line is a parametric fit to the points using a chi-squared minimisation.

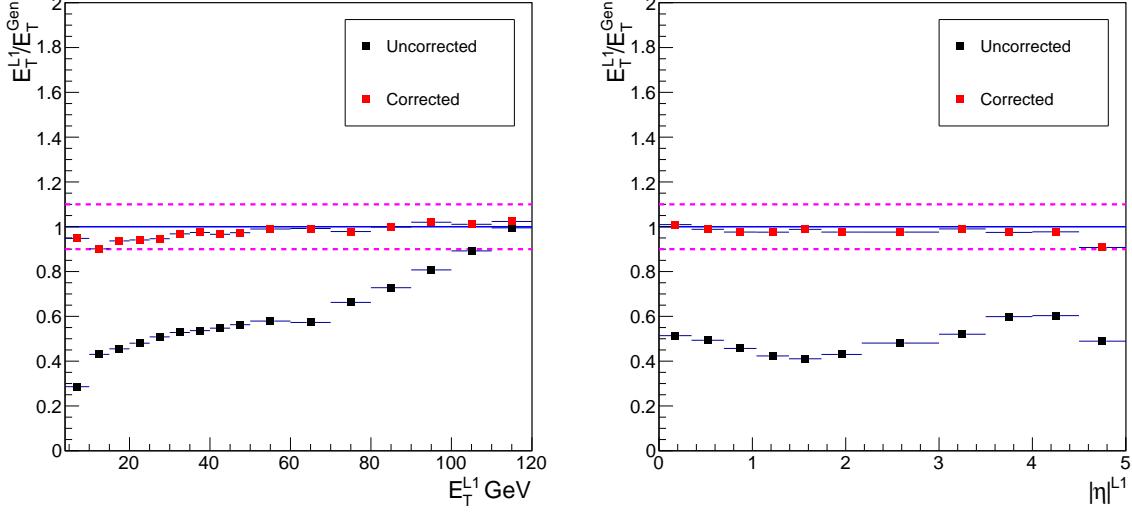
is then parameterised by performing a chi-squared fit of the functional form given in Equation 2.3.1.

$$\langle E_T^{L1}/E_T^{Gen} \rangle^{-1} = E_T^{L1} \cdot \left( p_0 + \frac{p_1}{(\log E_T^{L1})^2 + p_2} + p_3 \exp(-p_4(\log E_T^{L1} - p_5)^2) \right) \quad (2.3)$$

The functional form chosen is the same as that used for offline jet calibration at CMS [? ]. The parameterisation provides a multiplicative correction to be applied to L1 jets online. Figure 2.12 is an example of the fit in the  $0.348 < |\eta^{Gen}| < 0.695$  bin.

### 2.3.2. Calibration Performance

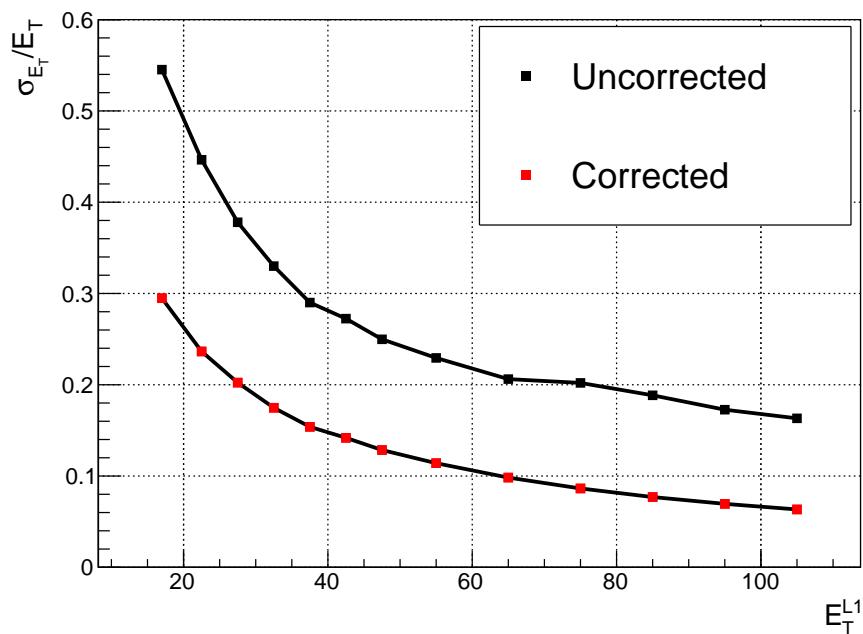
The calibrations derived were applied using the GCT emulation software to the same MC sample used to derive them to provide a closure test of their performance. The response is shown in Figure 2.11 as a function of  $E_T^{L1}$  and  $\eta^{Gen}$ . The results show that



**Figure 2.11.:** Closure tests performed in MC as a function of  $E_T^{L1}$  (left) and  $\eta^{Gen}$  (right). The test shows that after applying the corrections, the response is within 10% (dashed lines) of unity.

the procedure closes to a precision between 5% and 10%. The improvement in L1 jet resolution expected from MC is demonstrated in Figure 2.12. The resolution, calculated by fitting a Gaussian to the distribution of the difference in  $E_T$  measured at L1 to that of the generator jet in bins of  $E_T^{L1}$ , for L1 jets is shown before and after applying the corrections.

The calibrations were applied to data during the Run2011B data taking period at CMS and were found to improve the resolution of the L1 jets for all values of  $E_T^{L1}$ . An independent set of corrections were derived and applied during Run2011A but were found to give worse performance than the ones described here in particular for jets with  $E_T > 130$  GeV [20].



**Figure 2.12.:** Jet energy resolution at L1 as a function of  $E_T^{L1}$  before and after application of the derived calibrations.

# Chapter 3.

## Higgs Decay to Two Photons

*The two photon channel is one of the most promising decay modes in the search for the SM Higgs at the LHC. Despite having a relatively small branching ratio, the decay  $H \rightarrow \gamma\gamma$  provides a very clean, fully reconstructible 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,  $pp \rightarrow \gamma\gamma$  (prompt-prompt). In addition, there is a contribution from  $pp \rightarrow \gamma + \text{jet}$  (prompt-fake) and  $pp \rightarrow \text{jet} + \text{jet}$  (fake-fake) in which jets are mis-identified as photons. As the signal rate in the  $H \rightarrow \gamma\gamma$  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 techniques were developed. The first uses a fully parametric description of the background from data designed to incorporate systematic uncertainties as additional degrees of freedom in its functional form [23]. The second uses a binned model constructed from sidebands in the  $m_{\gamma\gamma}$  spectrum. The latter of these two serves as an independent cross-check of the former, in particular by allowing direct inclusion of the systematic uncertainties in the signal extraction thereby 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  $5.1 \text{ fb}^{-1}$  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 satisfy 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 [27]. 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 integrated luminosity of  $5.1 \text{ fb}^{-1}$ . However, the thresholds applied offline are chosen to be much tighter than those of the trigger; the trigger efficiency is  $>99\%$  with respect to the analysis selection [27].

Signal Monte Carlo (MC) events are generated for a Higgs decaying to two photons via the four main production processes, gluon-gluon fusion ( $ggH$ ), vector boson fusion ( $qqH$ ) and associated  $W/Z$  ( $VH$ ) and  $t\bar{t}$  ( $ttH$ ) production. The gluon-gluon fusion and vector boson fusion were generated with **POWHEG** [36] 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 [19]. The production cross-sections and branching ratios are taken from the LHC Cross-section Working Group [32].

MC for background processes were generated at LO using **POWHEG** interfaced with **PYTHIA** [40]. The QCD dijet and  $\gamma + jet$  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 [17]. The MC includes a simulation of additional interaction vertices expected in data from pileup. The distribution in the number of reconstructed vertices in MC is corrected to match that observed in data as described in Section ??.

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

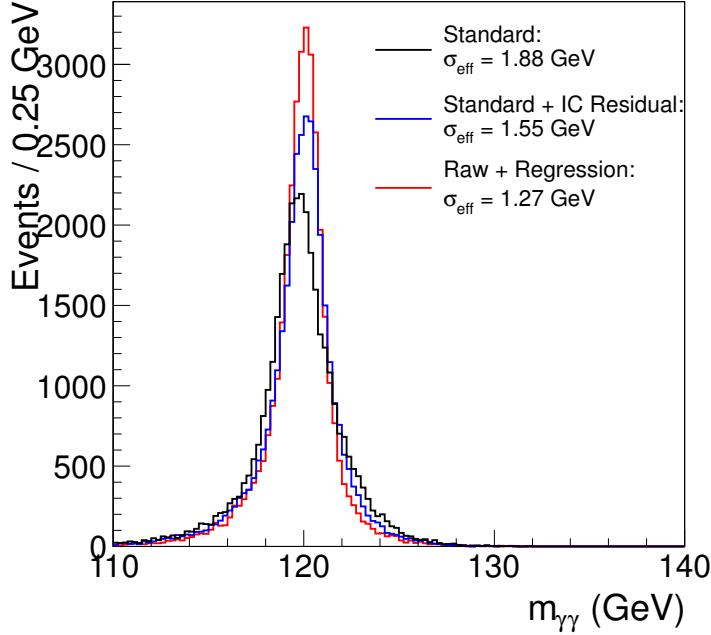
**Table 3.1.:** Background MC used throughput the analysis with production cross-sections and corresponding equivalent integrated luminosity.

## 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` [35]. Additional sensitivity can be gained by refining the object selection and reconstruction specifically to the search for  $H \rightarrow \gamma\gamma$ .

### 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  $H \rightarrow \gamma\gamma$  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 variables include both the  $\eta$  and  $\phi$  positions of the supercluster. In addition, several variables are included which describe the shower shape:  $r_9$ , the energy weighted widths in  $\eta$  and  $\phi$  of the supercluster, the energy weighted crystal width ( $\sigma_{i\eta i\phi}$ ) and the ratio of hadronic energy behind the supercluster to the energy of the supercluster itself ( $H/E$ ). In the endcaps, there is additional information available from the pre-shower measurement. The ratio of the energy in the pre-shower to the raw supercluster energy is included for superclusters in the ECAL endcaps. 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



**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 regression method. The quantity  $\sigma_{\text{eff}}$ , the narrowest range in  $m_{\gamma\gamma}$  which contains 68% of the distribution, is given for each peak.

of the residual energy difference between the generated and reconstructed photon energy as a function of supercluster energy, position and  $r_9$  [25]. The regression technique reduces the effective resolution of the Higgs mass peak ( $\sigma_{\text{eff}}$ ) 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,  $\sigma_E$ , is obtained by training a second regression BDT targeting 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 independent set of events to the first. The per-photon resolution is used to calculate 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  $Z \rightarrow e^+e^-$  MC which is used to compare the supercluster energy scale in data and MC [27].

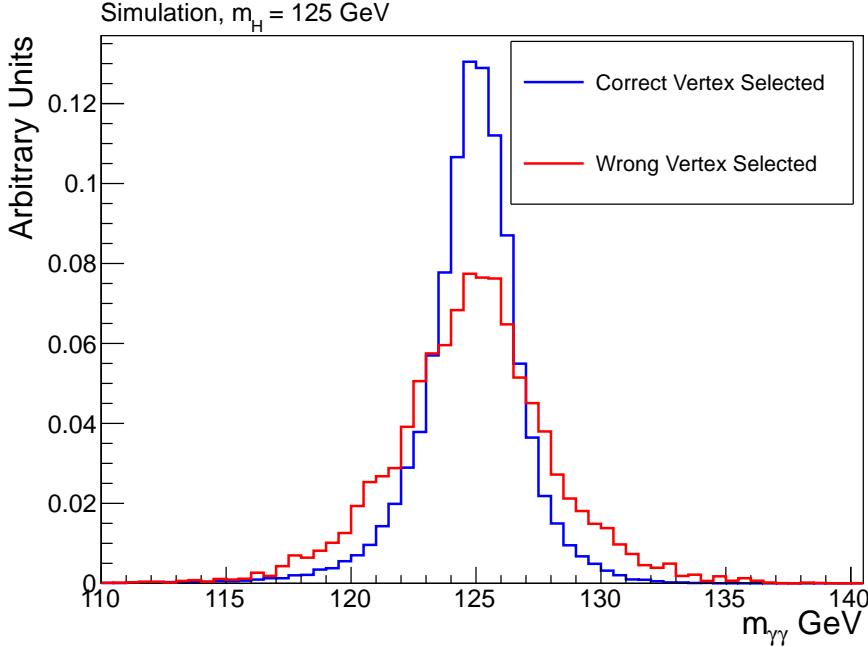
## Energy Scale Measured in Data

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 dependence of the ECAL crystal transparency [35]. Further corrections are derived based on  $Z \rightarrow e^+e^-$  events which provide 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 achieved using an analytic fit to the  $Z \rightarrow e^+e^-$  peak in data and MC separately. The natural peak of the  $Z$  is described using a Breit-Wigner distribution whose parameters are fixed to those given by the Particle Data Group,  $m_Z = 91.188$ ,  $\Gamma_Z = 2.495$  [18]. 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 [41]. The CB parameter  $\Delta m$  is a free parameter of the fit giving the offset of the peak position from the  $Z$  pole.

The values of these fitted parameters varies with the position 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,

$$1 - \Delta P = 1 - \frac{\Delta m_{data} - \Delta m_{MC}}{m_Z} \quad (3.1)$$

is applied to the photons in data. The values for the scale in each category,  $\Delta P$ , 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 systematics are included due to the variation of the measurements when changing the electron selection and between using the electron-trained and photon-trained regression



**Figure 3.2.:** Invariant mass peak in  $H \rightarrow \gamma\gamma$  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.

corrections. These uncertainties are incorporated 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 boson 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 vertices. 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,

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

and

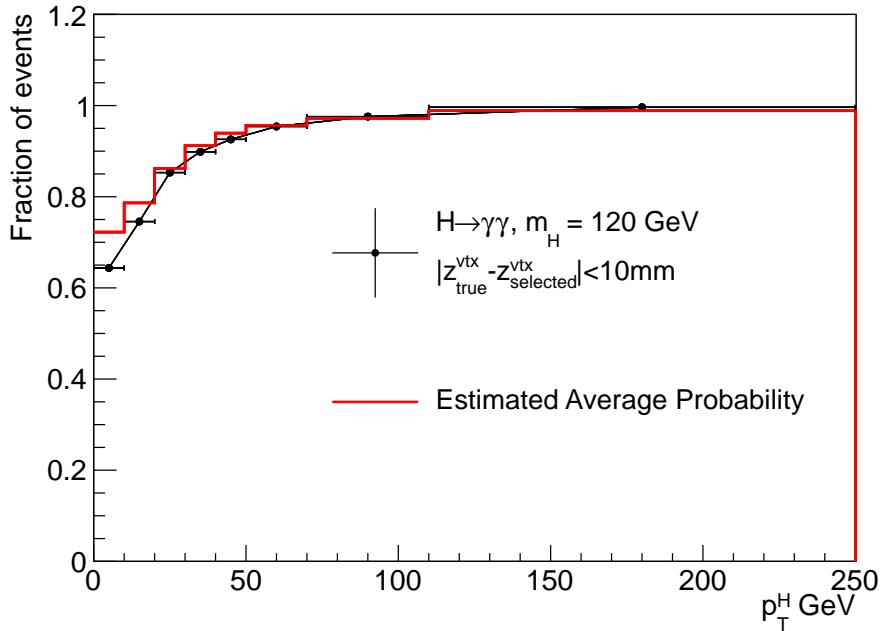
$$\frac{\left| \sum_{alltracks} \mathbf{p}_T^{track} \right| - p_T^{\gamma\gamma}}{\left| \sum_{alltracks} \mathbf{p}_T^{track} \right|} \quad (3.3)$$

respectively. 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 calculated 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 trained on  $H \rightarrow \gamma\gamma$  MC with a mass of 120 GeV. Figure ?? 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  $Z \rightarrow \mu^+\mu^-$  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 against the fraction of the events in which the selected vertex is located within 10mm from the true vertex.

### 3.2.3. Photon Identification

A large portion of the fake background in the  $H \rightarrow \gamma\gamma$  search is due to high momentum neutral mesons which decay to two photons where both the photons are combined into the same supercluster [23]. Information from the shower shape of the photon supercluster can be used, in addition to the energy isolation within the calorimeter, in order to distinguish these from prompt photons from the primary interaction point. A BDT was trained on MC events to combine the relevant information into a single photon identification (ID)



**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.

discriminator. The signal used for the training was taken from simulated  $H \rightarrow \gamma\gamma$  events with a Higgs boson mass of 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 [27]. 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  $Z \rightarrow e^+e^-$  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  $Z \rightarrow e^+e^-$  data and MC using a tag-and-probe method [? ]. The results are shown in Table 3.2.

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  $\eta$  and  $\phi$  ( $\sigma_\eta$ ,  $\sigma_\phi$ ). The  $\eta$  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 $r_9 > 0.9$	$0.9267 \pm 0.0012$	$0.9275 \pm 0.0006$	$0.999 \pm 0.0013$
EB $r_9 < 0.9$	$0.8882 \pm 0.0023$	$0.9025 \pm 0.0010$	$0.984 \pm 0.0025$
EE $r_9 > 0.9$	$0.9442 \pm 0.0010$	$0.9387 \pm 0.0009$	$1.006 \pm 0.0014$
EE $r_9 < 0.9$	$0.8639 \pm 0.0010$	$0.8517 \pm 0.0011$	$1.014 \pm 0.0015$

**Table 3.2.:** Signal efficiency for the preselection measured in data and MC using tag-and-probe in  $Z \rightarrow e^+e^-$  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,  $\rho$  [24]. 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  $\rho$  times the effective area of the cone [24], 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 dependence 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 chosen to be loose as the output of the photon ID will be used as input for the event selection (diphoton BDT) as described 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 photon respectively. Where more than one diphoton pair satisfies 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 diphotons from QCD processes. In order to distinguish these from a Higgs signal, the specific kinematics and topology of the event are exploited.

### 3.3.1. Diphoton BDT

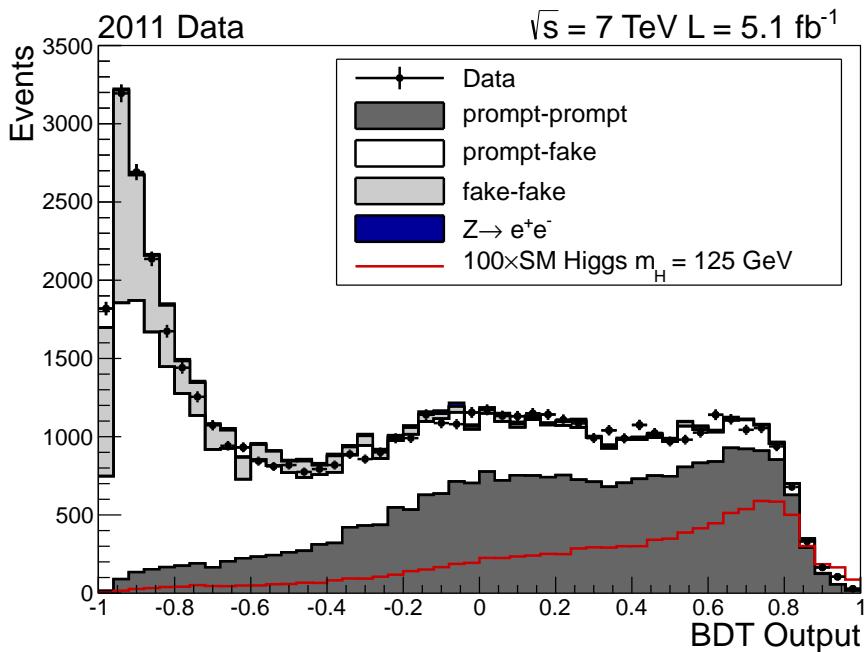
A BDT was trained to utilise the kinematics of the selected diphoton pair to discriminate prompt photons from QCD background from those produced by the decay  $H \rightarrow \gamma\gamma$ . The BDT was trained using the QCD Dijet, Gamma+Jet, DiphotonJets and Diphoton Born samples for background and Higgs MC with a mass of 123 GeV. As the mass of the Higgs boson is unknown, 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-independent. In addition, this allows for a fully data-driven estimation of the background shape as described in Section 3.4.4. The input variables which describe the kinematics are: the relative transverse momenta of the photons,  $p_T^1$ ,  $p_T^2$ , their pseudo-rapides,  $\eta^1$ ,  $\eta^2$  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,  $\sigma_E$  is combined for each photon to produce a per-event mass resolution estimate ( $\sigma_{m_{\gamma\gamma}}$ ) under the assumption that the correct vertex is selected;

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

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

Since the correct vertex is not always selected, the mass resolution assuming the incorrect vertex is chosen is calculated using the average beamspot length in data,  $\sigma_Z = 5.8\text{cm}$ . 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 analytically 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 [27].

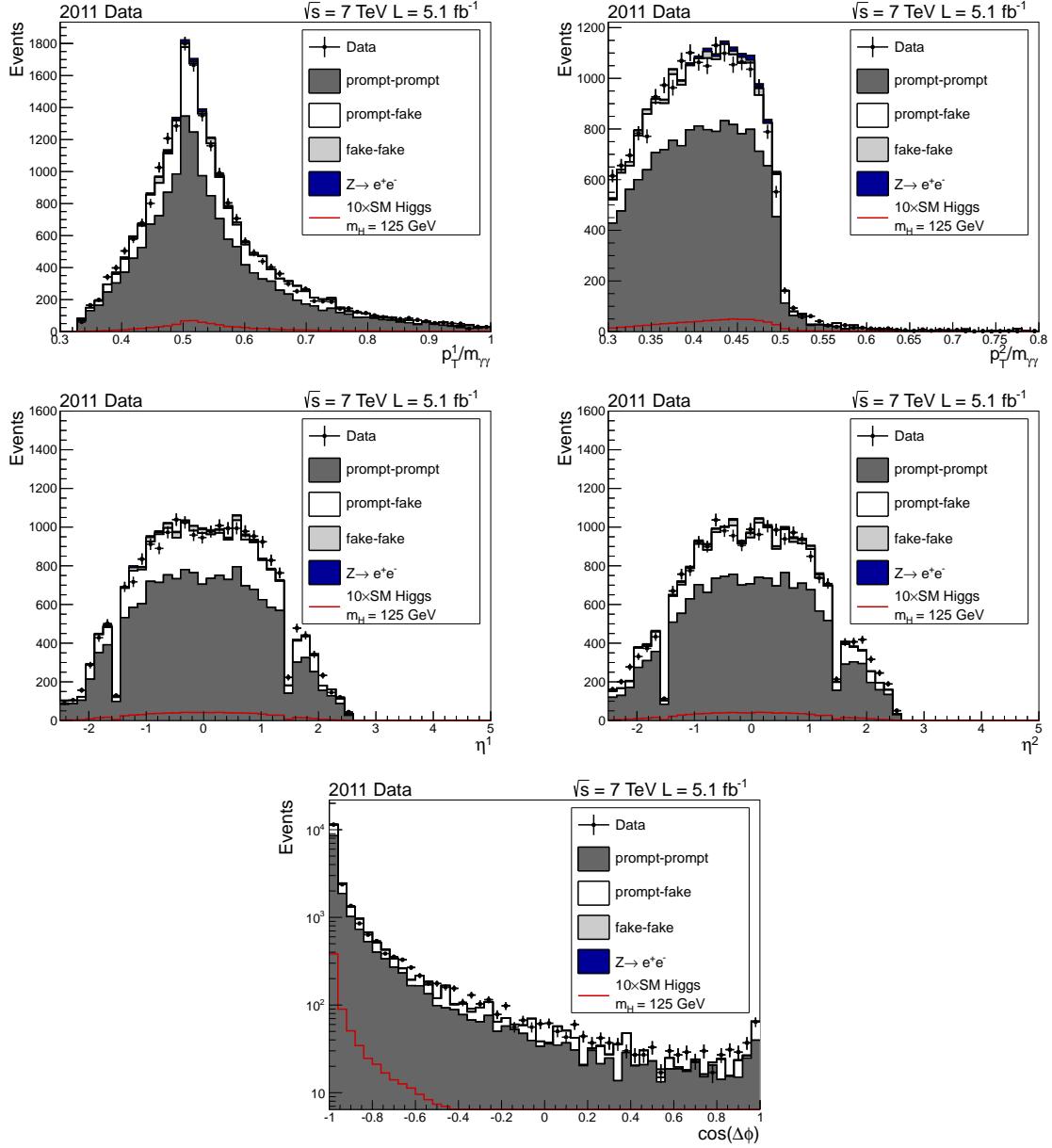


**Figure 3.4.:** Diphoton BDT distribution in data and MC. The contribution expected from a SM Higgs with mass 125 GeV, scaled by 100, is shown in red.

Figures ?? and ?? 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. The invariant mass distribution in data and MC for events passing the full selection is given in Figure 3.7. After the application of the full selection, the total background contains around 76% prompt diphoton events.

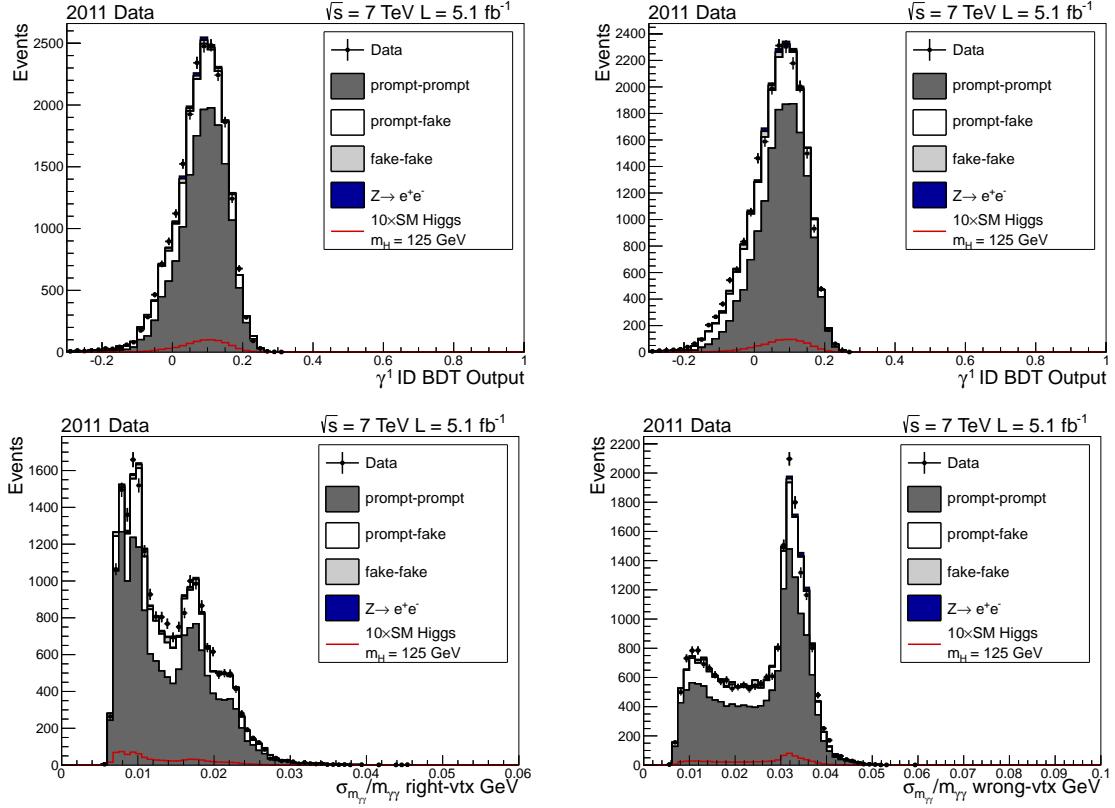
### Diphoton BDT Validation with $Z \rightarrow e^+e^-$ Data

By using a BDT for the full event selection, subtle correlations between the input variables are accounted for which improves the separation between the signal and background. Unlike



**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.

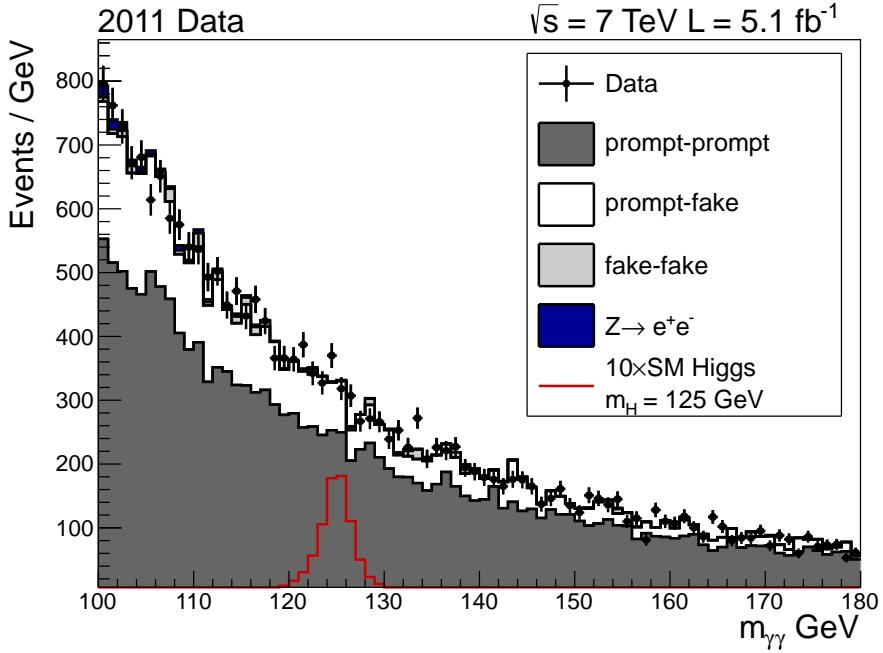
the background model, the signal model will be taken from corrected MC simulation. It is important therefore to ensure that the BDT will respond in the same way in data as for the signal MC used for the signal extraction. The MC can be validated using  $Z \rightarrow e^+e^-$  data-MC comparisons by inverting the electron veto and treating the electrons as though they were photons. This is done by using the supercluster associated to the electron for



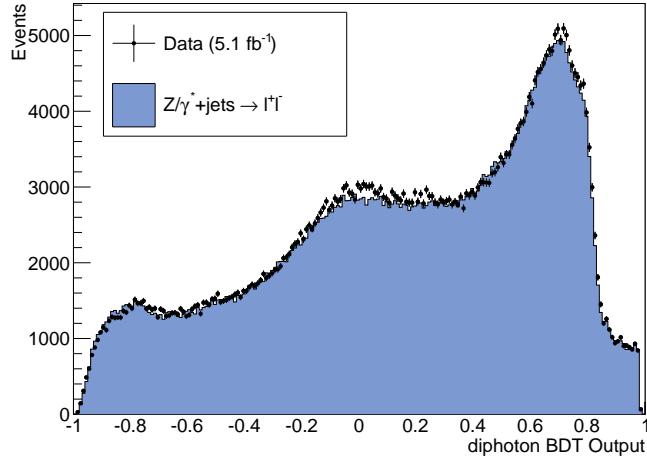
**Figure 3.6.:** Additional diphoton BDT input variable distributions in data and MC. The distributions are for events which pass the full selection including a cut on the diphoton BDT output of 0.05. The expectation from a SM Higgs with 125 GeV is shown in red.

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 for validation of the BDT's response to real photons from a resonance decay [35]. Figure ?? shows the diphoton BDT distribution in  $Z \rightarrow e^+e^-$  MC and data after applying the full selection using this technique.

Both the photon ID and regression BDT rely on a detailed simulation of electromagnetic showering in MC to correctly describe the data. Due to imperfections 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. These uncertainties are validated using  $Z \rightarrow e^+e^-$  in the same way as the diphoton BDT. Figures ?? and ?? show the distributions of the per photon energy resolution estimator  $\sigma_E$  relative to the photon energy and the output of the photon ID BDT in  $Z \rightarrow e^+e^-$  MC and data treating the electrons as photons. The red lines show the  $\pm 1\sigma$  error envelope attributed to the

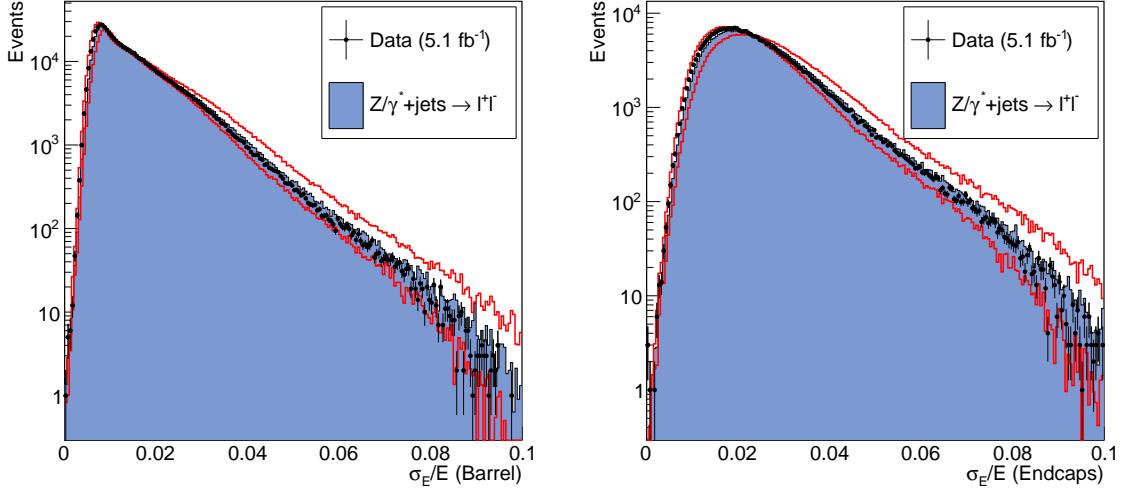


**Figure 3.7.:** Invariant mass distribution in data and MC after applying the full event selection in the range 100 to 180 GeV. The contribution expected from a SM Higgs with mass 125 GeV, scaled by 10, is shown in red.

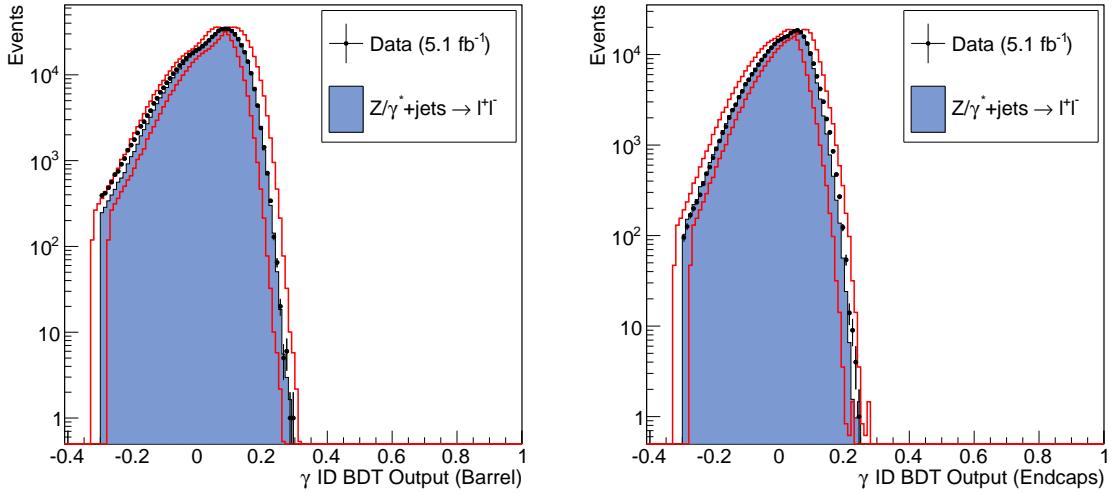


**Figure 3.8.:** Diphoton BDT output distribution in  $Z \rightarrow e^+e^-$  MC and data after the full selection treating the electrons as photons for the purposes of energy reconstruction. The electron veto is inverted to preferentially select electrons.

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.



**Figure 3.9.:** Upper: Per-photon resolution estimator,  $\sigma_E$  relative to the measured energy in  $Z \rightarrow e^+e^-$  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  $\sigma_E$  by  $\pm 10\%$ .



**Figure 3.10.:** Photon ID BDT output in  $Z \rightarrow e^+e^-$  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  $\sigma_E$  by  $\pm 0.025\%$ .

Variable	Cut Value
$E_T^{j^1}$	$> 30 \text{ GeV}$
$E_T^{j^2}$	$> 20 \text{ GeV}$
$m_{jj}$	$> 350 \text{ GeV}$
$ \eta_{j^1} - \eta_{j^2} $	$> 3.5$
$ \phi_{jj} - \phi_{\gamma\gamma} $	$> 2.6$
$ \frac{1}{2}(\eta_{j^1} + \eta_{j^2}) - \eta_{\gamma\gamma} $	$< 2.5$

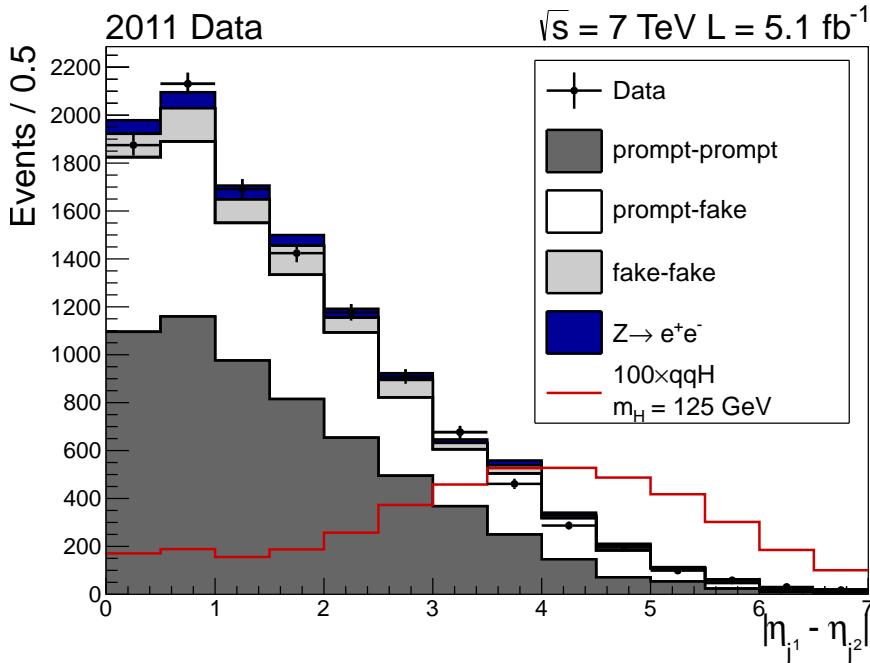
**Table 3.3.:** Dijet selection criteria for the two identified jets to be considered likely associated to  $qqH$  production. The leading and subleading  $E_T$  jets are denoted  $j^1$  and  $j^2$  respectively.

### 3.3.2. Dijet Tagging

The contribution to Higgs production from vector boson fusion is around a factor ten smaller than that of gluon-gluon fusion. However, additional information from the two jets associated with  $qqH$  production allows for further reduction of the diphoton background ???. Events containing two jets which pass the full selection and in addition satisfy a series of criteria designed to target the specific topology of the dijet system are tagged as likely to have originated from  $qqH$  production. For example, Figure ?? shows the separation in  $\eta$  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 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.

## 3.4. Signal Extraction

The signature for the decay  $H \rightarrow \gamma\gamma$  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,  $m_{\gamma\gamma}$ .



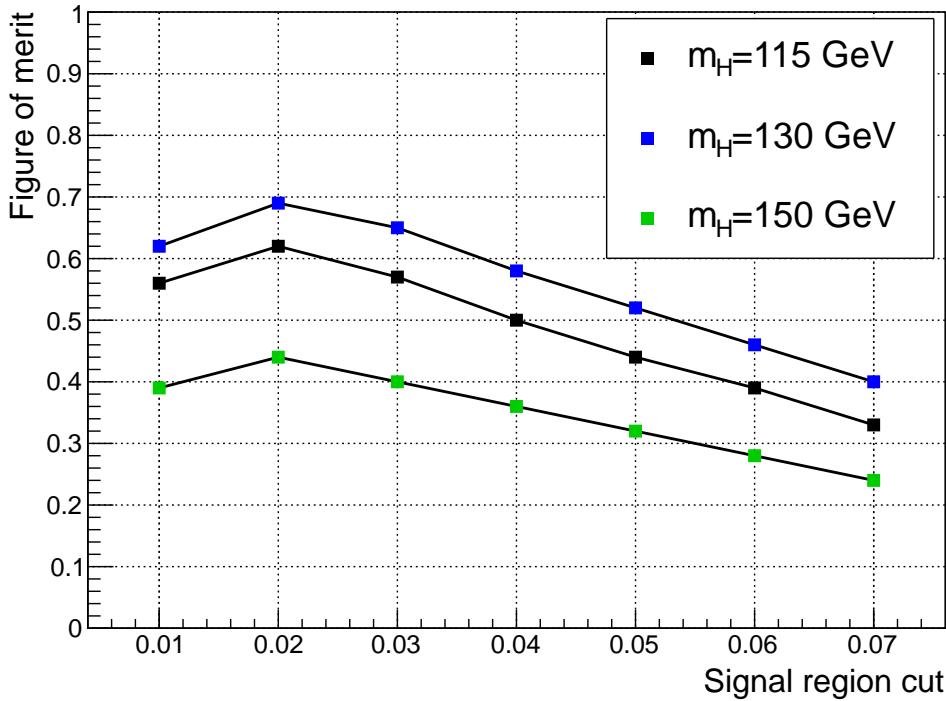
**Figure 3.11.:** Separation in  $\eta$  between two identified jets in data and MC. The expectation from a SM Higgs produced via vector boson fusion ( $qqH$ ), scaled by 100, is shown in red. All cuts other than the one on  $\Delta\eta(Jet1, Jet2)$  are applied to these distributions.

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 multivariate analysis (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



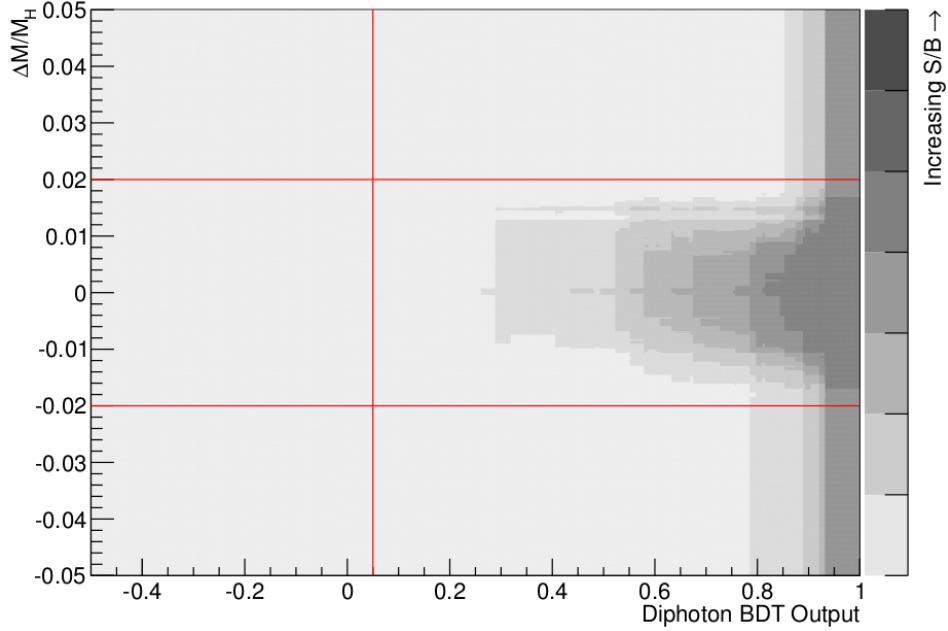
**Figure 3.12.:** Figure of merit for selection of the signal region cut value,  $w$ . Each color shows the evaluation under different Higgs mass hypotheses.

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 number of background events,  $\sigma_B$ , is calculated using the procedure described in Section 3.4.4 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. Event Categorisation BDT

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 BDT 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 3.13 shows the variation in the signal to background ratio ( $S/B$ ) across different regions in the two-dimensional plane defined by the output of the diphoton BDT output

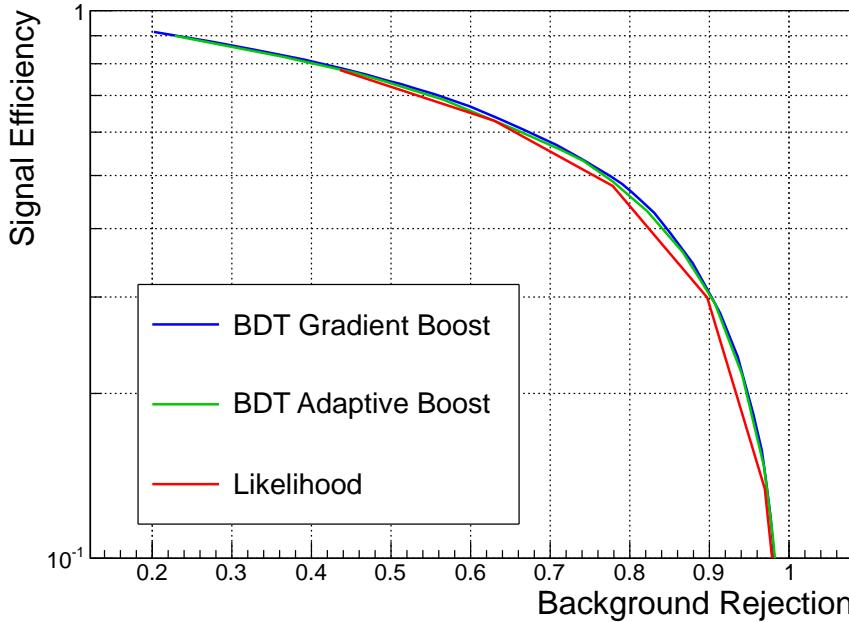


**Figure 3.13.:** Signal to background ratio as a function of diphoton BDT output and  $\Delta m/m_H$ . The red lines indicate the cuts applied before the training and for applying the event selection.

and  $\Delta m/m_H$ . Events close to centre of the peak ( $\Delta m/m_H = 0$ ) with a high score in the diphoton BDT are more likely signal events than those far from the high  $S/B$  regions.

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, adaptive and gradient boosting, both of which weight decision trees to optimize the performance in terms of signal-background separation [30]. 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



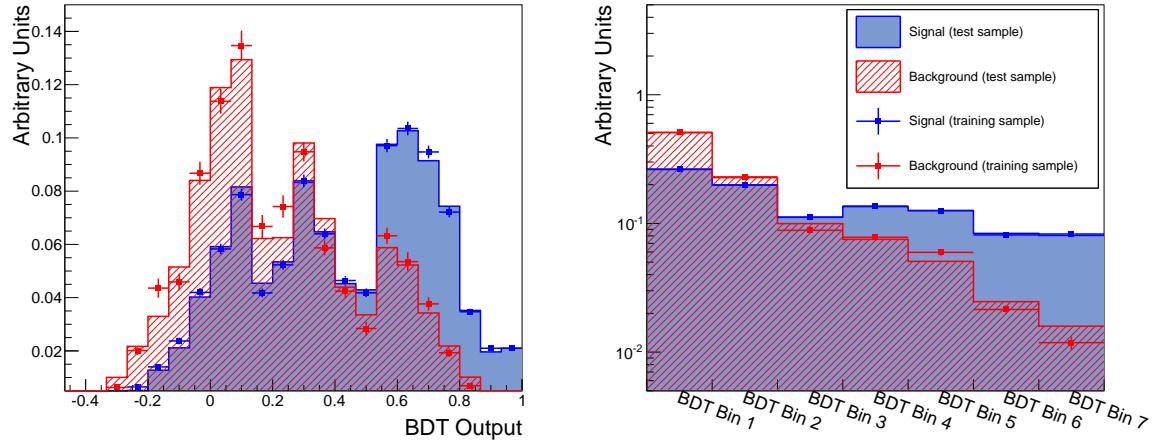
**Figure 3.14.:** Signal efficiency vs background rejection curves for three different MVA techniques used to train the signal-background event discriminator. The curves give the (in)efficiencies for signal (background) after applying sequentially tighter cuts on the discriminator output.

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 the final set of bins derived in Section 3.4.3. A  $\chi^2$  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.

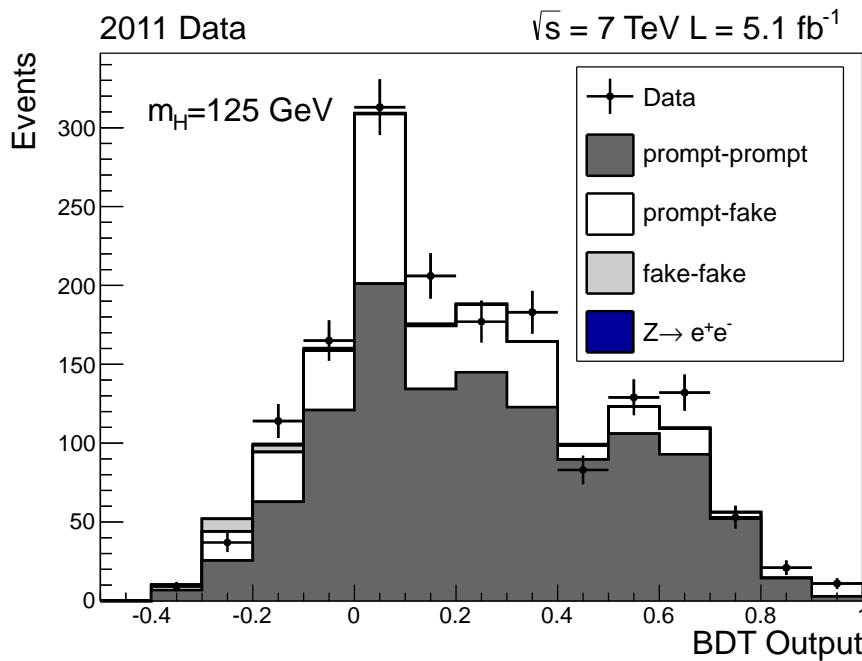
In this analysis, the background is estimated entirely from data. This means that any disagreement between data and MC will only effect the performance of the BDT and not the validity of the final results. The agreement between the data and MC is shown in Figure 3.16 for a mass hypothesis,  $m_H = 125$  GeV. The level of agreement is sufficient so as not to require in-depth study of the BDT output distributions of the background MC.

### 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



**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 the bins used for extracting the signal (right).



**Figure 3.16.:** Comparison of the distributions of BDT output at  $m_H = 125$  for data and background MC. The distributions are arbitrarily binned for the purposes of comparison only.

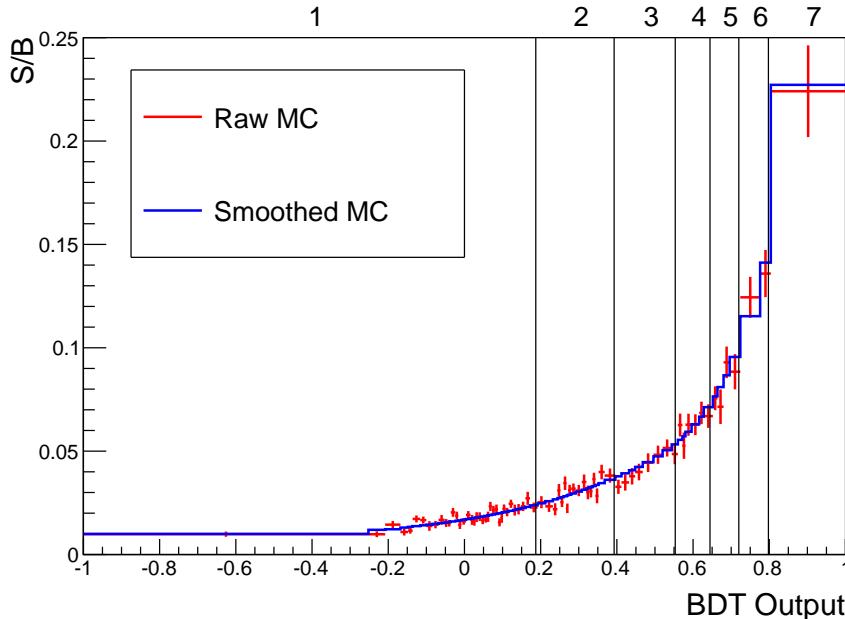
on the number of trees used. The boosting procedure provides a pseudo-continuous distribution which is used to model the signal and background. However, the resulting distribution will still be only pseudo-continuous. In addition, the BDT response does not directly correspond to a physical distribution 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. These requirements are desirable such that the expected background yield in each bin can be derived using data outside of the signal region as described in Section 3.4.4.

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,  $\sigma_{exp}$ , is given by

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

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  $5.1 \text{ fb}^{-1}$ .
2. Smoothed versions of the signal (at each 5 GeV step mass) and background MC templates 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.
3.  $N$  bin edges (boundaries),  $b_i$ , are defined on the remaining bins such that  $N + 1$  bins are formed with  $b_1 < b_2 < \dots < b_N$ . 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.



**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.

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 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 SM background is expected to have a smoothly varying invariant mass spectrum. However, detector effects such as selection, trigger efficiencies and energy resolution shape this distribution in ways which are imperfectly modelled in MC simulation. Moreover, the background contains fakes whose contribution varies 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 for modelling the background is used.

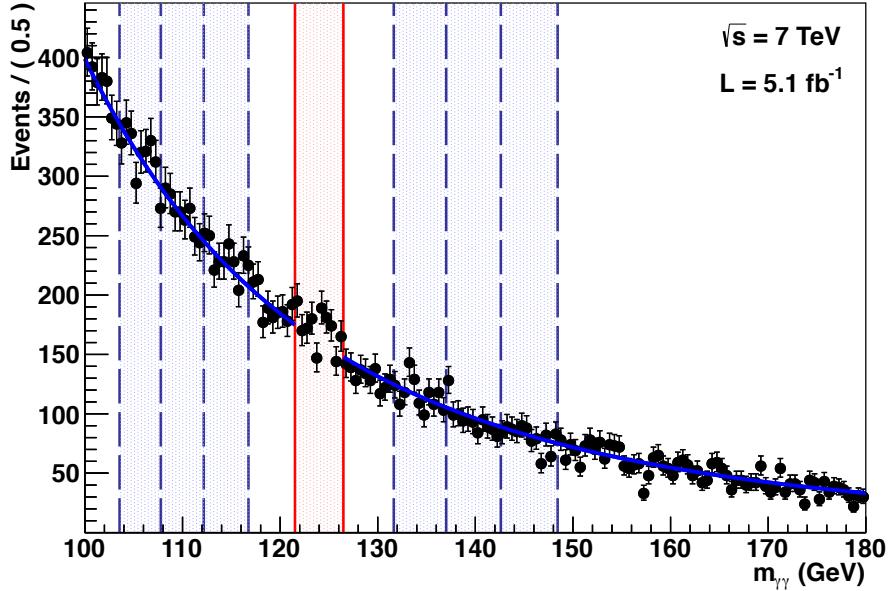
For a given mass hypothesis, the shape and normalization of the background model are obtained separately. The shape, meaning the fraction of events in each BDT output bin, is extracted from the BDT output distributions in mass-sidebands, while the overall normalization is obtained from a parametric fit 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 determine the shape of the background model. The blue line indicates the fit of a double power law used to determine the normalisation of the background in the signal region.

#### 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  $\pm 2\%$  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 particular parameterization used 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



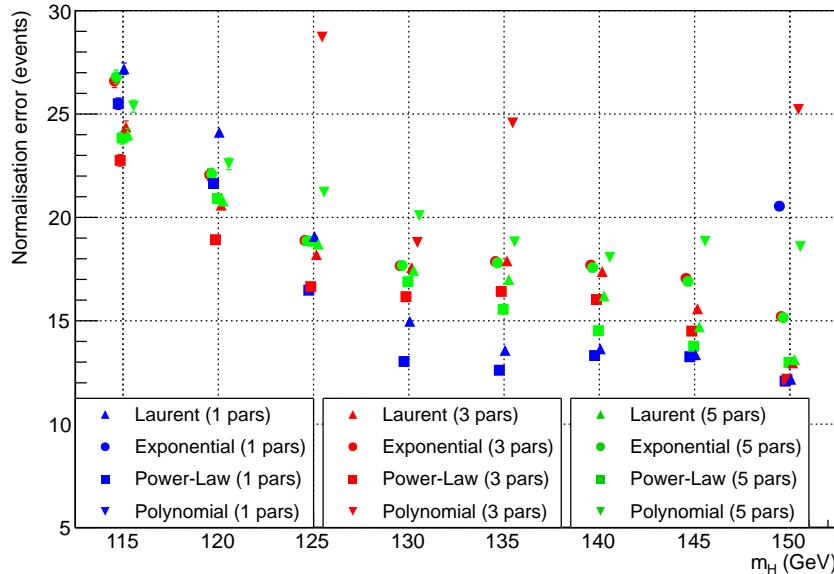
**Figure 3.18.:** Invariant mass distribution of the full 2011 dataset after selection over the mass range used in the analysis (100 to 180 GeV). The  $\pm 2\%$  signal region for  $m_H = 124$  is indicated in red, while the six corresponding sidebands are indicated as blue bands. The blue line is the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

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 assess 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



**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.

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 to the data 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. This study is repeated at 5 GeV intervals in  $m_H$  as the overall uncertainty varies as a function of mass hypothesis. Figure 3.19 shows the total error determined for each of the twelve functions at each value of  $m_H$  tested. 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 3.4.6).

## Obtaining the Shape of the Background

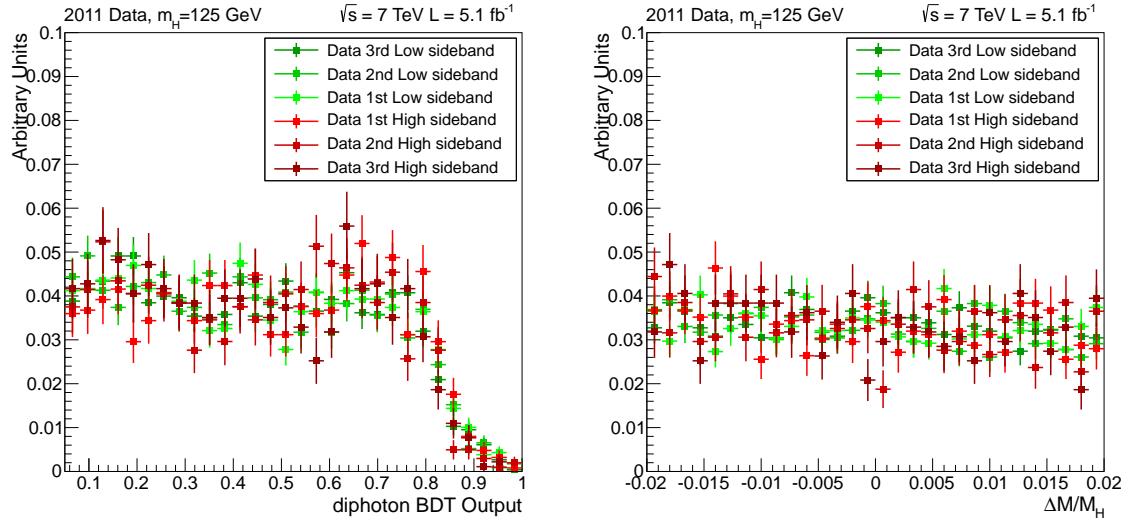
As the signal yield expected from a SM Higgs is small compared to the background, the sensitivity of the search is strongly dependant on how well the relative contribution from the background in each bin is understood. Both inputs to the BDT are designed to be insensitive to the invariant mass of the 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  $w = 0.02$ .

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

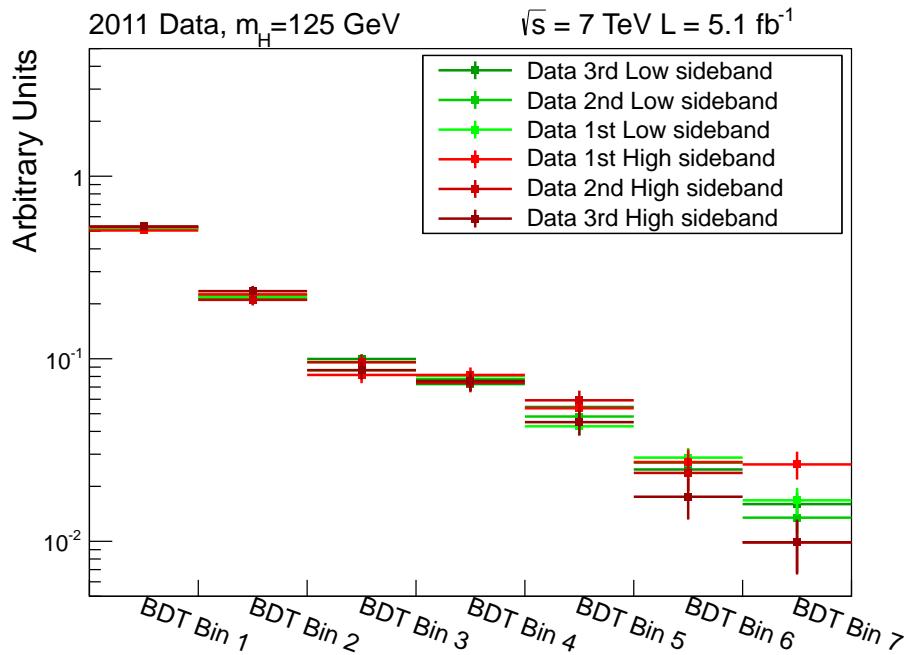
The two sidebands adjacent to the signal window (corresponding to  $i = \pm 1$  in Equation 3.6) are not used in order to avoid signal contamination. Dijet tagged events are treated in the same way as the rest of the events by introducing an eighth bin containing dijet tagged events inside the range  $\Delta m/m_H < w$ . The distributions for the two input variables, diphoton BDT output and  $\Delta m/m_H$ , for each of the six sidebands corresponding to  $m_H = 125$  are shown in Figure 3.20. 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 BDT 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  $j \in \{1, 8\}$  and  $i \in \{\dots, -4, -3, -2, 2, 3, 4, \dots\}$ .

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



**Figure 3.20.:** Distribution in data from the six sidebands corresponding to  $m_H = 125 \text{ GeV}$  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  $m_H = 125 \text{ GeV}$  of the BDT output binned in the 7 BDT output bins used for signal extraction.

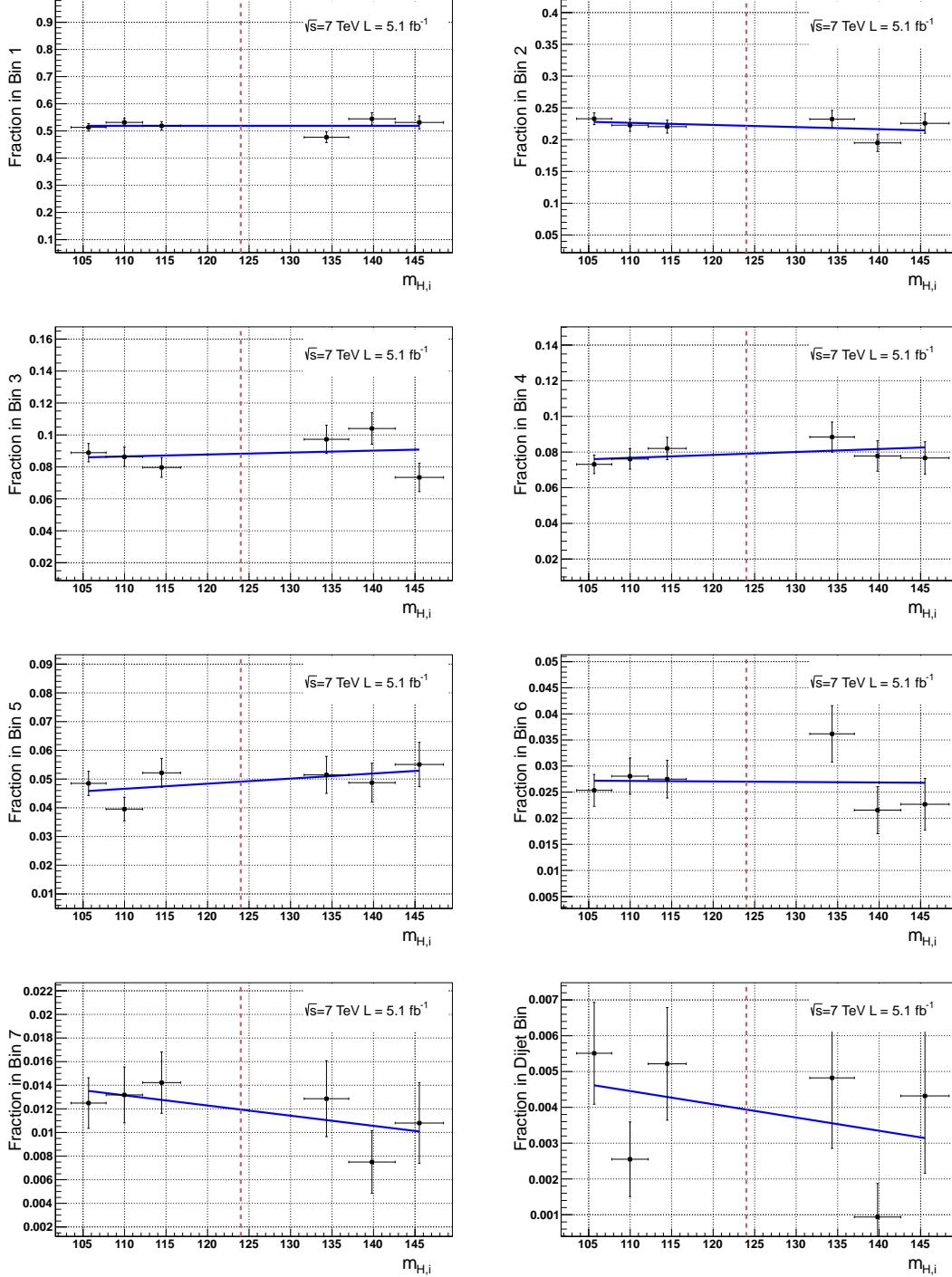
Since the normalisation for the background model is determined independently, the sum over all bins is constrained to be one. The expectation value for the background in each bin,  $j$ , is then determined as  $Nf_j$  where  $N$  is the normalisation estimated in section 3.4.4. This constraint is imposed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

$$p_{0,1} = 1 - \sum_{i=2}^8 p_{0,j} \quad p_{1,1} = - \sum_{j=2}^8 p_{1,j} \quad (3.8)$$

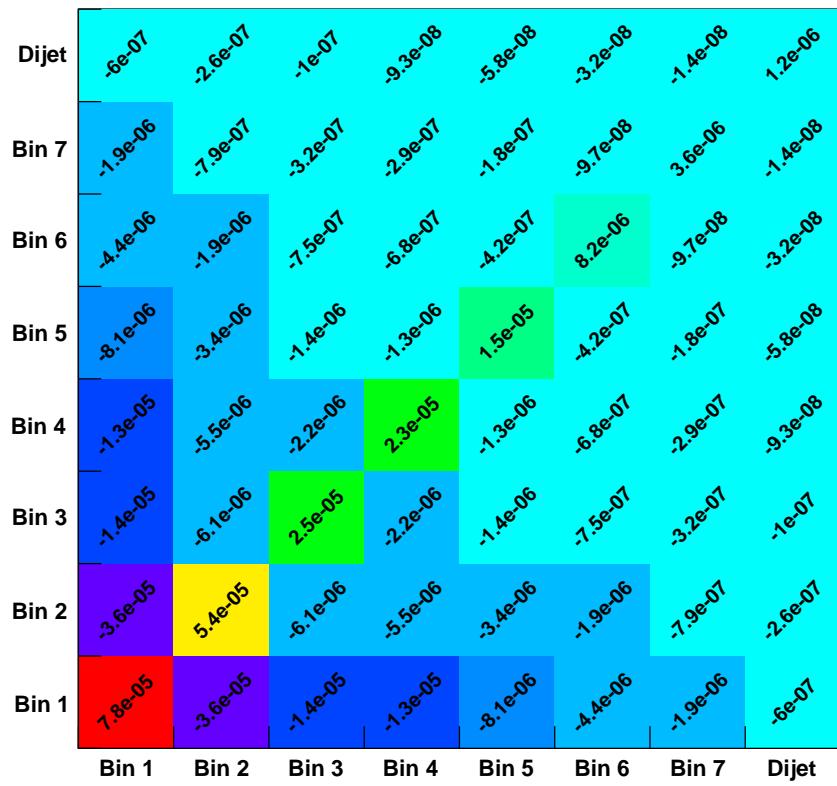
The coefficients  $p_{0,j}, p_{1,j}$  of Equation 3.7 are determined by performing a binned maximum likelihood fit to the observed fractions in the data assuming the contents of each bin in each sideband are Poisson distributed. The results of the fit for  $m_H = 124$  GeV are shown in Figure 3.22 and the resulting covariance matrix obtained is shown in Figure 3.23. The fit was performed using TMinuit under ROOT 5.2.0 [31].

There are seven degrees of freedom (eight bins minus one constraint) which are correlated in the simultaneous fit. In order to account for the statistical uncertainty on this fit, a set of seven uncorrelated variables are determined from the covariance matrix using eigenvector decomposition [14]. These variables provide are treated as seven independent sources of systematic uncertainty on the background shape for the purpose of signal extraction (Section 3.4.6). Figure 3.24 shows the total relative fit error for each bin, at  $m_H = 130$  GeV, 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 statistical uncertainty. In order to avoid contamination from  $Z \rightarrow e^+e^-$  at the lower mass hypotheses any lower sideband whose lower boundary is less than 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.

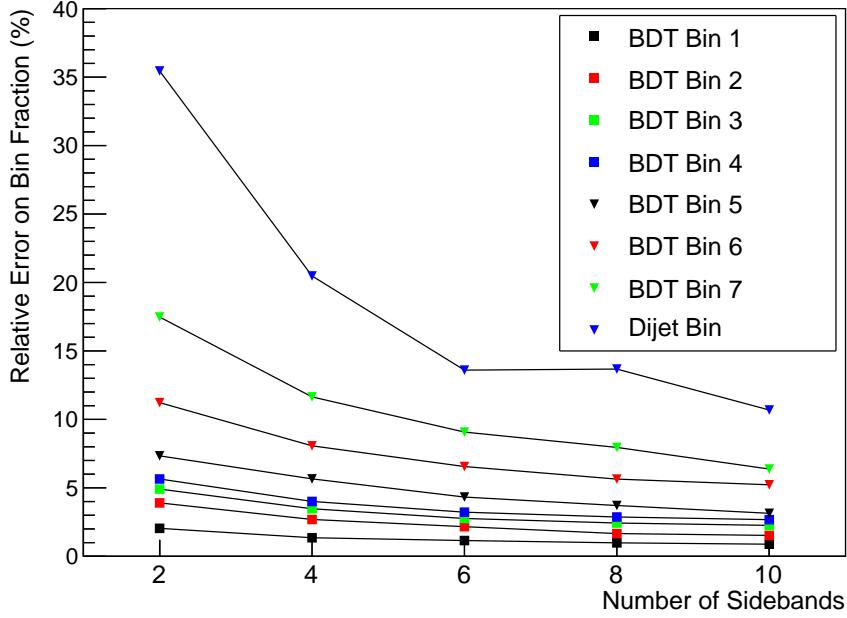
At most linear variations with mass are considered for the background BDT output distribution. This corresponds to evaluating the first term in a Taylor series for the true shape of the distribution about  $m_H$ . Higher terms can be introduced but the statistical precision of the fit will be reduced 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,



**Figure 3.22.:** Simultaneous fits to the six sidebands in data to determine the background shape for  $m_H = 124$  GeV. 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.



**Figure 3.23.:** Covariance matrix from the sideband fit to determine the background shape at  $m_H = 124$  GeV. 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.

$$f_j = p_{0,j} + p_{1,j}(m_{H,i} - m_H) + \frac{1}{2}p_{2,j}(m_{H,i} - m_H)^2. \quad (3.9)$$

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 between 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 root mean square (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 boson decay to two photons at a given mass is constructed by binning the BDT response from MC simulation of the four production processes,  $ggH$ ,  $qqH$ ,  $VH$  and  $t\bar{t}H$ . 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 and diphoton level corrections.

#### Photon Level Corrections

The energy resolution of the calorimeter is measured in data using  $Z \rightarrow e^+e^-$  events in categories defined by the position and  $r_9$  of the supercluster. 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  $\sigma_{cat}$  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 B.1.

The efficiency for a photon to pass the pre-selection is measured in  $Z \rightarrow e^+e^-$  data in four categories. These are defined by whether or not the supercluster is in the ECAL barrel or either endcap and the value of  $r_9$  being greater than or less than 0.94. The ratio of the efficiency measured in data to that measured in MC provides a scale factor which is applied to signal MC. Each signal event is reweighted by the product of the scale factors for each photon in the selected diphoton pair. In addition to these corrections, the value of  $\sigma_E$  and the photon ID BDT for each photon is shifted in each signal event to account for imperfections in detector simulations as described in Section 3.3.1.

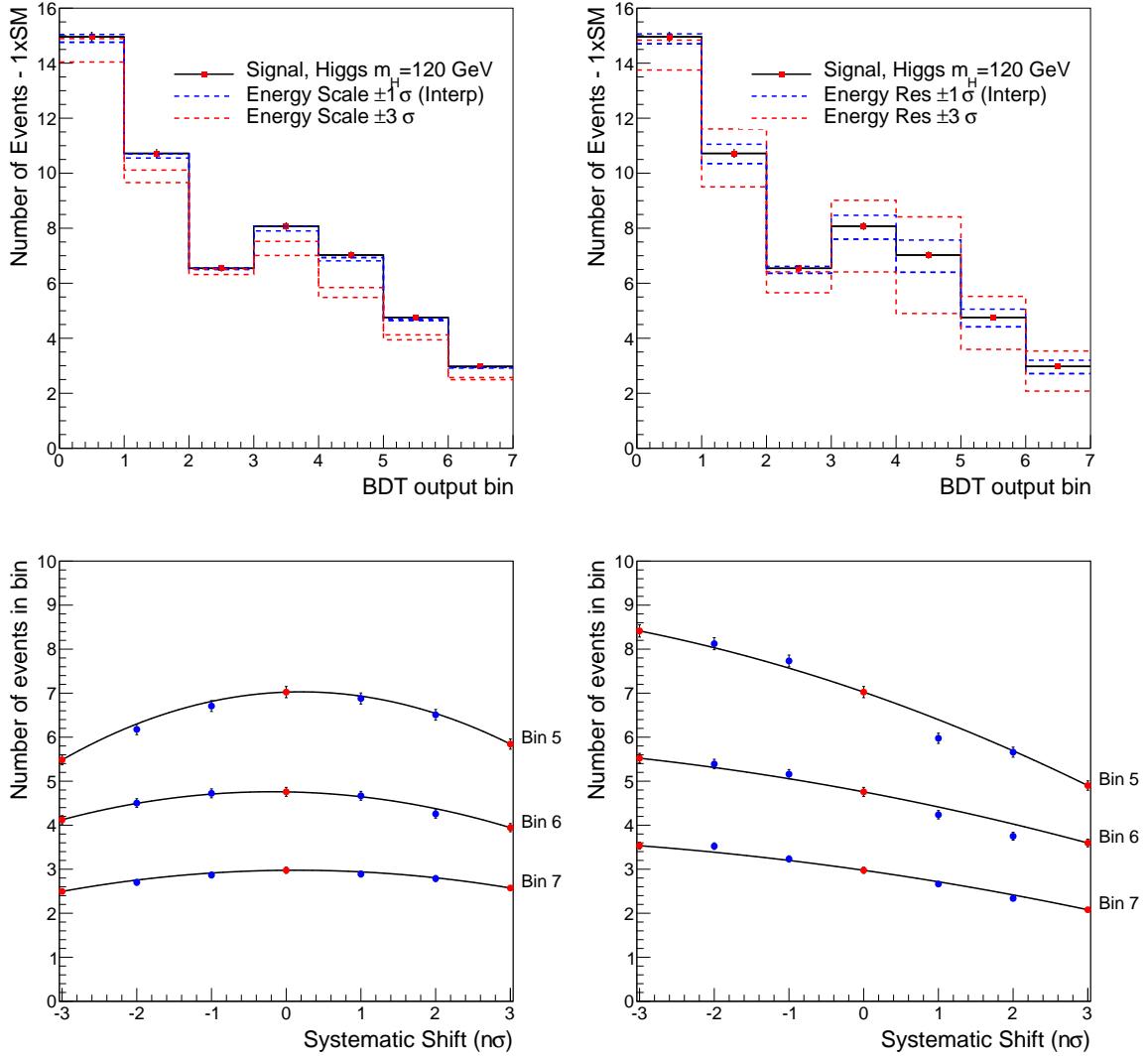
## Diphoton Level Corrections

The efficiency to select the correct vertex in the event is measured using  $Z \rightarrow \mu^+ \mu^-$  events as a function of the boson  $p_T$  as described in Section 3.2.2. 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 by the ratio of the probability that the event lies in a particular category as measured in  $Z \rightarrow \mu^+ \mu^-$  data to that measured in  $Z \rightarrow \mu^+ \mu^-$  MC. The L1/HLT efficiency is measured in four diphoton categories depending on the maximum supercluster  $\eta$  and minimum  $r_9$  value of the two photons using  $Z \rightarrow e^+ e^-$  data. As the simulation does not include the trigger, the efficiency is applied directly as a weight to each MC event.

## 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 variation due to that uncertainty. In practise, these quantities are derived by applying shifts to the MC corresponding to  $3\sigma$  variation of each uncertainty and interpolating the difference from the nominal values back to the  $1\sigma$  level. This is done so that the evaluation of the variation in each bin is more robust for systematics 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  $\sigma_E$  distributions obtained from the respective BDTs 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



**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 with red dashed lines while the interpolated  $\pm 3\sigma$  are shown with 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.

chosen to be that which covers the maximal difference in the ratio of each distribution in high  $p_T$  photons between data and MC. This is then validated using  $Z \rightarrow e^+e^-$  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 as the uncertainty in the fraction of gluon-gluon fusion and vector boson fusion events respectively which are expected to pass the dijet tag [23].

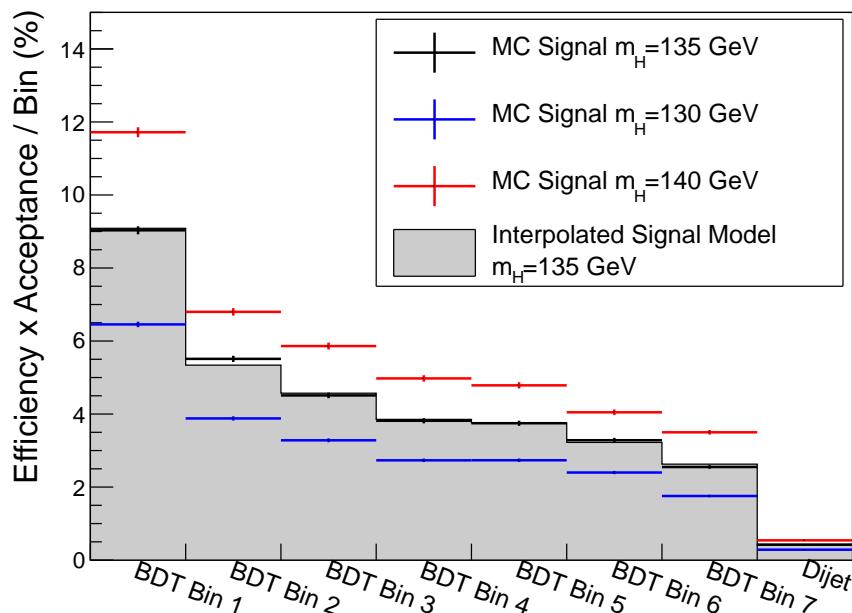
In addition to the shape systematics, theoretical errors on the SM Higgs boson cross-section are included due to uncertainties on the QCD scale and pdf variations of the various production modes [32]. A 2.2% luminosity error is also included as an uncertainty on the overall signal yield. A complete table of the systematics included in the signal model is given in Table 3.4.

### Interpolation to Intermediate Mass Points

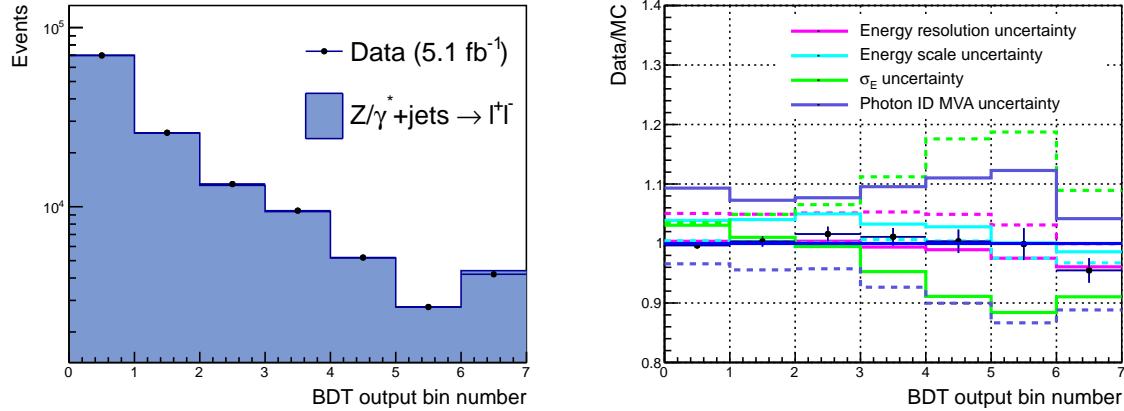
Signal Monte Carlo is available in  $m_H$  steps of 5 GeV in the range of 110 to 150 GeV. Due to the high resolution of the signal peak in the  $H \rightarrow \gamma\gamma$  channel, it is necessary to interpolate between these generated mass points in order to construct the signal model at intermediate 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.

Source of systematic uncertainty	Uncertainty	
<b>Per photon</b>	Barrel	Endcap
Photon identification efficiency	1.0%	2.6%
Energy resolution $(\Delta\sigma/E_{MC})$	$r_9 > 0.94$ (low $\eta$ , high $\eta$ ) 0.22%, 0.61% $r_9 < 0.94$ (low $\eta$ , high $\eta$ ) 0.24%, 0.59%	0.91%, 0.34% 0.30%, 0.53%
Energy scale $(E_{data} - E_{MC})/E_{MC}$	$r_9 > 0.94$ (low $\eta$ , high $\eta$ ) 0.19%, 0.71% $r_9 < 0.94$ (low $\eta$ , high $\eta$ ) 0.13%, 0.51%	0.88%, 0.19% 0.18%, 0.28%
Photon identification MVA	$\pm 0.025$ (output shift)	
Photon energy resolution MVA	10% (output scaling)	
<b>Per Event</b>		
Integrated luminosity	4.5%	
Vertex finding efficiency	$p_T^{\gamma\gamma}$ -differential	
Trigger efficiency	either photon, $r_9 < 0.94$ in endcap Other events	0.4% 0.1%
Dijet-tagging efficiency	Vector boson fusion process	10%
Dijet-tagging efficiency	Gluon-gluon fusion process	70%
<b>Production cross sections</b>	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%
<b>Scale and PDF uncertainties</b>	$p_T$ -differential	

**Table 3.4.:** 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 interpolation 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.



**Figure 3.27.:** BDT output distribution for  $Z \rightarrow e^+e^-$  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.

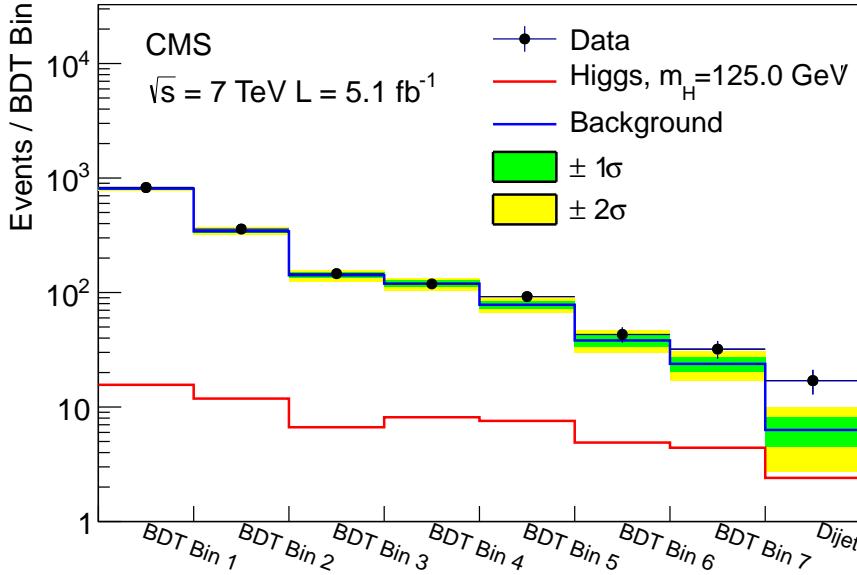
### Validation with $Z \rightarrow e^+e^-$ data

As with the other MVA discriminators in the  $H \rightarrow \gamma\gamma$  analysis, the signal model is validated by running the BDT in both  $Zee$  MC and data with the electron veto inverted. A 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  $Z \rightarrow e^+e^-$  events as for  $H \rightarrow \gamma\gamma$  events, the agreement seen between data and MC for  $Z \rightarrow e^+e^-$  events indicates that the reconstruction and kinematics of a potential signal in data will be well modelled in the signal MC.

#### 3.4.6. Statistical Interpretations of the Data

The  $H \rightarrow \gamma\gamma$  analysis was performed on the full 2011 dataset collected at CMS corresponding to 5.1  $\text{fb}^{-1}$  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  $\pm 2\%$  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.

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)



**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.

parameterises the relative compatibility of the data with the signal and background models as a function signal strength  $\mu$ , where  $\boldsymbol{\theta} = (\boldsymbol{\theta}^s, \boldsymbol{\theta}^b)$  are the nuisance parameters and  $p$  is a product of unit width Gaussian distributions centered at  $\boldsymbol{\theta}$ .

$$\mathcal{L}(\text{data}|\mu, \boldsymbol{\theta}) = p(\boldsymbol{\theta}_0|\boldsymbol{\theta}) \cdot \prod_{j=1}^8 \text{Poisson} \left( d_j | \mu \sum_p s_j^p(\boldsymbol{\theta}) + b_j(\boldsymbol{\theta}) \right) \quad (3.10)$$

The observed number of events in each bin,  $d_j$ , and expected contributions from each signal production process and background,  $s_j^p$  ( $p \in \{ggH, qqH, VH, ttH\}$ ) and  $b_j$ , correspond to one mass hypothesis although the general form is applicable to all values of  $m_H$ .

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  $\boldsymbol{\theta}^b = (\theta_N, \theta_1^b \dots \theta_7^b)$  represent the eight independent nuisances of the background

model.

$$s_j(\theta^s) = s_j^{p,mc} \cdot \prod_k \left(1 + \frac{\sigma_k^{s,p}}{s_j^{p,mc}}\right)^{\theta_k^s} \quad (3.11)$$

$$b_j(\theta^b) = N \left(1 + \frac{\sigma_N}{N}\right)^{\theta_N^b} \cdot f_j \prod_{k=1}^7 \left(1 + \frac{\sqrt{\lambda_k} V_{kj}}{f_j}\right)^{\theta_k^b} \quad (3.12)$$

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  $\theta_k^s$  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  $\lambda_k$  in Equation 3.4.6 are the eigenvectors and corresponding eigenvalues of the covariance matrix determined in Section 3.4.4. Finally,  $\sigma_N$  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,  $q_\mu$ , is constructed as the ratio of two values of the likelihood given in Equation 3.4.6,

$$q_\mu = -2 \ln \frac{\mathcal{L}(data|\mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(data|\hat{\mu}, \hat{\boldsymbol{\theta}})}, \quad (3.13)$$

where  $\hat{\mu}, \hat{\boldsymbol{\theta}}$  denote the values for  $\mu$  and  $\boldsymbol{\theta}$  at which the likelihood attains its maximum and  $\hat{\boldsymbol{\theta}}_\mu$  is the value at which the likelihood is maximal under the condition that  $\mu$  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  $\mu$  by comparing the compatibility of the data against different hypotheses for  $\mu$ . 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  $\mu$  is expressed using the  $CL_s$  procedure which is known to give conservative limits in the case of downward fluctuations

of the background [38]. This procedure involves computing two p-values (tail probabilities) under two hypothesis,  $\mu = 0$  and  $\mu \neq 0$  given by,

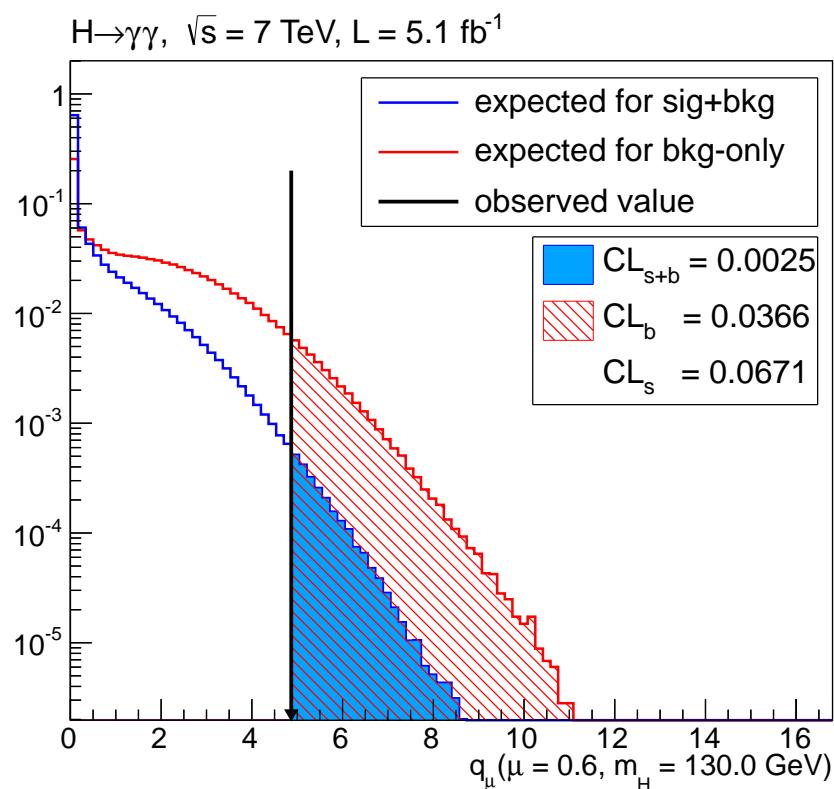
$$CL_{s+b} = \int_{q_\mu^{obs}}^{\infty} f(q_\mu | \mu, \boldsymbol{\theta} = \boldsymbol{\theta}_\mu^{obs}) dq_\mu \quad (3.14)$$

$$CL_b = \int_{q_\mu^{obs}}^{\infty} f(q_\mu | 0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs}) dq_\mu \quad (3.15)$$

where  $q_\mu^{obs}$ . The value of  $\mu$  for which the ratio  $CL_s = \frac{CL_{s+b}}{CL_b} = 0.05$  is the 95% confidence upper limit on  $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$ . When the upper limit on  $\mu$  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  $\mu$  and setting  $\mu = 0$  respectively. Pseudo data,  $d_j$ , for each bin are generated according to a Poisson distribution with expectation value  $\mu s_j(\theta_\mu^{obs}) + b_j(\theta_\mu^{obs})$ . Pseudo-measurements nuisance parameters,  $\boldsymbol{\theta}$ , are then r before evaluating the test statistic  $q_\mu$  in order to model the effect of systematic uncertainties. Examples of the normalised distributions of  $q_\mu$  for  $\mu = 0.6$  and  $\mu = 0$  are shown in Figure 3.29.

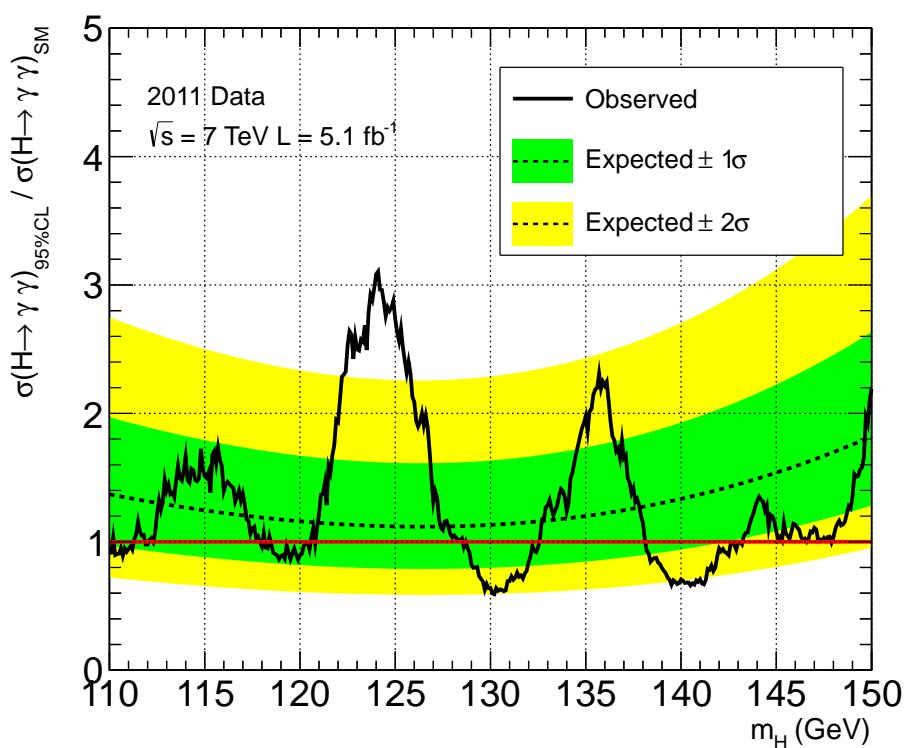
The 95% confidence upper limit on  $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$  was determined using the full 2011 dataset for different values of  $m_H$  in the range to which the channel  $H \rightarrow \gamma\gamma$  is most sensitive. Since the resolution of the signal 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 95% confidence level. The limits were calculated using an asymptotic approximation for the distribution of  $q_\mu$  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 ?? show this comparison for the median expected, 68% and 95% quantile ranges at different values of  $m_H$ .



**Figure 3.29.:** Distributions of the test statistic  $q_\mu$  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
$m_H = 120 \text{ GeV}$		
2.5%	$0.534 \pm 0.044$	0.533162
16%	$0.777 \pm 0.012$	0.778268
median	$1.175 \pm 0.020$	1.17444
84%	$1.785 \pm 0.021$	1.79479
97.5%	$2.592 \pm 0.213$	2.63468
$m_H = 130 \text{ GeV}$		
2.5%	$0.629 \pm 0.051$	0.605412
16%	$0.822 \pm 0.012$	0.797828
median	$1.149 \pm 0.019$	1.14546
84%	$1.665 \pm 0.019$	1.66279
97.5%	$2.349 \pm 0.192$	2.37195
$m_H = 140 \text{ GeV}$		
2.5%	$0.855 \pm 0.070$	0.81724
16%	$1.040 \pm 0.015$	1.00112
median	$1.361 \pm 0.022$	1.34578
84%	$1.869 \pm 0.021$	1.84936
97.5%	$2.540 \pm 0.208$	2.54582

**Table 3.5.:** Comparison of expected median upper limit and quantiles obtained using the asymptotic calculation of  $CL_s$  and toys. The error quoted in the toys column is the statistical uncertainty from only generating 1000 toys at each value of  $\mu$ . The comparison is made at three mass hypotheses in the range 120 to 140 GeV.



**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 dashed line indicates the median expected value for the upper limit on  $\mu$  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.

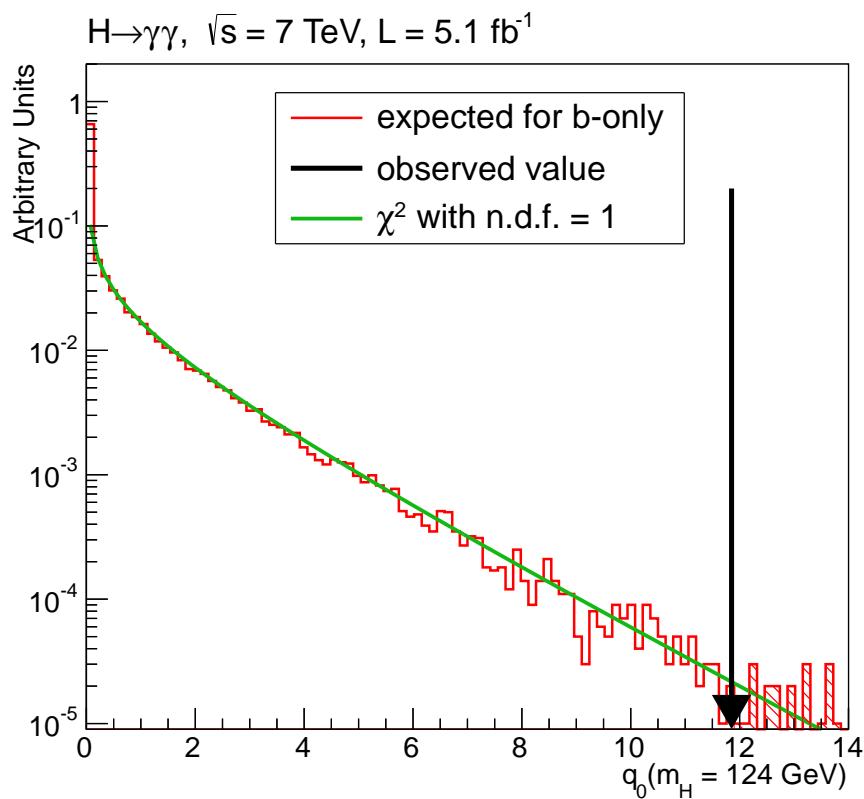
## Quatifying Excesses in the Observed Data

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.

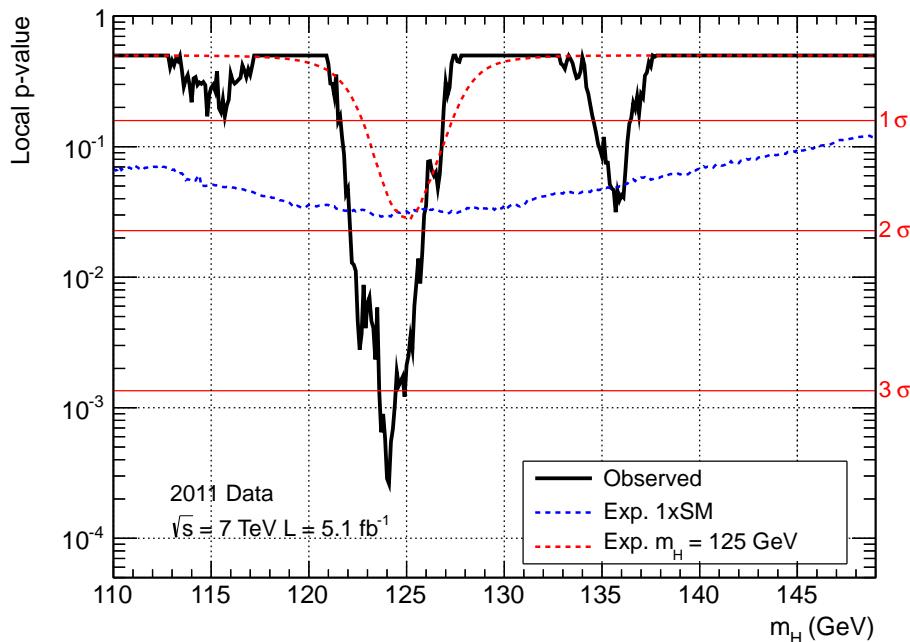
The test statistic which quantifies the relative compatibility of the data with the background only hypothesis and the presence of a signal, with any signal strength, is  $q_0$ . This is obtained by setting  $\mu = 0$  in Equation 3.13 and removing the upper bound on  $\hat{\mu}$ . Again, there is an implicit assumption that the test statistic is defined only given a particular value of  $m_H$ . The test statistic designed this way means that only excesses which are compatible in shape with that of a  $H \rightarrow \gamma\gamma$  signal at some  $m_H$  are considered significant. As the mass peak of  $H \rightarrow \gamma\gamma$  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 3.16 where  $q_0^{obs}$  is the value of the test statistic obtained in data.

$$p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs}) dq_0 \quad (3.16)$$

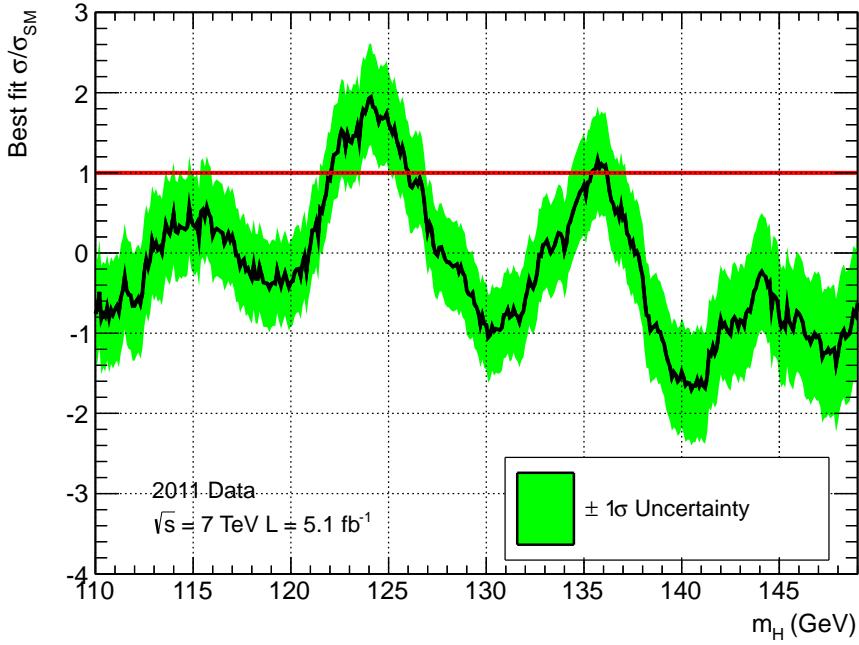
Analogous to calculating limits, the distribution  $f(q_0|0, \boldsymbol{\theta} = \boldsymbol{\theta}_0^{obs})$  can be obtained either through generating toys or using an analytic form. Figure 3.31 shows the normalised distribution of  $q_0$  under the background only hypothesis generated from pseudo-experiments compared with the analytic form, in this case a  $\chi^2$  distribution 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  $p_0 = \sqrt{q_0^{obs}}$  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  $3.4\sigma$ . 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  $\mu$  at which the likelihood attains its maximum,  $\hat{\mu}$ , as a function of  $m_H$ . The excess observed at 124 GeV corresponds to  $\hat{\mu} = 1.93_{-0.60}^{+0.67}$ , that is nearly twice the expectation from a SM Higgs.



**Figure 3.31.:** Normalised distribution of  $q_0$  at  $m_H = 124 \text{ GeV}$  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.



**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  $m_H$ .



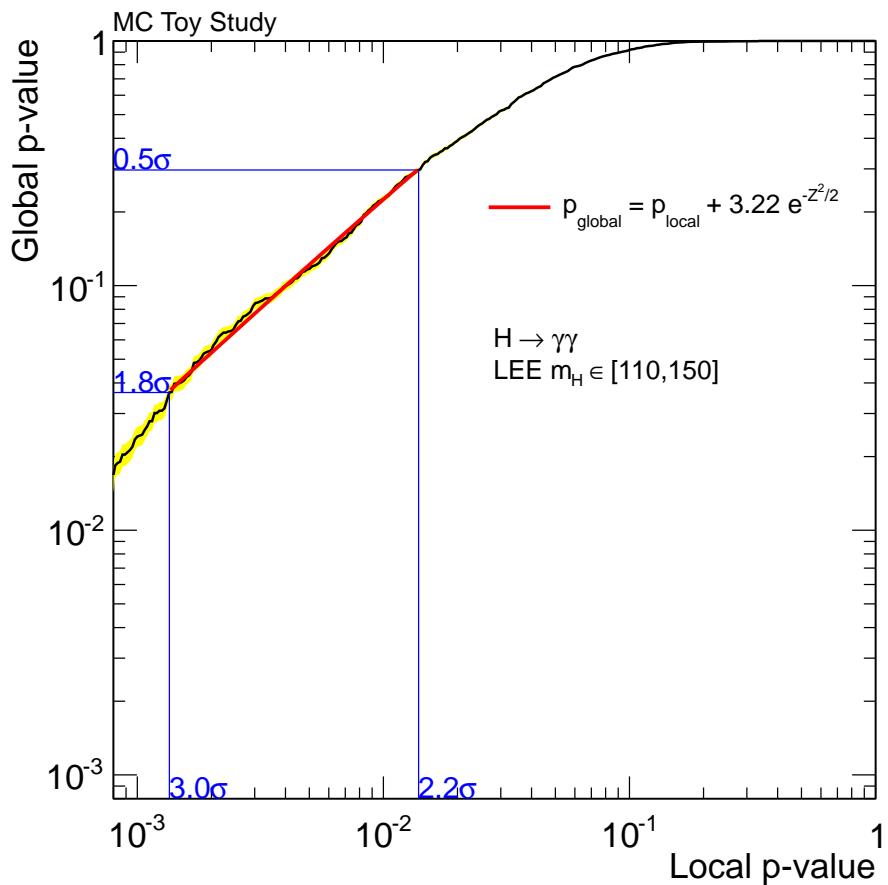
**Figure 3.33.:** Best fit for the signal strength,  $\hat{\mu}$ , in steps of 100 MeV in the range  $110 < m_H < 150$ . The green bands indicate the 68% uncertainty on  $\hat{\mu}$  for a fixed  $m_H$ . The red line at 1 represents the expectation for a SM Higgs.

### The Look-Elsewhere Effect

As the signal for the decay  $H \rightarrow \gamma\gamma$  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 [33]. 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 minimum  $p_0$  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,

$$p_{global} = p_{local} + Ce^{-\frac{Z^2}{2}}, \quad (3.17)$$

where  $Z$  is the local significance and  $C$  is a free parameter [29]. This function is then used to determine the look-elsewhere effect for larger significances. The excess observed at 124 GeV corresponds to a  $2.4\sigma$  global significance.



**Figure 3.34.:** Relationship between local and global p-values to determine the look-elsewhere effect in the  $H \rightarrow \gamma\gamma$  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.

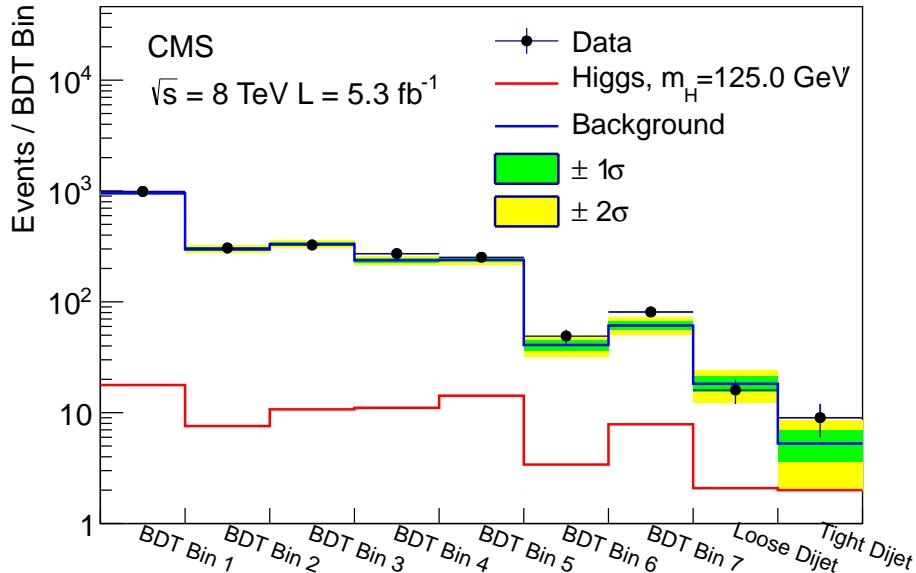
In order to generate suitable background only toys, pseudo-data are generated in two variables,  $m_{\gamma\gamma}$  and the diphoton BDT output. The value of  $m_{\gamma\gamma}$  for each event in the pseudo-data is generated from a double power law fit to the full  $m_{\gamma\gamma}$  spectrum in data in the range  $100 < m_{\gamma\gamma} < 180$  GeV. The value of the diphoton BDT is generated by fitting a kernel density estimator 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. Inclusion of 2012 Data

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  $10.4 \text{ fb}^{-1}$  [22].

#### 3.5.1. Updates for the 8 TeV Analysis

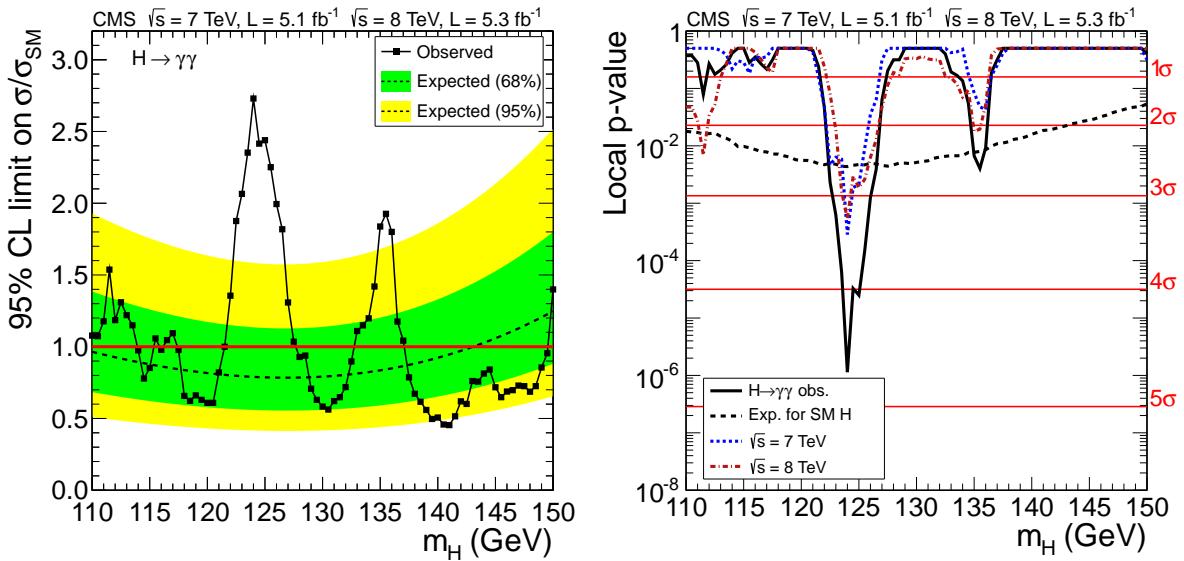
The majority of the analysis remains unmodified between the two data taking periods. Due to increased pileup conditions in the 2012 data, the regression BDTs and vertex BDTs were re-trained using MC weighted to a higher average number of pileup vertices. As a result of this, both the diphoton and event categorisation BDTs were re-trained to incorporate 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 appropriate. The invariant mass cut on the dijet system for dijet tagged events category was reduced to 250 GeV. The dijet events are then further subdivided by separating events with a large reconstructed dijet mass,  $m_{jj} > 500$  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 derived 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.



**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.

### 3.5.2. Results from the Combined Datasets

The 2011 and 2012 datasets were combined statistically by extending the likelihood in Equation 3.4.6 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 calculated as described in Section 3.4.6. Figure 3.36 (left) shows the expected and observed 95% upper limits on  $\sigma(H \rightarrow \gamma\gamma)/\sigma(H \rightarrow \gamma\gamma)_{SM}$  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  $m_H$  shown in Figure 3.36 (right). The largest excess is observed at  $m_H = 124$  GeV corresponding to a local significance of  $4.8\sigma$ . This is reduced to a global significance of  $3.9\sigma$  when considering the look-elsewhere effect in the range 110 to 150 GeV.



**Figure 3.36.:** Exclusion limits on SM Higgs production and subsequent decay to two photons (left) and Local p-value,  $p_0$  (right) in the range  $110 < m_H < 150$  GeV from the combined 2011 (7 TeV) and 2012 (8 TeV) datasets. In the left figure, the black dashed lines indicate the median expected value for the upper limit on  $\mu$  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. In the right figure, 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 black dashed line. The observed  $p_0$  from the 2011 (7 TeV) and 2012 (8 TeV) datasets individually are shown in the blue and red dashed lines respectively. The right hand scale shows the significance in standard deviations at each  $m_H$ .

# Chapter 4.

## Higgs Combinations and Properties

*For a particular hypothesis of the SM Higgs boson mass, the sensitivity of the search depends not only on the production cross-section and branching ratio to a particular decay channel, but also on the efficiency of the selection, the experimental resolution and the relative proportions of signal to SM background processes. By combining results from searches in many decay channels, the overall sensitivity to a SM Higgs is greatly improved. This chapter describes the combined search for the SM Higgs In Section 4.1, a short review on the statistical treatment of data from different analyses used at CMS is provided and a set of diagnostic tools developed by the author are discussed. The results of the search using the International Conference on High Energy Physics (ICHEP) dataset are included which led to the announcement of the discovery of a new particle by Atlas and CMS in June 2012 [22]. Section 4.2 deals with early studies of the properties of the newly discovered particle presented at the Hadron Collider Physics (HCP) symposium in November 2012. This includes a discussion of the Feldman-Cousins technique which was in part implemented and performed by the author to extract information on the compatibility of the new state with the SM Higgs boson.*

### 4.1. Combined Higgs Searches

In High Energy physics, it is often desireable to search for new physics which has been motivated by some theoretical considerations. Where the theory provides well defined predictions, experimental data can be used to verify or reject the theory by means of Hypothesis testing. The goal is to use the data to reject one of two hyptotheses,  $H_0$  and  $H_1$ , known as the null and alternate hypotheses respectively. A function is defined,

$t(data)$ , which characterises the observed data as a single value. The critical region,  $w$ , is defined as the set of possible values of  $t$  which indicate that  $H_0$  is not true. The probability then to observe  $t \in w$  when  $H_0$  is true ( $\alpha$ ),

$$\alpha = P(t \in w | H_0) \quad (4.1)$$

is the probability that  $H_0$  would be rejected even if it was true. This is also known as the significance of a test. The strength of a test (referred to as its power) is quantified by the probability,  $1 - \beta$  that  $t \in w$  when  $H_1$  is true.

$$1 - \beta = P(t \in w | H_1) \quad (4.2)$$

In the case of the search of the SM Higgs boson, the two hypotheses can be parameterised in terms of a production cross-section relative to that predicted by the SM,  $\sigma/\sigma_{SM}$ . The null hypothesis ( $H_0$ ) is then that under which no SM Higgs boson exists,  $\sigma/\sigma_{SM} = 0$ , where the alternate ( $H_1$ ) is characterized by  $\sigma/\sigma_{SM} = 1$ . In this case,  $H_0$  is referred to as the background only hypothesis and  $H_1$  the signal plus background hypothesis. The possible outcomes values of  $t$  are assumed to be random with a probability density function  $f(t|\sigma/\sigma_{SM})$ . The values of  $\alpha$  and  $1 - \beta$  are then obtained by;

$$1 - \beta = \int_w f(t|\sigma/\sigma_{SM} = 1) dt \quad (4.3)$$

$$\alpha = \int_w f(t|\sigma/\sigma_{SM} = 0) dt, \quad (4.4)$$

It can be shown that the choice of  $w$  which maximises the power of the test for a given level of significance are the set of points for which,

$$q = \frac{f(t|\sigma/\sigma_{SM} = 1)}{f(t|\sigma/\sigma_{SM} = 0)} \geq c_\alpha, \quad (4.5)$$

where  $c_\alpha$  is chosen such that Equation 4.4 holds [35]. The quantity  $q$ , known as the “test-statistic”, is formed of the ratio of the likelihoods under the two hypotheses.

In the search for the SM Higgs boson, the compatibility of the data with the presence of a Higgs boson are interpreted in terms of the continuous parameter  $\mu$  known as the signal strength relative to that expected from the Standard Model. Again, the null hypothesis,  $H_0$ , is characterized by setting  $\mu = 0$  however, an infinite number of alternate hypotheses exist for any value  $\mu \geq 0$ . This is known as a nested set of hypotheses in the sense that the null is contained in the set of alternate hypotheses [35].

In order to combine data from all decay channels relevant in the search for the SM Higgs, the likelihood for a particular outcome of the data given a particular value of  $\mu$  is the product of the individual likelihoods in each channel  $i$ ,

$$\mathcal{L}(\text{data}|\mu, \boldsymbol{\theta}) = \prod_i \mathcal{L}_i(\text{data}_i|\mu, \boldsymbol{\theta}) = \prod_i P(\text{data}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}_0|\boldsymbol{\theta}). \quad (4.6)$$

, where the data are either the real observed data or pseudo-data. The relative signal strength  $\mu$  is a single parameter which scales the signal yield in all sub-channels simultaneously. Each value of  $\mu$  represents an hypothesis which can be rejected using the data at some confidence level  $c \in (0, 1)$  in favour of the background only hypothesis. Systematic uncertainties in the signal and background models in each channel are modelled through the nuisance parameters,  $\boldsymbol{\theta}$ . Typically these nuisances will be constrained by some external measurements  $\theta_0$ , such as the energy scale measured in  $Z \rightarrow e^+e^-$  events in the two-photon decay channel described in Section 3.2.1. The probability density function  $p(\theta_0|\boldsymbol{\theta})$  is a product of each of the nuisance pdfs, usually Gaussian distributions, from each independent source of systematic uncertainty. Although each event observed in data are exclusive to a particular each channel, many sources of systematic uncertainty are common to several analyses. For this reason, the nuisance parameters,  $\boldsymbol{\theta}$  are correlated between sub-channels. The test-statistic for a given  $\mu$  is defined as the ratio of “profiled” likelihoods  $q_\mu$  in Equation 4.7.

$$q_\mu = \begin{cases} -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\boldsymbol{\theta}})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} < 0 \end{cases} \quad (4.7)$$

An immediate consequence of this definition is that the value attained by the test statistic is always positive. Small values of the test statistic indicate outcomes which are in favour of the signal plus background hypothesis, where large values indicate outcomes which disfavour it. Due to this, the critical region  $w$  can always be defined as the right hand tail of the normalized distribution of the test-statistic  $f(q_\mu|\mu)$ ,

$$w = \{q_\mu : q_{mu} \in (c_\alpha, +\infty)\}, \quad (4.8)$$

Commonly the integral of  $f(q_\mu|\mu)$  above the observed value of the test-statistic in data ( $q_\mu^{obs}$ ), known as a  $p$ -value, is calculated to provide a measure of how well the data disfavours a value of  $\mu$ .

For each specific value of  $\mu$ , the  $p$ -value  $CL_{s+b}$  (Equation ??) is calculated. The largest value of  $\mu$  for which  $CL_{s+b} < \alpha$ , at a fixed value of  $\alpha$  is quoted as the upper limit on  $\mu$  with confidence level  $1 - \alpha$ . An additional constraint of  $\hat{\mu} \leq \mu$  is imposed in Equation 4.7 when calculating upper limits which forces the limit on  $\mu$  to be one-sided. At CMS, the upper limit on  $\mu$  is determined using the  $CL_s$  procedure described in Section 3.4.6, which is designed to provide weaker exclusion limits in analyses which are less sensitive to signal [37].

In the presence of a sizeable excess in data, the background only hypothesis can be rejected in favour of an SM like one. Specifically, the excess will be compatible with the presence of a SM Higgs excepting in the rate at which it is produced. In order to quantify the excess, the test-statistic is replaced with  $q_0$  by setting  $\mu = 0$  in Equation ???. The background only hypothesis is rejected in favour of a signal plus background one when the  $p$ -value  $p_0$ , given in Equation ??, is less than some pre-determined critical level  $\alpha$ . Since  $p_0$  is uniformly distributed between 0 and 1 under the background hypothesis,  $p_0$  is exactly the probability  $\alpha$  of falsely rejecting the background only hypothesis. The critical value for  $\alpha$  is typically  $2.87 \times 10^{-7}$  (corresponding to a significance of  $5\sigma$ ) when searching for new physics. The procedures used at CMS for determining the distributions,  $f(q_0)$  and  $f(q_\mu)$  to calculate  $p$ -values was discussed in Section 3.4.6.

The likelihood is coded using the C++ based statistical packages `RooFit/RooStats` version 5.3.0. A framework for automating the procedure of combining datasets, generating toys and evaluating likelihoods in the context of the combined search for the SM Higgs boson was developed within `CMSSW` under the package `HiggsAnalysis/CombinedLimit`. All of the results shown in the following sections are obtained with this package.

#### 4.1.1. Diagnostics with Toy Datasets

Frequentest statistical techniques often involve generating many pesudo-datasets (toys) in order to build the distribution of a test-statistic. As described in Chapter 3, these distributions are used to set confidance intervals or determine the significance of some observed excess in experimental data. The combined Higgs searches at CMS employs the profiled likelihood test-statistic (Equation 3.13) in which the nuisance parameters,  $\theta$ , are profiled (fit) from the data. For calculating the significance of an excess in data, the distribution of the test-statistic  $q_0$  under the background only hypothesis is required. The procedure for determining this distribution proceeds as follows;

- Fit the observed data fixing  $\mu = 0$ . The values of the nuisance parameters at which the likelihood attains its maximum are denoted  $\boldsymbol{\theta}_{obs}$  and will represent the expectation value of the nuisance parameters in the likelihood.
- Generate a toy dataset under the background only hypothesis. For the purposes of generating data, the nuisance parameters are fixed to  $\boldsymbol{\theta} = \boldsymbol{\theta}_{obs}$ .
- Fit the toy dataset twice, once fixing  $\mu = 0$ ,  $\mathcal{L}(data|0, \hat{\boldsymbol{\theta}}_0)$  and once more letting  $\mu$  float freely,  $\mathcal{L}(data|\hat{\mu}, \hat{\boldsymbol{\theta}})$ . When evaluating the likelihood, the values  $\boldsymbol{\theta}_{obs}$  are re-generated in order to model the systematic uncertainties.

A realistic example of a search for an hypothesised particle,  $H$ , decaying to two  $\tau$  leptons was produced in the form of a simple counting experiment. The search is performed as a combination of three channels arising from the possible subsequent decays of the two  $\tau$  leptons;  $\tau_h - e$ ,  $\tau_h - \mu$  and  $e - \mu$ . In each channel the expected background is estimated either from simulation or some control region in data. The observed data in each channel is represented by a simple count of events falling in that category. Several sources of systematic uncertainty are included which effect the expected signal and background in one or more of the channels. Systematics are incorporated into the likelihood in the form of nuisance parameters as described previously. The analysis is summarized in Table 4.1 which details the number of expected events from each background and signal process in each channel as well as the observed count in data. The analysis is available from the CMSSW package `HiggsAnalysis/CombinedLimit` under the directory `data/tutorials/realistic-counting-experiment.txt` [35].

Around 90,000 toy datasets were generated under the background only hypothesis as is appropriate to determine the distribution of  $q_0$  (setting the value of  $\mu$  of Equation 3.13 to zero). Each toy dataset was fit twice, once fixing the signal strength,  $\mu$ , to zero and a second allowing  $\mu$  to float freely. The results of the fits can be used to diagnose the fits and highlight potentially problematic channels or nuisance parameters. Figure 4.1 shows a summary of the fit results in the nuisance parameter `lumi`, which models the systematic associated to uncertainty of the total luminosity of the dataset. The upper left panel shows two pull distributions of the values from the fit defined as the difference between the value of the fitted parameter and the value from the best fit to data,  $\theta_{obs}$ , divided by the  $1\sigma$  uncertainty on the parameter before fitting to the data. The blue histogram includes all toys while the red shows the results only for which the best fit signal strength is positive. Since the test-statistic  $q_0$  is designed to report only excesses

channel	$\tau_h - e$			$\tau_h - \mu$			$e - \mu$		
observed	517			540			101		
expected	Sig	$Z \rightarrow \tau\tau$	QCD	Sig	$Z \rightarrow \tau\tau$	QCD	Sig	$Z \rightarrow \tau\tau$	other
	0.34	190	327	0.57	329	259	0.15	88	14
systematics									
lumi	11%	-	-	11%	-	-	11%	-	11%
tauid	23%	23%	-	23%	23%	-	-	-	-
ZtoLL	-	4%	-	-	4%	-	-	4%	-
effic	4%	4%	-	4%	4%	-	4%	4%	4%
QCDel	-	-	20%	-	-	-	-	-	-
QCDmu	-	-	-	-	-	10%	-	-	-
other	-	-	-	-	-	-	-	-	10%

**Table 4.1.:** A realistic counting experiment across several channels. The number of observed events and that expected from signal and background processes are given per channel. Several sources of systematic are included which effect the expected rate of each signal or background process. Where a dash is entered, the systematic uncertainty has no effect on that process or channel.

in the data, it is important to check that nuisance parameters correlated to the signal strength are well behaved.

The pull distributions are fitted with a Gaussian and the width and mean are reported in the upper right panel. Since the constrain terms in the likelihood for nuisances are generated around the best fit to data, the pulls are expected to be centered around 0. In general, nuisance parameters are constrained from external measurements so it is expected that the width of the pull distribution is 1. Nuisance parameters which are further constrained by the observed data will typically have a pull distribution with width less than unity. The parameter `lumi` does not show signs of being constrained by the data. This is reflected in the lower left panel which shows the correlation between the nusiance parameter and the value generated for the expectation of this parameter (`lumi_In`) in each toy. This behaviour is expected since this nuisance parameter mostly effects the signal process and is correlated across all channels so that only the overall normalization is altered. Since these fits allow  $\mu$  to float freely, any parameter which alters only the overall normalization of the signal should achieve its expectation value at the maximum of the likelihood. The lower right panel shows the shape of the negative

log-likelihood as a function of the nuisance parameter ( $\theta$ ),

$$-\log \frac{\mathcal{L}(\text{data}|\mu = \hat{\mu}, \theta = \theta_{S+B})}{\mathcal{L}(\text{data}|\mu = \hat{\mu}, \theta)}, \quad (4.9)$$

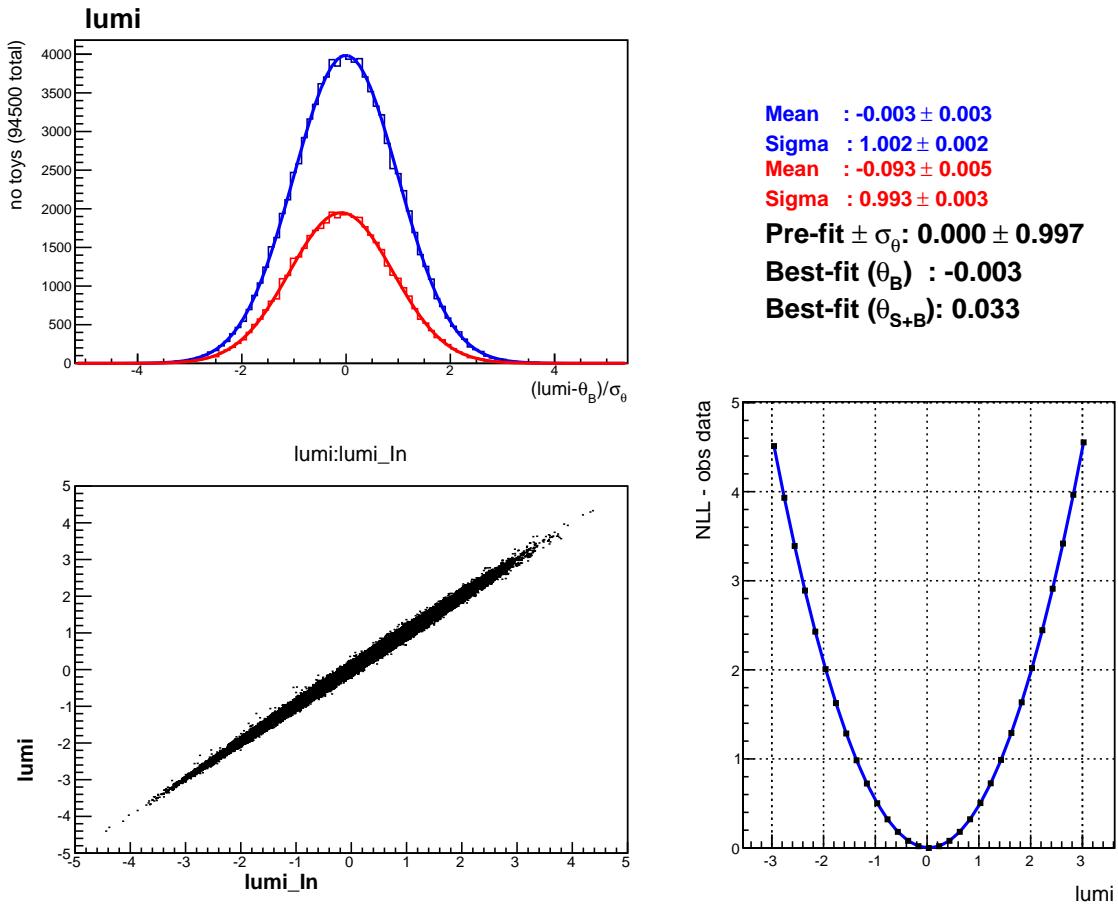
near its minimum value ( $\theta_{S+B}$ ). At each point, all other parameters are fixed to those of the best fit to the data (in this case, from the fit allowing  $\mu$  to float). The likelihood is expected to be parabolic around its minima with no secondary (local) minima present. Degenerate minima, which can cause instabilities in the fitting procedure, will be visible in the shape of the negative log-likelihood. The diagnostic tools described here were applied to the ICHEP 2012 combination as documented in a CMS Analysis note by the author [?]. The full set of diagnostic summary plots can be found here [?].

#### 4.1.2. Higgs Search Combination

A search for the SM Higgs boson was performed by combining data recorded at CMS at a centre of mass energy of 7 and 8 TeV. The search was performed in five decay modes,  $H \rightarrow \gamma\gamma$ ,  $H \rightarrow ZZ$ ,  $H \rightarrow WW$ ,  $H \rightarrow \tau\tau$  and  $H \rightarrow bb$  with datasets of integrated luminosities of  $4.9 - 5.1 \text{ fb}^{-1}$  and  $5.1 - 5.3 \text{ fb}^{-1}$  from the 2011 and 2012 data taking periods of the LHC respectively. The search is performed across a wide range in Higgs mass hypothesis ( $m_H$ ) from 110 to 600 GeV. For  $m_H > 150$ , the  $H \rightarrow \gamma\gamma$ ,  $H \rightarrow \tau\tau$  and  $H \rightarrow bb$  decay modes are not used as their expected sensitivities are no longer sensitive compared to the  $H \rightarrow WW$  and  $H \rightarrow ZZ$  channels. The exclusive final state decays from each of the five modes used in the combination, including the size of the dataset used and the mass range to which they are sensitive, are given in Table 4.2.

#### Combined Search Channels

The  $H \rightarrow \gamma\gamma$  analysis is one of the most sensitive channels at low  $m_H$ . The analysis is that described in Chapter 3 with the exception of the signal extraction. Events are categorized using the diphoton BDT into four classes chosen so as to optimise the search in terms of the expected limit at  $m_H = 125 \text{ GeV}$ . The diphoton invariant mass spectrum in each category is fit with polynomial functions whose order is determined following a procedure designed to reduce any potential bias to less than 20% of the statistical uncertainty on the background [?]. The dijet selected events are categorised separately and treated in the same way as the inclusive events.



**Figure 4.1.:** Summary plots for the parameter `lumi` of the realistic counting experiment. The entries in the histograms are for fits to toys generated under the background only hypothesis letting  $\mu$  float freely. The bottom, left panel shows the correlation between the value generated for the expectation value of the nuisance `lumi_In` and the fitted value of the parameter. The lower right panel shows the shape of the negative log-likelihood (NLL) as a function of the nuisance parameter. The the parameters of the fitted Gaussian for each histogram is given as Mean and Sigma. The value and error of the nuisance are given before fitting to the data (Pre-fit), followed by the best fit value of the parameter under the background only and signal plus background hypotheses.

Due to the extremely high cross-section of  $b\bar{b}$  production in p-p collisions, the  $H \rightarrow bb$  search focuses on Higgs boson production in association with a  $W$  or  $Z$ , which are identified by the presence of leptons or, in the case of neutrino final states such as  $Z \rightarrow \nu\nu$ , by requiring large  $/E_T$  in the event [5]. The Higgs candidate itself is reconstructed by looking for two  $b$ -tagged jets indicated by their production at secondary vertices. Events are categorized into those where the  $W$  or  $Z$  boson is recoiling away from the  $b\bar{b}$  system with high momentum. The main backgrounds are from  $W/Z+jets$  and  $t\bar{t}$  as well as from  $WZ$  and  $ZZ$  in which the  $Z$  decays to a pair of  $b$ -quarks. The backgrounds are suppressed by use of a multivariate analysis technique trained on MC simulation. The search is also performed in events in which the Higgs is produced in association with a pair of top-quarks ( $t\bar{t}H$ ) categorized into either lepton plus jet or dilepton final states [5]. This mode was not included in the 8 TeV dataset for the ICHEP 2012 combination.

In the  $H \rightarrow \tau\tau$  decay channel, the search is performed using events with leptonic final states and events in which one of the tau-leptons decays hadronically ( $\tau_h$ ) [6]. Events are divided into categories based on the number and type of jets in the event and by transverse momentum of the visible part of the tau decay. A signal in this channel will be visible as a broad excess in the invariant mass of the  $\tau\tau$  system ( $m_{\tau\tau}$ ). The main backgrounds are from  $Z \rightarrow \tau\tau$  events and  $W+jet$  production. The  $VH$  modes are exploited by selecting events which have one or more additional leptons consistent with a  $W$  or  $Z$  boson decay.

The  $H \rightarrow WW$  analysis is one of the most sensitive analyses at CMS for values of  $m_H$  between 150 and 200 GeV [7]. The  $WW \rightarrow 2l2\nu$  sub-channel consists of events with two opposite charged leptons, a large  $/E_T$  and up to two jets (to target  $qqH$  production). These are sub-divided into categories in which the two leptons are the same flavour and opposite flavour to exploit the different background contributions from  $Z$  decays. For the 7TeV analysis, an MVA classifier was trained on signal and background MC to separate signal from background. The search is conducted by looking for an excess of events in the output distribution of the MVA. In the  $WW \rightarrow l\nu2q$  sub-channel, a broad excess is searched for in the four-body invariant mass [? ]. The invariant mass is reconstructed from the lepton four-vector and  $/E_T$  assuming the mass of the  $l\nu$  is that of a  $W$  boson and choosing the neutrino's longitudinal component to be that which minimises the transverse momentum of the  $l\nu$  system. Associated production of the Higgs with a  $W$  boson is searched for by looking for an excess of events with three leptons and large  $/E_T$  [8].

The  $H \rightarrow ZZ$  analysis focuses on four final state topologies. The  $ZZ \rightarrow 4l$  is a search for a narrow four-lepton invariant mass peak over a small background [9]. The kinematics of the  $4l$  system are used to assign a probability that the event is from either a signal or background process to improve the sensitivity. For the lower mass region ( $m_H < 180$ ), only one of the lepton pairs is required to have a mass consistent with an on-shell  $Z$  boson. The  $4e$ ,  $4\mu$  and  $2e2\mu$  sub-channels are categorised separately as the mass resolutions and the background rates differ between the three final states. In the  $ZZ \rightarrow 2l2\tau$  and  $ZZ \rightarrow 2l2q$ , a broader peak is searched for in the dilepton-ditau and dilepton-dijet mass respectively [9, 10]. The limited resolution in jet energy reconstruction and the effect of the neutrino escaping detection in leptonic tau decays reduces the mass resolution in these channels compared to the  $4l$  decay. The  $ZZ \rightarrow 2l2\nu$  search looks for a leptonic  $Z$  decay and a large  $/E_T$  [11]. A broad excess of events in the  $m_T$  distribution is used to signal the presence of a SM Higg boson.

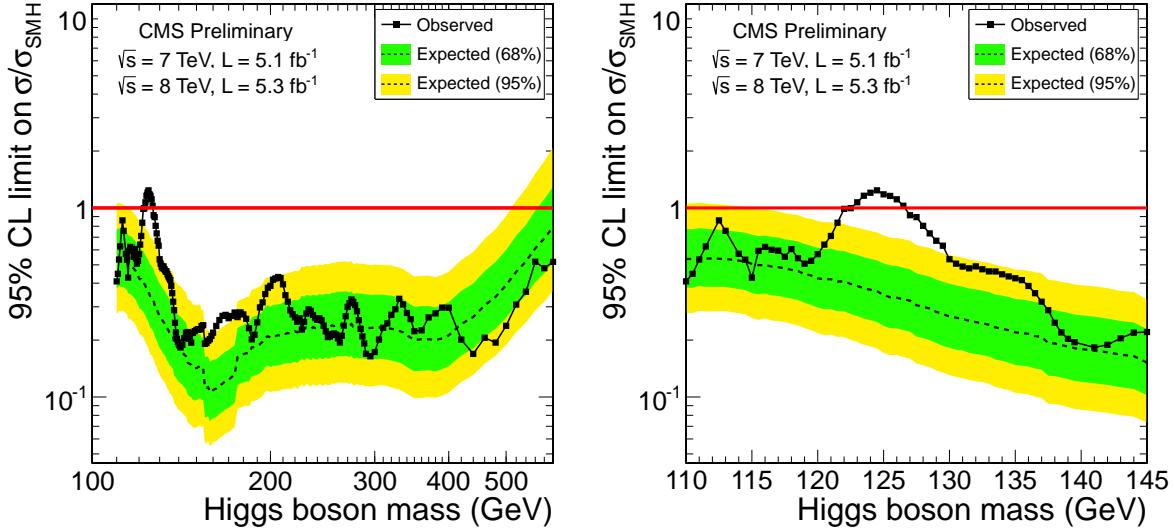
## Combined Results

The 95% upper limits on the signal strength  $\mu = \sigma/\sigma_{SM}$  as a function of the hypothesised Higgs boson mass,  $m_H$ , are shown in Figure 4.2. The right hand figure is an enlargement of the region  $110 < m_H < 145 GeV$ . The median expected limit in the absence of a SM Higgs boson is less than 1 for the range  $110 < m_H < 600 GeV$ . The observed limits are consistent with statistical fluctuations given the size of the dataset in most of the range as indicated by the fact that the observed line lies within the 68% or 95% quantiles. However an excess of events is observed at low mass in the range  $122.5 < m_H < 127 GeV$  so that exclusion of a SM Higgs boson with a mass in that range cannot be excluded at the 95% confidence level. The significance of the excess is quantified as a function of  $m_H$  by calculating the local  $p$ -value,  $p_0$  as shown n Figure 4.3. For the overall combination, the local  $p_0$  is around  $5.5 \times 10^{-7}$ , equivalent to a significance of  $4.9\sigma$ . The test indicates that the observed excess is incompatible with the background only hypothesis indicating the presence of a new state with a mass near 125 GeV. The largest contributions to the excess are from the  $H \rightarrow \gamma\gamma$  and  $H \rightarrow ZZ \rightarrow 4l$  channels, both of which have good mass resolutions localisation of the excess. The combination of the two high mass resolution channels results in a local significance of  $5\sigma$ . Of the lower resolution channels, only  $H \rightarrow WW$  shows an excess at 125 GeV. The inclusion of the  $H \rightarrow bb$  and  $H \rightarrow \tau\tau$  channels reduced the overall significance. The overall global  $p$ -value in the range 115-130 GeV is calculated by generating 10,000 pseudo-datasets and fitting for the constant  $C$  in the relationship to the local  $p$ -value given in Equation 3.17. The look-elsewhere effect,

$H$ decay	$H$ prod	Final state	No. sub-chans	$m_H$ (GeV)	Lumi ( $fb^{-1}$ ) 7/8TeV
$\gamma\gamma$	untagged $qqH$ -tag	$\gamma\gamma$ (kinematic classes) $\gamma\gamma + jj$ ( $m_{jj}$ classes in 8TeV)	4 1 or 2	110-150 110-150	5.1/5.3 5.1/5.3
$bb$	$VH$ -tag $ttH$ -tag	$(\nu\nu, ee, \mu\mu, e\nu, \mu\nu + 2j_b) \times (\text{low/high } p_T^V)$ $l + (4, 5, \geq 6j) \times (3, \geq 4j_b), l + 4j + 2j_b,$ $ll + (2, \geq 3j_b)$	10 9	110-135 110-140	5.0/5.1 5.0/-
$\tau\tau$	0/1 - jets $qqH$ -tag $ZH$ -tag $WH$ -tag	$(e\tau_h, \mu\tau_h, e\mu, \mu\mu) \times (\text{low/high } p_T^{\tau\tau}) \times (0/1j)$ $(e\tau_h, \mu\tau_h, e\mu, \mu\mu) + jj$ $(ee, \mu\mu) \times (\tau_h\tau_h, \mu\tau_h, \mu\tau_h, e\mu)$ $ee\tau_h, \mu\mu\tau_h, e\mu\tau_h$	16 4 8 3	110-145 110-145 110-160 110-140	4.9/5.1 4.9/5.1 5.0/- 4.9/-
$WW \rightarrow ll\nu\nu$	0/1 - jets $qqH$ -tag $WH$ -tag $VH$ -tag	$(ee/\mu\mu, e\mu) \times (0/1j)$ $(l\nu l\nu + jj) \text{ (SF or OF } ll \text{ in 8TeV)}$ $3l3\nu$ $(l\nu l\nu + jj) \text{ (SF or OF } ll \text{)}$	4 1 or 2 1 2	110-600 110-600 110-200 118-190	4.9/5.1 4.9/5.1 4.9/- 4.9/-
$WW \rightarrow llqq$	untagged	$(e\mu) \times (jj + 0/1j)$	4	170-600	5.0/5.1
$ZZ \rightarrow 4l$	untagged	$4e, 4\mu, 2e2\mu$	3	110-600	5.0/5.3
$ZZ \rightarrow 2l2\tau$	untagged	$(ee, \mu\mu) \times (\tau_h\tau_h, e\tau_h, \mu, \tau_h, e\mu)$	8	200-600	5.0/5.3
$ZZ \rightarrow 2l2q$	untagged	$(ee, \mu\mu) + jj(0, 1, 2j_b)$	6	200-600	4.9/-
$ZZ \rightarrow 2l2\nu$	untagged $qqH$ -tag	$(ee, \mu\mu) + /E_T(0, 1, 2j) \text{ (not } qqH \text{ jets)}$ $(ee, \mu\mu) + /E_T + jj$	6 2	200-600 200-600	4.9/5.1 4.9/5.1

**Table 4.2:** Summary of analyses included in the ICHEP 2012 combination [12]. The column for  $H$  prod indicates the production process targetted by the sub-channel. A label “untagged” indicates that the main contribution is from the  $ggH$  production process. The final states for each channel are exclusive (no events lie in more than one sub-channel). The notations used here are:  $jj$  indicating a dijet pair whether from a  $W, Z$  decay or being consistent the vector-boson fusion process;  $j_h$  denotes a jet which is identified as a  $b$ -jet;  $l$  is either a muon ( $\mu$ ) or electron ( $e$ ); OF and SF are dilepton pairs with opposite flavour ( $e\mu$ ) and same flavour ( $ee$  or  $\mu\mu$ ) respectively.

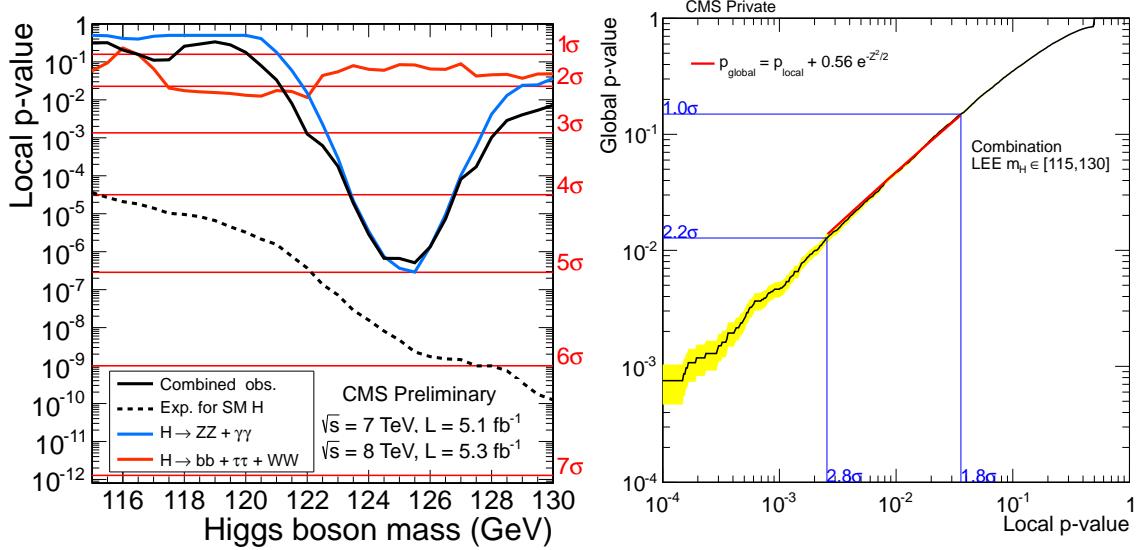
calculated as the ratio between the local and global  $p_0$ , is around 11 such that the global significance remains high at  $4.4\sigma$ .



**Figure 4.2.:** Combined 95% upper limits on the production cross-section of Higgs boson production relative to that of the Standard Model in the  $m_H$  ranges 110–600 GeV (left) and 110–145 GeV (right). The median upper limits expected in the absence of a SM Higgs is indicated by the dashed black line and the 68% and 95% quantiles by the green and yellow bands respectively. The observed upper limits from the combined ICHEP 2012 dataset is shown by the black solid line. Where the observed limit is lower than 1 (red line) A SM Higgs boson with that  $m_H$  is excluded at the 95% confidence level.

## 4.2. Higgs Properties

With the announcement of the discovery of a new state near 125 GeV, attention at CMS and ATLAS turned to the characterization of the particle through measurements of its properties. In particular, emphasis is placed on ascertaining the compatibility of the new state with the SM Higgs boson. This section includes discussions of some of the techniques used at CMS to determine the properties of the newly discovered state and results presented at the HCP symposium in November 2012. All of the analyses described in Section ??, with exception of  $H \rightarrow \gamma\gamma$ , were updated to improve their sensitivity and include the additional data collected at CMS. The total integrated luminosity of the data sample used is up to  $12.2 fb^{-1}$  depending on the specific channel.



**Figure 4.3.:** Left: The observed local  $p$ -value,  $p_0$  for sub-combinations of the low and high resolution channels and the overall combination as a function of  $m_H$ . The dashed line shows the expected  $p_0$  at each  $m_H$  should a SM Higgs boson exist with mass  $m_H$ . Right: Relationship between the local and global  $p_0$  in the range 115–130 GeV. The red line indicates the analytic expression given fit to the relationship derived from 10,000 pseudo-datasets.

#### 4.2.1. Extracting Signal Parameters

The best fit value for the signal strength is evaluated by scanning for the value of  $\mu$  at which the likelihood (Equation ??) attains its maximum in data. This can be extended where more than one signal parameter is of interest by generalising to the profiled likelihood ratio,

$$q_{\boldsymbol{\mu}} = -2 \ln \frac{\mathcal{L}(\text{data}|\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})}{\mathcal{L}(\text{data}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}, \quad (4.10)$$

where  $\boldsymbol{\mu} = \mu_1, \mu_2, \dots, \mu_N$  represents the  $N$  parameters of interest in the signal model. The values of the nuisance parameters  $\boldsymbol{\theta}$ , which maximise the value of  $\mathcal{L}(\text{data}|\boldsymbol{\mu}, \boldsymbol{\theta})$  first fixing the values of  $\boldsymbol{\mu}$  and then letting them float freely are denoted  $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\theta}}$  respectively. The values for which  $q_{\boldsymbol{\mu}} = 0$  in the observed data are the best fit values. The contour defined by the set of points for which  $q_{\boldsymbol{\mu}} = X$  is interpreted as the 68% confidence interval for the parameter set [35]. In one dimension, this represents the usual 68% confidence interval.

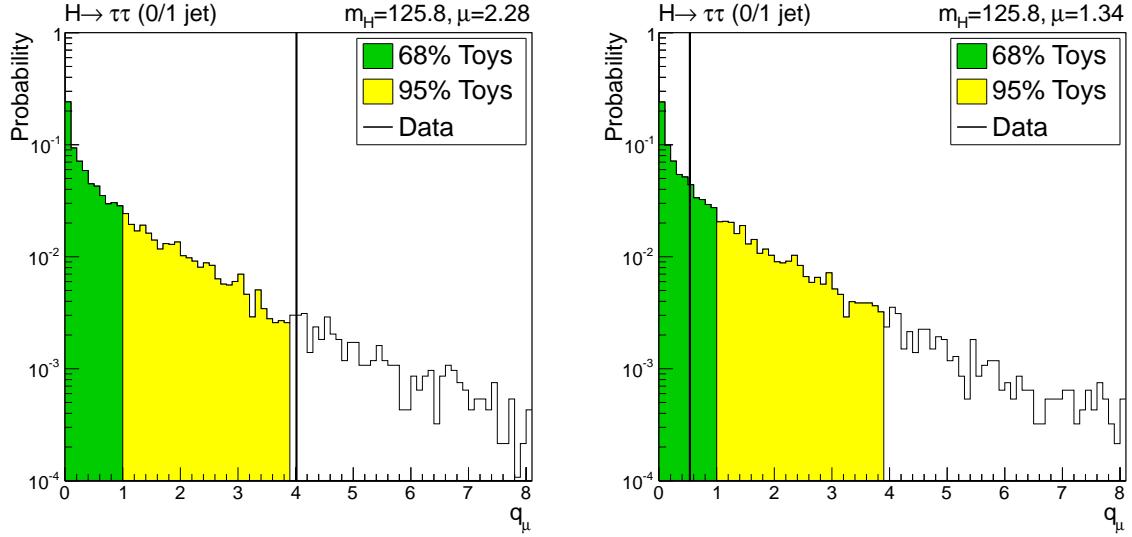
### The Feldman-Cousins Procedure for Evaluating Confidence Intervals

For parameters such as the relative production cross-section,  $\mu$ , negative values are not considered physical. Constraints on the fit can be imposed to avoid quoting unphysical values. However, where the best-fit values for the signal model parameters lie outside physically allowed regions, the relationship between the values of  $q_\mu$  and the 68% confidence interval no longer holds. In order to assign confidence intervals which provide the correct coverage, the Feldman-Cousins procedure is used [28]. The procedure involves throwing pseudo-datasets and evaluating an appropriate test-statistic to evaluate the compatibility of the data with each point in the  $N$ -dimensional parameter space. The test-statistic used in the one-dimensional case of the signal strength is defined using the ratio of profiled likelihoods,  $q_\mu$  (Equation ??). The physical constraint on the parameter is imposed in this case by requiring that  $\hat{\mu} \geq 0$ . The probability to obtain a value of the test-statistic larger than the one observed in data ( $CL_{s+b}$ ) is calculated as in Equation 3.14, where the distribution  $f(q_\mu|\mu, \boldsymbol{\theta} = \boldsymbol{\theta}_\mu^{obs})$  is generated from evaluating the test-statistic in pseudo-datasets generated from the expected background and signal models. As with calculating upper limits, for generating the pseudo-data, the nuisances ( $\boldsymbol{\theta}$ ) as set to the values obtained from a fit to the data. Figure 4.4 shows an example of this distribution for two values of  $\mu$  from the (0/1)-jet bin of the  $H \rightarrow \tau\tau$  analysis and the values of  $q_\mu^{obs}$  obtained from the observed data. The 68% confidence interval for  $\mu$  is determined as the union of all values of  $\mu$  for which  $1 - CL_{s+b} < 0.68$ . Figure 4.5 shows the values of  $1 - CL_{s+b}$  for different values of  $\mu$  in the 0/1 jet bin of the  $H \rightarrow \tau\tau$  analysis. The vertical red line indicates the  $CL_{s+b} = 0.68$  and the values at which the curve crosses this line (indicated by the horizontal red lines) form the 68% confidence interval for  $\mu$ .

The procedure is easily extended to a higher number of dimensions by exchanging the test-statistic for that given in Equation 4.10. Pseudo-datasets are generated and fit as before and the union of points for which  $1 - CL_{s+b} < 0.68$  defines a confidence-contours in an  $n$ -dimensional parameter space.

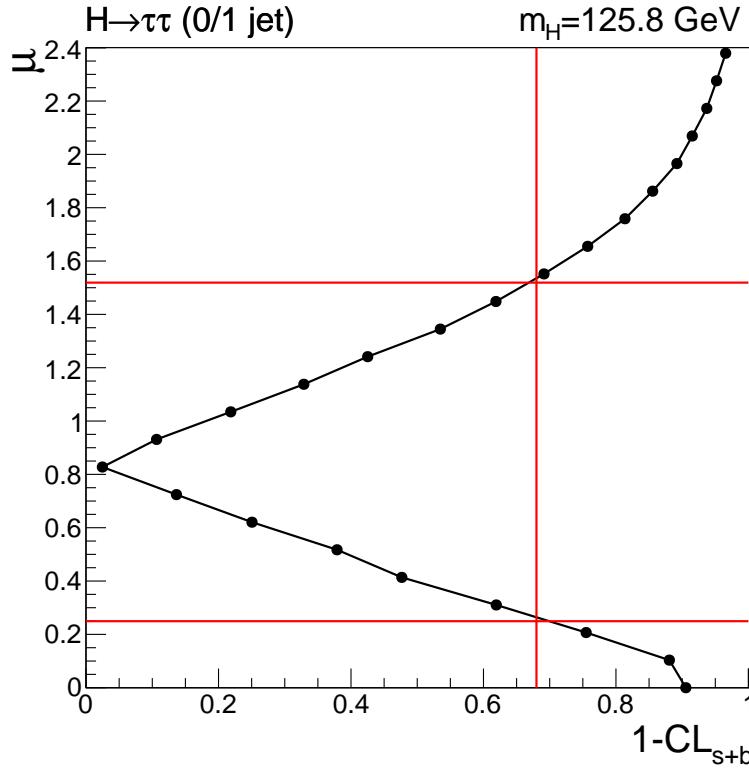
#### 4.2.2. Combined Mass Measurement

The mass of the Higgs boson is a free parameter in the context of the Standard Model. The high resolution channels,  $H \rightarrow \gamma\gamma$  and  $H \rightarrow ZZ \rightarrow 4l$ , provide the largest constraint on the mass of the new particle as the signal is visible as a narrow peak in the invariant



**Figure 4.4.:** Distributions of the test statistic  $q_\mu$  for the 0/1 jet bin of the  $H \rightarrow \tau\tau$  analysis at the combined best fit mass,  $m_H = 125.8$  GeV. The green and yellow filled regions indicate the 68% and 95% quantiles of the distribution respectively. The left distribution is generated at  $\mu = 2.28$  which lies outside of the 68% confidence interval while the right distribution is generated at  $\mu = 1.34$  which lies inside the 68% confidence interval. The values of the test statistic obtained from the observed data,  $q_\mu^{obs}$  is indicated by the solid vertical lines.

mass of the decay products. To measure the mass,  $m_X$  of the particle in a model-independent way, the signal strengths for the  $gg \rightarrow H \rightarrow \gamma\gamma$ ,  $qqH + VH \rightarrow H \rightarrow \gamma\gamma$  and  $H \rightarrow ZZ \rightarrow 4l$  are assumed independent and thus are treated as nuisance parameters in the likelihood. Each of the signals in these channels are assumed to be due to the presence of a single state with mass  $m_X$ . Figure 4.6 (left) shows the value of the test-statistic  $q_{m_X}$  for the  $H \rightarrow \gamma\gamma$ ,  $H \rightarrow ZZ$  channels and their combination near the best fit points. From the combination, the mass is determined to be  $m_X = 125.8 \pm 0.5$  GeV. The 68% confidence interval is determined from the values of  $m_X$  at which the curve crosses the horizontal red line at 1. Large background fluctuations in the  $H \rightarrow \gamma\gamma$  channel can result in large variations of the measured mass of a small signal. Conversely, the kinematic constraints on the  $4l$  system cause a large variation in the branching ratio of  $H \rightarrow ZZ \rightarrow 4l$  as a function of  $m_H$ . Figure 4.6 (right) shows the two-dimensional 68% confidence intervals in  $m_X$  and  $\sigma/\sigma_{SM}$  for the  $H \rightarrow ZZ \rightarrow 4l$ ,  $H \rightarrow \gamma\gamma$  and their combination. For this combination, the relative signal strength between the two channels are kept fixed to the SM expectation; only the overall signal strength is kept as a free parameter. The best fit value of  $m_X$  is consistent with the value determined in the

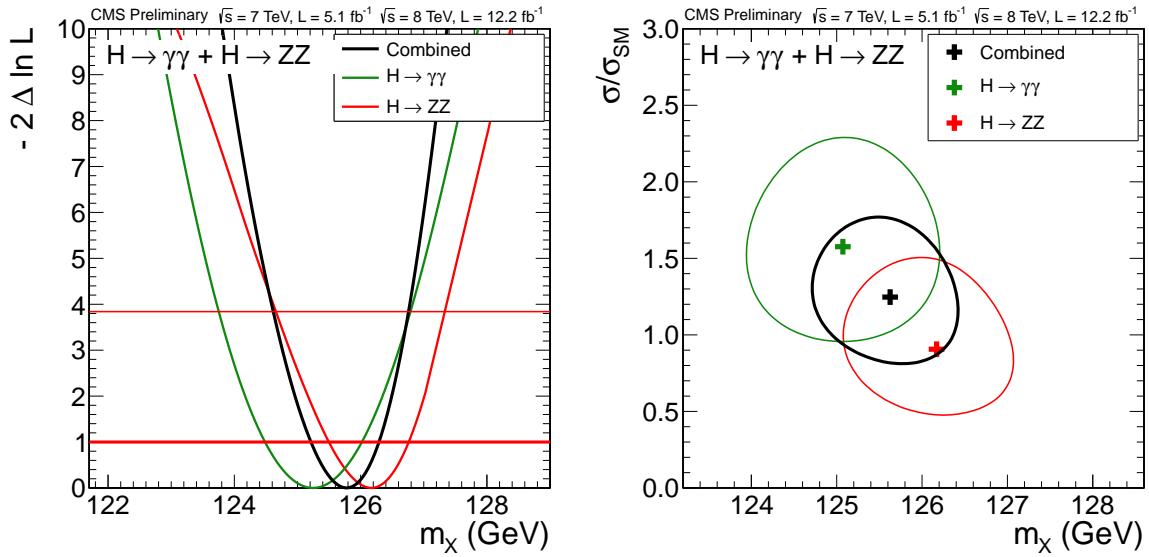


**Figure 4.5.:** Confidence level evaluation curve for the  $H \rightarrow \tau\tau$  analysis in the (0/1) jet bin. At each point, pseudo-data are generated with signal injected at the given value of  $\mu$  and its confidence level (CL) calculated. Linear interpolation between the generated points is used to determine the 68% confidence interval; the two values of  $\mu$  (red horizontal lines) which cross the curve at  $CL = 0.68$  (vertical red line).

one-dimensional case. The best fit value for the combined signal strength relative to the Standard Model is  $0.88 \pm 0.21$  for a mass of 125.8 GeV.

#### 4.2.3. Compatibility with the Standard Model

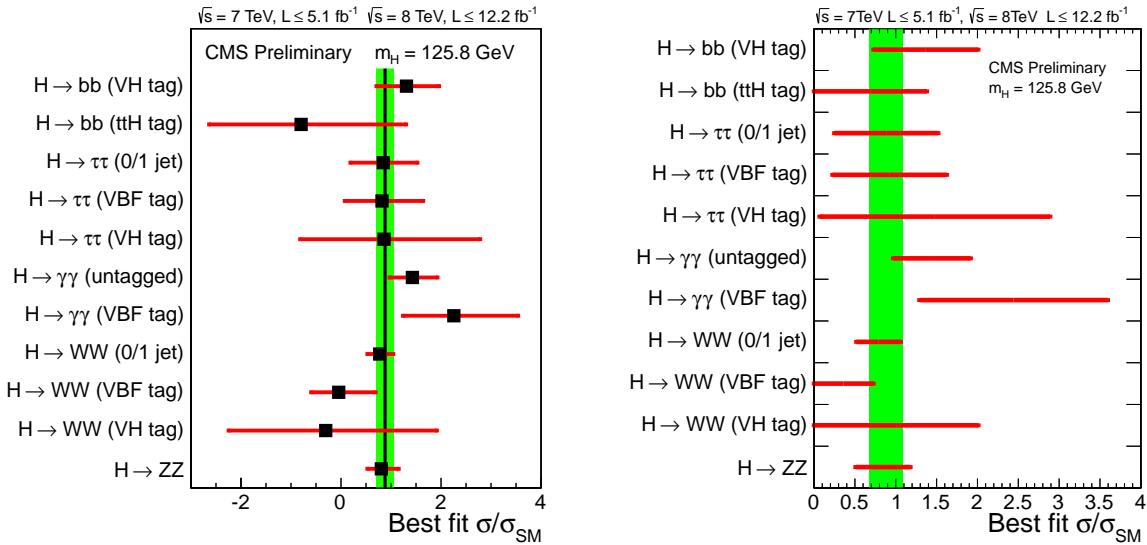
The Standard Model makes very precise predictions for the coupling of the Higgs boson to all of the known fundamental particles which influence its various rates of decay and production. Precise measurements of these rates in the combined search channels therefore provides information on the couplings relative to those predicted by the SM. Significant deviations from those values would indicate the presence of new physics which cannot be explained by the SM.



**Figure 4.6.:** Left: One-dimensional scan of  $q_{m_x}$  for the  $H \rightarrow \gamma\gamma$ ,  $H \rightarrow ZZ$  channels and their combination. For the combination, the relative signal strengths between the channels are allowed to float. The 68% and 95% confidence intervals for  $m_X$  are determined as the values at which the curves cross the horizontal red lines. Right: 68% confidence contours in  $m_X$  and  $\sigma/\sigma_{SM}$  for the  $H \rightarrow \gamma\gamma$  and  $H \rightarrow ZZ$  channels and their combination. For this combination, the relative signal strengths of the channels are kept fixed to the SM expectation.

### Channel Compatibility

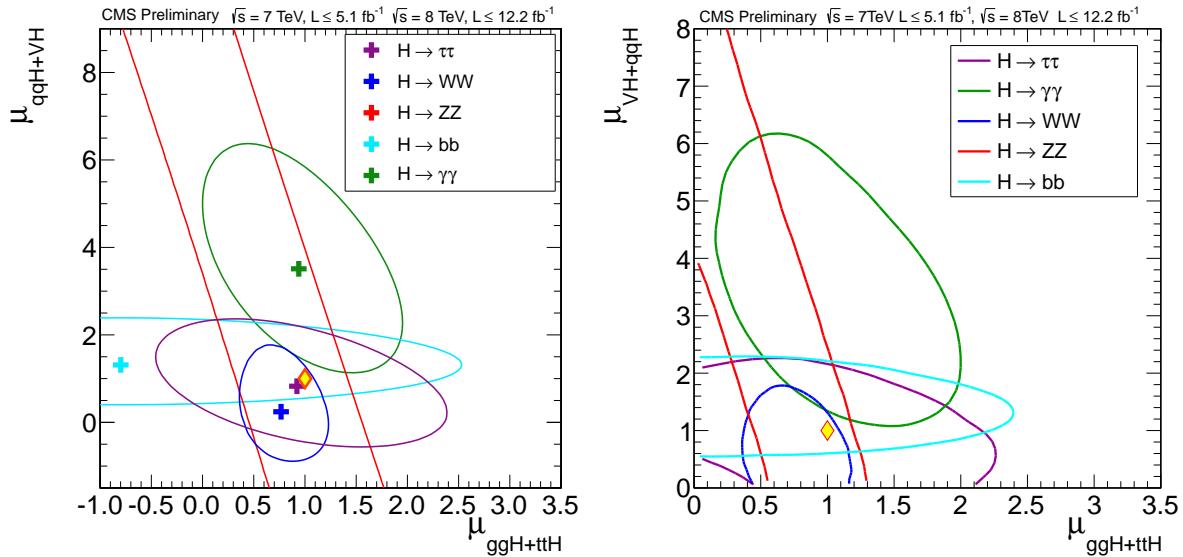
When determining the preferred value of  $\mu$  in the combined data, the ratios of decay rates to each contributing channel relative to that predicted by the SM are kept constant. By relaxing this constraint, the compatibility of the new state with the SM Higgs boson can be studied on a per-decay/per-production level. Due to the limited amount of data collected at CMS, some of the channels and sub-channels entering the combination have a negative value for the best fit signal strength ( $\mu = \sigma/\sigma_{SM}$ ). In order to avoid quoting unphysical values in each channel, the Feldman-Cousins [28] procedure is used to determine 68% confidence intervals for  $\sigma/\sigma_{SM}$  separately in the different channels/sub-channels entering the combination. Figure 4.7 shows the 68% confidence intervals on  $\sigma/\sigma_{SM}$  for the sub-channels included in the combination obtained from the HCP dataset. The intervals are extracted for a Higgs boson mass  $m_H = 125.8$  GeV, which is the overall best fit mass of the new state obtained from the same dataset. The 68% confidence interval on  $\sigma/\sigma_{SM}$  for the full combination is indicated by the green band. With the exception of the dijet (VBF) tagged channel in the  $H \rightarrow \gamma\gamma$  analysis, all of the intervals contain the value  $\mu = 1$  which is the expected value for a SM Higgs boson. The results



**Figure 4.7.:** 68% confidence intervals for  $\mu = \sigma/\sigma_{SM}$  for individual channels or combination of sub-channels determined by scanning the likelihood (left) and the Feldman-Cousins procedure (right). The value of  $\sigma/\sigma_{SM}$  denotes the production cross-section times the relevant branching fraction for a given channel, relative to the SM. The green band indicates the 68% confidence interval on  $\sigma/\sigma_{SM}$  for all channels combined. The intervals are determined at the best fit mass,  $m_H = 125.8$  GeV.

are compared with the intervals determined directly from a scan of  $q_\mu$ , as shown in the same figure. The two methods are found to be in good agreement.

Several of the analyses which are combined in the search for the Higgs boson use selections (tags) which are specifically designed to enhance the sensitivity to particular Higgs boson production modes. The  $H \rightarrow WW$ ,  $H \rightarrow \tau\tau$  and  $H \rightarrow \gamma\gamma$  analyses all include a dijet (or VBF tagged) category which are designed predominantly select events produced via vector-boson fusion ( $qqH$ ). Additional sensitivity is gained in the  $H \rightarrow WW$ ,  $H \rightarrow \tau\tau$  and  $H \rightarrow bb$  channels by looking for additional leptons or  $E_T$  in association with production of a vector boson ( $VH$ ). The production rates associated to couplings with top-quarks ( $ggH$  and  $ttH$ ) and vector bosons ( $qqH$  and  $VH$ ) are determined by removing the requirement that the relative production cross-sections  $\mu_{ggH+ttH}$  and  $\mu_{VH+qqH}$  are equal. The compatibility of the rates observed in data with respect to those predicted by the Standard Model are tested using the Feldman-Cousins procedure. The relative branching ratios to each of the five observable final states are left unconstrained. Figure 4.8 shows the 68% confidence contours for each of the five decay processes. With the exception of the  $H \rightarrow ZZ$  analysis, the explicit exploitation of the different production modes leads to elliptical contours. The SM point (1, 1), indicated by



**Figure 4.8.:** 68% confidence contours for the production cross-section in  $ggH$  and  $ttH$  modes ( $\mu_{ggH+ttH}$ ), and  $VH$  and  $qqH$  modes ( $\mu_{VH+qqH}$ ), relative to the SM determined by scanning the likelihood (left) and the Feldman-Cousins procedure (right). Each colour indicates the result by combining all sub-channels in a particular decay mode. The yellow diamond at (1, 1) indicates the SM values. The contours are determined at the best fit mass,  $m_H = 125.8 \text{ GeV}$ .

the yellow diamond, is contained within the 68% confidence contours from each decay channel with the exception of  $H \rightarrow \gamma\gamma$ . The same contours are also determined from a scan of  $q_\mu$ , as shown in the same figure. Good agreement is found when comparing the two methods.

### Coupling Measurements

The couplings of the Higgs boson cannot be directly measured in the experimental data. In order to measure these couplings, the rates of production and decay in the various channels must be interpreted in terms of the underlying couplings to the SM particles. For the purposes of measuring the couplings, the following simplifications are made:

- Signals observed in each of the different search channels originate from a single resonance near 125 GeV.

	$H \rightarrow \gamma\gamma$	$H \rightarrow ZZ/H \rightarrow WW$	$H \rightarrow bb/H \rightarrow \tau\tau$
$ggH/ttH$	$\frac{\kappa_f^2 \kappa_V^2 (\kappa_f, \kappa_V)}{\kappa_H (\kappa_f, \kappa_V)^2}$	$\frac{\kappa_f^2 \kappa_V^2}{\kappa_H (\kappa_f, \kappa_V)^2}$	$\frac{\kappa_f^2 \kappa_f^2}{\kappa_H (\kappa_f, \kappa_V)^2}$
$qqH/VH$	$\frac{\kappa_V^2 \kappa_f^2 (\kappa_f, \kappa_V)}{\kappa_H (\kappa_f, \kappa_V)^2}$	$\frac{\kappa_V^2 \kappa_V^2}{\kappa_H (\kappa_f, \kappa_V)^2}$	$\frac{\kappa_V^2 \kappa_f^2}{\kappa_H (\kappa_f, \kappa_V)^2}$

**Table 4.3.:** Boson and fermion vertex scaling as a function of  $\kappa_V$  and  $\kappa_f$  for each production/decay included in the combination. Each cell represents the scaling factor applied to the production (row) decay (column) combination.

- The natural width of the resonance is small enough to be neglected such that the cross-section of the signal in each channel can be expressed as

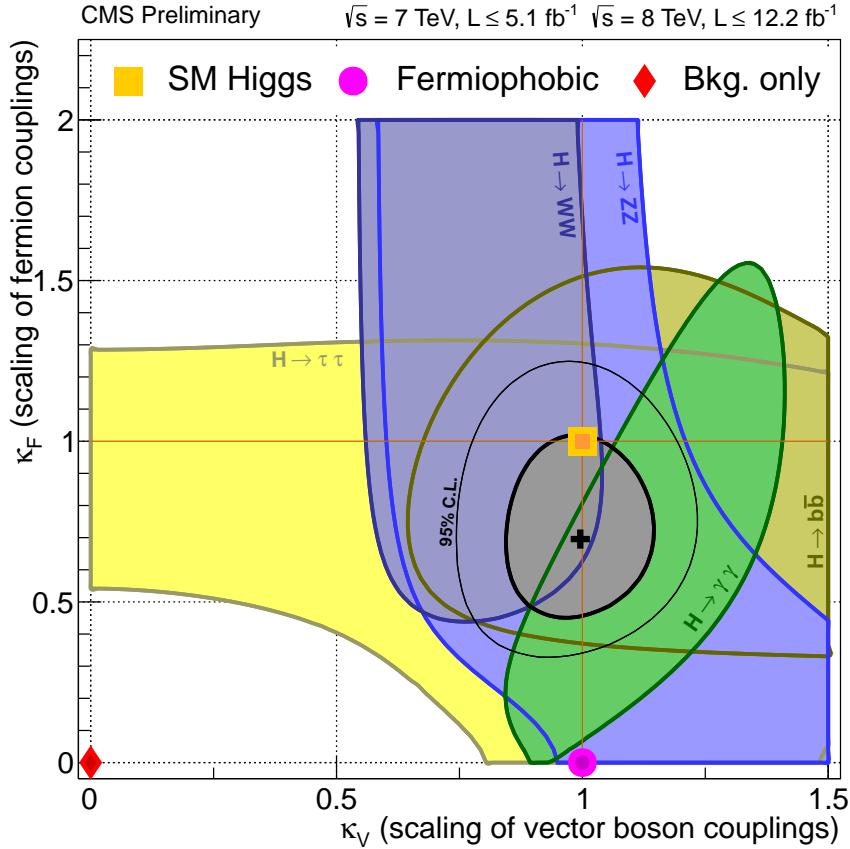
$$(\sigma \cdot BR)(ii \rightarrow H \rightarrow ff) = \frac{\sigma_{ii} \Gamma_{ff}}{\Gamma}, \quad (4.11)$$

where  $\sigma_{ii}$  is the production cross-section through the initial state  $ii$ ,  $\Gamma_{ff}$  is the partial decay width to the final state  $ff$  and  $\Gamma$  is the total width.

- Only modifications of the absolute values of the coupling strengths are allowed. The structure of the couplings is fixed to the SM, in particular this means the new state is assumed to be a CP-even scalar.

In general, no specific assumptions are made on any additional states of new physics which could influence the phenomenology of the 125 GeV state. A number of frameworks to investigate the coupling structure of the new particle are used at CMS [35]. The simplest of these is an unfolding of the production cross-section modifiers  $\mu_{ttH+ggH}$ ,  $\mu_{VH+qqH}$  by expressing them as functions of the couplings to fermions  $\kappa_f$  and vector bosons  $\kappa_V$ . The decay rates to each channel are also expressed as functions of these parameters such that the overall yield in each channel relative to the SM expectation is parameterized. The ratio of the total width to that predicted by the SM is denoted  $\kappa_H = \Gamma/\Gamma_{SM}$ . Table 4.3 shows the parameterization of  $(\sigma \cdot BR)(ii \rightarrow H \rightarrow ff)$  for each production/decay included in the combination. The parameters  $\kappa_f$  and  $\kappa_V$  are the couplings relative to the SM predictions for the Higgs such that the SM is recovered setting  $\kappa_V = \kappa_f = 1$ . No invisible final states are assumed so that the total width,  $\Gamma$ , is a function of  $\kappa_V$  and  $\kappa_f$ .

Figure 4.9 shows the best fit values in the observed data for  $\kappa_V$  and  $\kappa_f$  68% confidence contours determined from a scan of  $q_\mu$ . The values are extracted independently in each decay channel and from the full combination. In addition to the SM point, the fermiophobic Higgs scenario, in which the Higgs does not couple to fermions, is indicated.



**Figure 4.9.:** The 68% confidence contours extracted from data in the individual decay channels (coloured regions) and the full combination (solid line). The yellow square shows the SM value, while the fermiophobic and background only scenarios are indicated by the pink dot and red diamond respectively.

The data are compatible with the expectation of a SM Higgs boson; the SM point ( $\kappa_V = \kappa_f = 1$ ) lies within the 95% confidence contour defined by the data.

# **Chapter 5.**

## **Conclusions**

≈ 1 page

# **Appendix A.**

## **A.1. Common Tools**

A large number of physics analyses at CMS use techniques which are common to many areas of experimental particle physics. The following section contains brief descriptions of several tools which are commonly used at CMS to ensure a high level of accuracy and quality of physics analyses and obtain the most from the data available.

Should also include the l1 JEC closure tests (though they are in the AN) Definitely include the table of the coefficients...

### **A.1.1. Isolation Sums**

### **A.1.2. Boosted Decision Trees**

# Appendix B.

## B.1. Energy Scale and Resolution Measurements

The energy scale and resolution is measured in the 2011 dataset using  $Z \rightarrow e^+e^-$  events as described in Section 3.2.1. The additional resolution required to match the  $Z \rightarrow e^+e^-$  peak in MC to that of the data (Table ??) is used to correct the Higgs MC for modelling the signal in the  $H \rightarrow \gamma\gamma$  analysis. The scale measurements (Tables ?? and ?? ) are used to correct the energy of the photons in data.

Category	$\sigma_E/E$ (%)
EB, $ \eta  < 1$ , $R9 > 0.94$ , NOT GAP	$0.67^{+0.10}_{-0.33} \pm 0.22$
EB, $ \eta  < 1$ , $R9 > 0.94$ , GAP	$0.77^{+0.06}_{-0.12} \pm 0.22$
EB, $ \eta  < 1$ , $R9 < 0.94$	$0.96^{+0.05}_{-0.05} \pm 0.24$
EB, $ \eta  > 1$ , $R9 > 0.94$	$1.41^{+0.15}_{-0.33} \pm 0.60$
EB, $ \eta  > 1$ , $R9 < 0.94$	$1.96^{+0.06}_{-0.07} \pm 0.59$
EE, $ \eta  < 2$ , $R9 > 0.94$	$2.68^{+0.15}_{-0.20} \pm 0.90$
EE, $ \eta  < 2$ , $R9 < 0.94$	$2.79^{+0.09}_{-0.10} \pm 0.30$
EE, $ \eta  > 2$ , $R9 > 0.94$	$2.93^{+0.08}_{-0.08} \pm 0.34$
EE, $ \eta  > 2$ , $R9 < 0.94$	$3.01^{+0.11}_{-0.12} \pm 0.52$

**Table B.1.:** Additional energy resolution included in the  $H \rightarrow \gamma\gamma$  signal model measured from comparison of  $Z \rightarrow e^+e^-$  data and MC. The label “NOT GAP” indicates superclusters whose seed crystal is located more than 5 crystals away from an ECAL module boundary whereas the label “GAP” indicates superclusters for whose seed crystal is within 5 crystals of an ECAL module boundary [26].

Category	Run Range	$\Delta P$
EB, $ \eta  < 1, r_9 < 0.94$	160431 - 167913	$-0.0004 \pm 0.0002 \pm 0.0019$
EB, $ \eta  < 1, r_9 < 0.94$	170000 - 172619	$-0.0016 \pm 0.0002 \pm 0.0019$
EB, $ \eta  < 1, r_9 < 0.94$	172620 - 173692	$-0.0017 \pm 0.0002 \pm 0.0019$
EB, $ \eta  < 1, r_9 < 0.94$	175830 - 177139	$-0.0021 \pm 0.0002 \pm 0.0019$
EB, $ \eta  < 1, r_9 < 0.94$	177140 - 178421	$-0.0025 \pm 0.0002 \pm 0.0019$
EB, $ \eta  < 1, r_9 < 0.94$	178424 - 180252	$-0.0024 \pm 0.0002 \pm 0.0019$
EB, $ \eta  < 1, r_9 > 0.94$	160431 - 167913	$0.0059 \pm 0.0002 \pm 0.0013$
EB, $ \eta  < 1, r_9 > 0.94$	170000 - 172619	$0.0046 \pm 0.0002 \pm 0.0013$
EB, $ \eta  < 1, r_9 > 0.94$	172620 - 173692	$0.0045 \pm 0.0002 \pm 0.0013$
EB, $ \eta  < 1, r_9 > 0.94$	175830 - 177139	$0.0042 \pm 0.0002 \pm 0.0013$
EB, $ \eta  < 1, r_9 > 0.94$	177140 - 178421	$0.0038 \pm 0.0002 \pm 0.0013$
EB, $ \eta  < 1, r_9 > 0.94$	178424 - 180252	$0.0039 \pm 0.0002 \pm 0.0013$
EB, $ \eta  > 1, r_9 < 0.94$	160431 - 167913	$-0.0045 \pm 0.0006 \pm 0.0071$
EB, $ \eta  > 1, r_9 < 0.94$	170000 - 172619	$-0.0066 \pm 0.0008 \pm 0.0071$
EB, $ \eta  > 1, r_9 < 0.94$	172620 - 173692	$-0.0058 \pm 0.0007 \pm 0.0071$
EB, $ \eta  > 1, r_9 < 0.94$	175830 - 177139	$-0.0073 \pm 0.0006 \pm 0.0071$
EB, $ \eta  > 1, r_9 < 0.94$	177140 - 178421	$-0.0075 \pm 0.0006 \pm 0.0071$
EB, $ \eta  > 1, r_9 < 0.94$	178424 - 180252	$-0.0071 \pm 0.0007 \pm 0.0071$
EB, $ \eta  > 1, r_9 > 0.94$	160431 - 167913	$0.0084 \pm 0.0013 \pm 0.0051$
EB, $ \eta  > 1, r_9 > 0.94$	170000 - 172619	$0.0063 \pm 0.0014 \pm 0.0051$
EB, $ \eta  > 1, r_9 > 0.94$	172620 - 173692	$0.0071 \pm 0.0013 \pm 0.0051$
EB, $ \eta  > 1, r_9 > 0.94$	175830 - 177139	$0.0056 \pm 0.0013 \pm 0.0051$
EB, $ \eta  > 1, r_9 > 0.94$	177140 - 178421	$0.0054 \pm 0.0013 \pm 0.0051$
EB, $ \eta  > 1, r_9 > 0.94$	178424 - 180252	$0.0058 \pm 0.0013 \pm 0.0051$

**Table B.2.:** Relative energy scale difference in data and MC ( $\Delta P$ ) in the ECAL barrel, measured in  $Z \rightarrow e^+e^-$  data. The first uncertainty given is statistical while the second is the systematic assigned to cover the difference in the  $r_9$  distributions between electrons and photons [26].

Category	Run Range	$\Delta P$
EE, $ \eta  < 2, r_9 < 0.94$	160431 - 167913	$-0.0082 \pm 0.0008 \pm 0.0088$
EE, $ \eta  < 2, r_9 < 0.94$	170000 - 172619	$-0.0025 \pm 0.0011 \pm 0.0088$
EE, $ \eta  < 2, r_9 < 0.94$	172620 - 173692	$-0.0035 \pm 0.0010 \pm 0.0088$
EE, $ \eta  < 2, r_9 < 0.94$	175830 - 177139	$-0.0017 \pm 0.0009 \pm 0.0088$
EE, $ \eta  < 2, r_9 < 0.94$	177140 - 178421	$-0.0010 \pm 0.0009 \pm 0.0088$
EE, $ \eta  < 2, r_9 < 0.94$	178424 - 180252	$0.0030 \pm 0.0009 \pm 0.0088$
EE, $ \eta  < 2, r_9 > 0.94$	160431 - 167913	$-0.0033 \pm 0.0010 \pm 0.0018$
EE, $ \eta  < 2, r_9 > 0.94$	170000 - 172619	$0.0024 \pm 0.0012 \pm 0.0018$
EE, $ \eta  < 2, r_9 > 0.94$	172620 - 173692	$0.0014 \pm 0.0011 \pm 0.0018$
EE, $ \eta  < 2, r_9 > 0.94$	175830 - 177139	$0.0032 \pm 0.0010 \pm 0.0018$
EE, $ \eta  < 2, r_9 > 0.94$	177140 - 178421	$0.0040 \pm 0.0010 \pm 0.0018$
EE, $ \eta  < 2, r_9 > 0.94$	178424 - 180252	$0.0079 \pm 0.0010 \pm 0.0018$
EE, $ \eta  > 2, r_9 < 0.94$	160431 - 167913	$-0.0064 \pm 0.0008 \pm 0.0019$
EE, $ \eta  > 2, r_9 < 0.94$	170000 - 172619	$-0.0046 \pm 0.0009 \pm 0.0019$
EE, $ \eta  > 2, r_9 < 0.94$	172620 - 173692	$-0.0029 \pm 0.0009 \pm 0.0019$
EE, $ \eta  > 2, r_9 < 0.94$	175830 - 177139	$-0.0040 \pm 0.0009 \pm 0.0019$
EE, $ \eta  > 2, r_9 < 0.94$	177140 - 178421	$-0.0050 \pm 0.0008 \pm 0.0019$
EE, $ \eta  > 2, r_9 < 0.94$	178424 - 180252	$-0.0059 \pm 0.0009 \pm 0.0019$
EE, $ \eta  > 2, r_9 > 0.94$	160431 - 167913	$0.0042 \pm 0.0006 \pm 0.0028$
EE, $ \eta  > 2, r_9 > 0.94$	170000 - 172619	$0.0060 \pm 0.0008 \pm 0.0028$
EE, $ \eta  > 2, r_9 > 0.94$	172620 - 173692	$0.0077 \pm 0.0007 \pm 0.0028$
EE, $ \eta  > 2, r_9 > 0.94$	175830 - 177139	$0.0067 \pm 0.0007 \pm 0.0028$
EE, $ \eta  > 2, r_9 > 0.94$	177140 - 178421	$0.0056 \pm 0.0007 \pm 0.0028$
EE, $ \eta  > 2, r_9 > 0.94$	178424 - 180252	$0.0047 \pm 0.0007 \pm 0.0028$

**Table B.3.:** Relative energy scale difference in data and MC ( $\Delta P$ ) in the ECAL endcaps, measured in  $Z \rightarrow e^+e^-$  data. The first uncertainty given is statistical while the second is the systematic assigned to cover the difference in the  $r_9$  distributions between electrons and photons [26].

## B.2. Binning Algorithm Optimisation

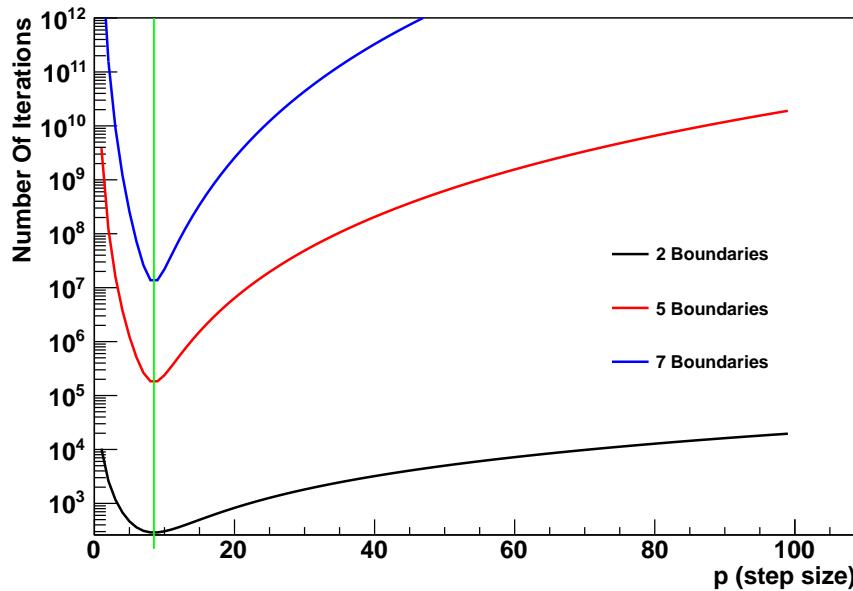
The optimisation procedure used to select the bin boundaries of the  $H \rightarrow \gamma\gamma$  categorisation BDT involves a full scan over all combinations of bin boundaries. As this scan can be very slow, the procedure is separated into two parts, first a broad scan in large steps to find the region containing the optimum point then using small steps to refine the scan. The first step in the binning procedure is designed to ensure that at least 20 background events are expected in every bin. This gives a total of  $B$  bins at a given luminosity. To maintain this feature, only boundaries which match any of the  $B - 1$  bin edges (remembering -1 and 1 are fixed boundaries) are scanned. The step size of the scan is therefore expressed as a step in number of bins so that for a given BDT output range,  $(b_i, b_j)$  includes an integer number of the  $B$  bins. The fine scan is defined to have a step size of 1, being the minimum step size defined this way. The step size for the broad scan,  $P$ , can be chosen to reduce the total time taken for the scan. For  $N$  BDT boundaries, the scan is  $N$ -dimensional and the total number of points to scan (combinatins of bin boundary values) assuming the two step procedure is given by,

$$\frac{1}{2^{N-1}} \left( \left( \frac{B}{P} \right)^N + (2P)^N \right) \quad (\text{B.1})$$

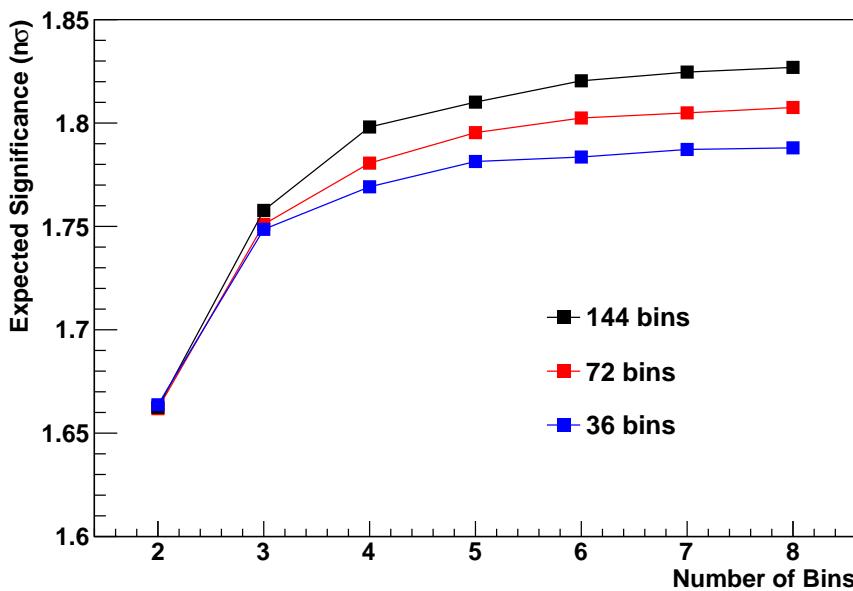
imposing the condition  $b_1 < b_2 < \dots < b_N$ . Figure B.1 shows the total number of iterations required to perform the full scan for different numbers of boundaries as a function of the broad step stize  $P$ . The value,  $P_{min}$ , which minimises the total number of iterations is the same for any value of  $N$  and is given by,

$$P_{min} = e^{\frac{1}{2} \ln(B/2)} \quad (\text{B.2})$$

The scan is repeated, increasing the number of boundaries until the improvement in terms of the maximum expected significance in the presence of a SM Higgs is less than 0.1%. Figure B.2 shows the additinal sensitivity gained as the number of final BDT output bins is increased for different starting values of  $B$ . The red curve is representative of the actual scan performed for the 2011 analysis.



**Figure B.1.:** Total number of iterations in binning optimization scan as a function of the broad step size  $P$ . The curve is shown for different numbers of final BDT boundaries. The minimum always occurs at the same value of  $P$  as indicated by the green vertical line.



**Figure B.2.:** Increase in expected significance in the presence of a SM Higgs as the number of final BDT output bins is increased. The three curves show the improvement for different numbers of initial bins,  $B$ . The red curve is representative of the result obtained from performing the optimization procedure in the 2011 analysis.

### B.3. Signal Systematics

The treatment of systematic variations in the signal modelling for the  $H \rightarrow \gamma\gamma$  analysis described in Chapter 3 is the same for all uncertainties except those due to theoretical uncertainty on the Higgs production cross-sections and the integrated luminosity measurement. For each uncertainty, the relevant quantity in the MC is varied by 3 standard deviations and the resulting BDT distributions are compared to the nominal one. The three “templates” (corresponding to nominal and  $\pm 3\sigma$  variations) are used to determine the  $1\sigma$  variations of the  $j$ -th BDT bin of the signal model due to the  $k$ -th signal systematic ( $\sigma_k^{s,p}$  used in Equation 3.4.6). The procedure is performed for each signal process,  $p$ , separately. The value for the  $1\sigma$  variation in each bin is given by,

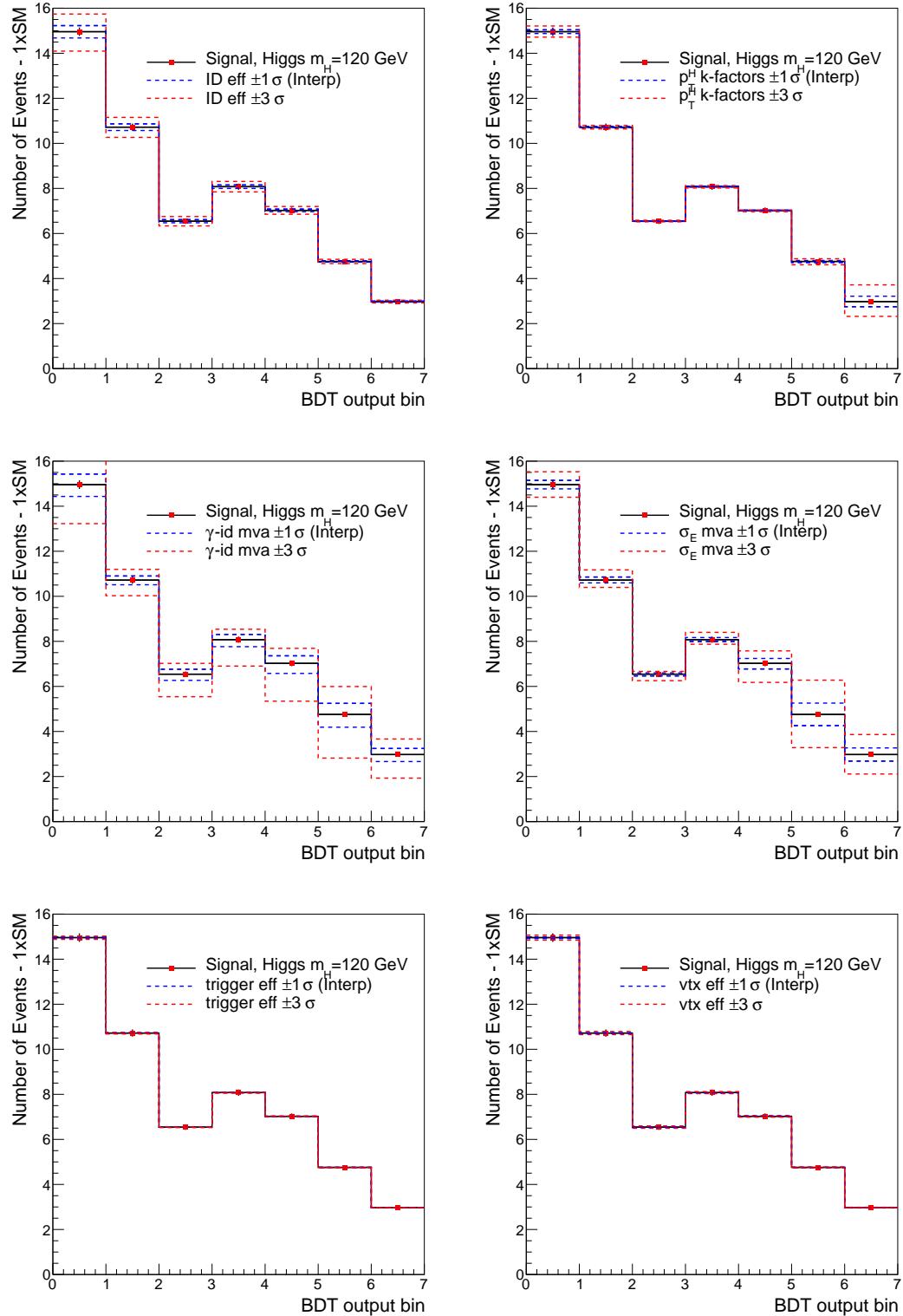
$$\sigma^\pm = a \pm b + c, \quad (\text{B.3})$$

where  $\sigma^+$  is the value of  $\sigma_k^{s,p}$  used for positive values of the associated nuisance parameter and  $\sigma^-$  is for negative values. The parameters  $a$  and  $b$  are determined for a particular bin by solving the set of simultaneous equations;

$$\begin{pmatrix} s^{-3\sigma} \\ s^{mc} \\ s^{3\sigma} \end{pmatrix} = \begin{bmatrix} 9 & -3 & 1 \\ 0 & 0 & 1 \\ 9 & 3 & 1 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad (\text{B.4})$$

where  $s^{mc}$  is the nominal value for the signal in that bin and  $s^{\pm 3\sigma}$  are the values determined from the to the  $\pm 3\sigma$  templates.

Figure B.3 shows the  $\pm 3\sigma$  amnd  $1\sigma$  variations of the BDT distribution expected from the  $ggH$  production process calculated from MC and using the interpolation procedure respectively. The distributions are normalised to the expectation in  $5.1 \text{ fb}^{-1}$ . The energy scale and resolution uncertainties can be found in Section 3.4.5 (Figure 3.25)



**Figure B.3.:** Systematic uncertainties on the  $ggH$  signal model. The effect of  $\pm 3\sigma$  variations derived in MC are shown with red dashed lines while the interpolated  $\pm 3\sigma$  are shown with blue.



# Bibliography

- [1]
- [2]
- [3]
- [4]
- [5]
- [6]
- [7]
- [8]
- [9]
- [10]
- [11]
- [12]
- [13] Tracking and primary vertex results in first 7 TeV collisions. 2010.
- [14] Herv Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.
- [15] D Acosta, Michel Della Negra, L Fo, A Herv, and Achille Petrilli. *CMS physics: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 2006.
- [16] W Adam, R Frhwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the cms tracker at the lhc. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9, 2005.

- [17] S. Agostinelli, J. Allison, and K. Amako et al. Geant4a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003.
- [18] J. Beringer, J. F. Arguin, and Barnett et al. Review of Particle Physics. *Phys. Rev. D*, 86:010001, Jul 2012.
- [19] G. Bozzi, S. Catani, D. de Florian, and M. Grazzini. The qT spectrum of the Higgs boson at the LHC in QCD perturbation theory. *Physics Letters B*, 564(12):65 – 72, 2003.
- [20] J. Brooke, Mathias B., Tapper A., and Wardle N. Calibration and Performance of the Jets and Energy Sums in the Level-1 Trigger. 2012.
- [21] CERN. CMS compact muon solenoid,[public.web.cern.ch/public/Objects/LHC/CMSnc.jpg](http://public.web.cern.ch/public/Objects/LHC/CMSnc.jpg), Feb 2010.
- [22] S. Chatrchyan and V. Khachatryan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30 – 61, 2012.
- [23] S. Chatrchyan and V. Khachatryan et al. Search for the standard model Higgs boson decaying into two photons in pp collisions at CMS. *Physics Letters B*, 710(3):403 – 425, 2012.
- [24] CMS Collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *Journal of Instrumentation*, 6:11002, November 2011.
- [25] P. D. Dauncey, Kenzie M., and C. Seez. Residual photon energy corrections and resolution from simulation. 2012.
- [26] A. Benaglia et al. Search for a Standard Model Higgs boson decaying into two photons. 2012.
- [27] A. Benaglia et al. Search for a Standard Model Higgs boson decaying into two photons employing multivariate methods. 2012.
- [28] Gary J. Feldman and Robert D. Cousins. Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, 57:3873–3889, Apr 1998.
- [29] Eilam Gross and Ofer Vitells. Trial factors for the look elsewhere effect in high energy physics. *European Physical Journal C*, 70:525–530, 2010.

- [30] A. Hoecker and Speckmayer et al. TMVA - Toolkit for Multivariate Data Analysis. *ArXiv Physics e-prints*, March 2007.
- [31] F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput.Phys.Commun.*, 10:343–367, 1975.
- [32] LHC Higgs Cross Section Working Group, S. Dittmaier, and Mariotti et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. *ArXiv e-prints*, January 2011.
- [33] L. Lyons. Open statistical issues in particle physics. *ArXiv e-prints*, November 2008.
- [34] E. Meschi, Monteiro T., Seez C., and Vikas P. Electron Reconstruction in the CMS Electromagnetic Calorimeter. 2001.
- [35] A. Nobody. Null for now.
- [36] C. Oleari. The POWHEG BOX. *Nuclear Physics B Proceedings Supplements*, 205:36–41, August 2010.
- [37] A L Read. Presentation of search results: the cl s technique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693, 2002.
- [38] A. L. Read. Presentation of search results: the CLs technique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693, 2002.
- [39] Baffioni S., Charlot C., Ferri F., Futyan D., Meridiani P., Puljak I., Rovelli C., Salerno R., and Sirois Y. Electron reconstruction in CMS. 2006.
- [40] T. Sjöstrand, S. Mrenna, and P. Skands. PYTHIA 6.4 physics and manual. *Journal of High Energy Physics*, 5:26, May 2006.
- [41] Tomasz Skwarnicki. *A study of the radiative cascade transitions between the Upsilon-prime and Upsilon resonances*. PhD thesis, Institute of Nuclear Physics, Krakow, 1986. <http://inspirehep.net/record/230779/files/230779.pdf> DESY-F31-86-02.
- [42] Adam W., S. Baffioni, F. Beaudette, D. Benedetti, C. Broutin, D. Chamont, C. Charlot, E. DiMarco, D. Futyan, S. Harper, D. Lelas g, A. Martelli, P. Meridiani, M. Pioppo, I. Puljak, D. Sabes, R. Salerno, M. Sani, C. Seez, Y. Sirois, P. Vanlaer, and D. Wardrope. Electron reconstruction in CMS. 2009.

# List of Figures

2.1.	LHC accelerator ring. The relative locations of the four main experiments are indicated along with their points of access to the beam. . . . .	4
2.2.	Diagram of the CMS Detector. The arrows indicate the main detector elements. The figure has been altered from its original source [21] . . . . .	6
2.3.	Cross-section of the pixel and silicon strip detector components of the CMS tracker [1]. . . . .	7
2.4.	Resolution of vertex $z$ -position as a function of the number of tracks associated to the vertex measured in simulation and 2010 data. The resolution is given for three different average track momenta. . . . .	8
2.5.	Sub-cluster construction of the Hybrid algorithm used to reconstruct photons and electrons in the ECAL barrel. . . . .	10
2.6.	Relative ECAL crystal response to blue laser light (440 nm) in bins of pseudorapidity, for the 2011 data taking period runs. The grey bands indicate periods during which there was no beam. . . . .	12
2.7.	Ratio $E/p$ in electrons reconstructed in the ECAL Barrel from $W \rightarrow e\nu$ events in 2011 data as a function of time before and after applying transparency corrections from the laser monitoring (LM) system. The blue line indicates the correction applied per point averaged over all crystals used in the electron energy measurement. . . . .	13
2.8.	Shower shape variable $r_9$ (left) and $\sigma_{i\eta i\eta}$ (right) distributions for super-clusters associated to simulated real and fake photons. The real photon is taken from simulated $H \rightarrow \gamma\gamma$ events while the fake photon is taken from a $\gamma + jet$ sample where the photon candidate is matched to a generated quark leg. . . . .	14

2.9. Response measured from matched generator-L1 jet pairs in MC simulation as a function of the generator jet pseudo-rapidity $ \eta^{Gen} $ . . . . .	16
2.10. Correction function for the $0.348 <  \eta^{Gen}  < 0.695$ . The points represent the average quantities as measured in MC simulation. The blue line is a parametric fit to the points using a chi-squared minimisation. . . . .	18
2.11. Closure tests performed in MC as a function of $E_T^{L1}$ (left) and $\eta^{Gen}$ (right). The test shows that after applying the corrections, the response is within 10% (dashed lines) of unity. . . . .	19
2.12. Jet energy resolution at L1 as a function of $E_T^{L1}$ before and after application of the derived calibrations. . . . .	20
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 regression method. The quantity $\sigma_{eff}$ , the narrowest range in $m_{\gamma\gamma}$ which contains 68% of the distribution, is given for each peak. . . . .	24
3.2. Invariant mass peak in $H \rightarrow \gamma\gamma$ 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. . . . .	27
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. . . . .	28
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. . . . .	32
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. . . . .	32

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. . . . .	32
3.7. Invariant mass distribution 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. . . . .	32
3.8. Diphoton BDT output distribution in $Z \rightarrow e^+e^-$ 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.	33
3.9. Upper: Per-photon resolution estimator, $\sigma_E$ relative to the measured energy in $Z \rightarrow e^+e^-$ 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 $\sigma_E$ by $\pm 10\%$ . . . . .	34
3.10. Photon ID BDT output in $Z \rightarrow e^+e^-$ 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 $\sigma_E$ by $\pm 0.025\%$ . . . . .	34
3.11. Separation in $\eta$ between two identified jets in data and MC. The expecta- tion 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. . . . .	36
3.12. Figure of merit for selection of the signal region cut value, $w$ . Each color shows the evaluation under different Higgs mass hypotheses. . . . .	38
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. . . . .	39
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. . . . .	40

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 the bins used for extracting the signal (right). . . . .	40
3.16. Comparison of the distributions of BDT output at $m_H = 125$ for data and background MC. The distributions are arbitrarily binned for the purposes of comparison only. . . . .	40
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. . . . .	43
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 $\pm 2\%$ 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. . . . .	44
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. . . . .	46
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). . . . .	46
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. . . . .	47
3.22. Simultaneous fits to the six sidebands in data to determine the background shape for $m_H = 124$ GeV. 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 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. . . . .	47

3.23. Covariance matrix from the sideband fit to determine the background shape at $m_H = 124$ GeV. The covariance matrix includes the additional 20% systematic attributed to possible second order variations in the BDT output background distribution with mass. . . . .	48
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. . . . .	49
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 with red dashed lines while the interpolated $\pm 3\sigma$ are shown with 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. . . . .	52
3.26. Closure test for signal interpolation to intermediate mass points. The solid grey histogram is the result of a linear interpolation between the efficiency $\times$ acceptance in each bin of the blue ( $m_H = 130$ ) and red ( $m_H = 140$ ) histograms. The efficiency $\times$ acceptance from $ggH$ MC generated with mass 135 GeV is shown in black for comparison. . . . .	55
3.27. BDT output distribution for $Z \rightarrow e^+e^-$ 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. . . . .	57
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. . . . .	58
3.29. Distributions of the test statistic $q_\mu$ 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.	61

---

3.30. Exclusion limits on SM higgs production and subsequent decay to two photons in the range $110 < m_H < 150$ GeV. The black dashed line indicates the median expected value for the upper limit on $\mu$ 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. . . . .	63
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. . . . .	65
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 $m_H$ . . . . .	66
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. . . . .	67
3.34. Relationship between local and global p-values to determine the look-elsewhere effect in the $H \rightarrow \gamma\gamma$ 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. . . . .	68
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. . . . .	70

- 3.36. Exclusion limits on SM Higgs production and subsequent decay to two photons (left) and Local p-value,  $p_0$  (right) in the range  $110 < m_H < 150$  GeV from the combined 2011 (7 TeV) and 2012 (8 TeV) datasets. In the left figure, the black dashed lines indicates the median expected value for the upper limit on  $\mu$  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. In the right figure, 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 black dashed line. The observed  $p_0$  from the 2011 (7 TeV) and 2012 (8 TeV) datasets individually are shown in the blue and red dashed lines respectively. The right hand scale shows the significance in standard deviations at each  $m_H$ . . . . .

4.1. Summary plots for the parameter `lumi` of the realistic counting experiment. The entries in the histograms are for fits to toys generated under the background only hypothesis letting  $\mu$  float freely. The bottom, left panel shows the correlation between the value generated for the expectation value of the nuisance `lumi_In` and the fitted value of the parameter. The lower right panel shows the shape of the negative log-likelihood (NLL) as a function of the nuisance parameter. The the parameters of the fitted Gaussian for each histogram is given as Mean and Sigma. The value and error of the nuisance are given before fitting to the data (Pre-fit), followed by the best fit value of the parameter under the background only and signal plus background hypotheses. . . . .

4.2. Combined 95% upper limits on the production cross-section of Higgs boson production relative to that of the Standard Model in the  $m_H$  ranges 110-600 GeV (left) and 110-145 GeV (right). The median upper limits expected in the absence of a SM Higgs is indicated by the dashed black line and the 68% and 95% quantiles by the green and yellow bands respectively. The observed upper limits from the combined ICHEP 2012 dataset is shown by the black solid line. Where the observed limit is lower than 1 (red line) A SM Higgs boson with that  $m_h$  is excluded at the 95% confidence level. . . . .

- 
- 4.3. Left: The observed local  $p$ -value,  $p_0$  for sub-combinations of the low and high resolution channels and the overall combination as a function of  $m_H$ . The dashed line shows the expected  $p_0$  at each  $m_H$  should a SM Higgs boson exist with mass  $m_H$ . Right: Relationship between the local and global  $p_0$  in the range 115-130 GeV. The red line indicates the analytic expression given fit to the relationship derived from 10,000 pseudo-datasets. . . . . 84
- 4.4. Distributions of the test statistic  $q_\mu$  for the 0/1 jet bin of the  $H \rightarrow \tau\tau$  analysis at the combined best fit mass,  $m_H = 125.8$  GeV. The green and yellow filled regions indicate the 68% and 95% quantiles of the distribution respectively. The left distribution is generated at  $\mu = 2.28$  which lies outside of the 68% confidence interval while the right distribution is generated at  $\mu = 1.34$  which lies inside the 68% confidence interval. The values of the test statistic obtained from the observed data,  $q_\mu^{obs}$  is indicated by the solid vertical lines. . . . . 86
- 4.5. Confidence level evaluation curve for the  $H \rightarrow \tau\tau$  analysis in the (0/1) jet bin. At each point, pseudo-data are generated with signal injected at the given value of  $\mu$  and its confidence level (CL) calculated. Linear interpolation between the generated points is used to determine the 68% confidence interval; the two values of  $\mu$  (red horizontal lines) which cross the curve at  $CL = 0.68$  (vertical red line). . . . . 87
- 4.6. Left: One-dimensional scan of  $q_{m_x}$  for the  $H \rightarrow \gamma\gamma$ ,  $H \rightarrow ZZ$  channels and their combination. For the combination, the relative signal strengths between the channels are allowed to float. The 68% and 95% confidence intervals for  $m_X$  are determined as the values at which the curves cross the horizontal red lines. Right: 68% confidence contours in  $m_X$  and  $\sigma/\sigma_{SM}$  for the  $H \rightarrow \gamma\gamma$  and  $H \rightarrow ZZ$  channels and their combination. For this combination, the relative signal strengths of the channels are kept fixed to the SM expectation. . . . . 88

---

4.7. 68% confidence intervals for $\mu = \sigma/\sigma_{SM}$ for individual channels or combination of sub-channels determined by scanning the likelihood (left) and the Feldman-Cousins procedure (right). The value of $\sigma/\sigma_{SM}$ denotes the production cross-section times the relevant branching fraction for a given channel, relative to the SM. The green band indicates the 68% confidence interval on $\sigma/\sigma_{SM}$ for all channels combined. The intervals are determined at the best fit mass, $m_H = 125.8$ GeV. . . . .	90
4.8. 68% confidence contours for the production cross-section in $ggH$ and $ttH$ modes ( $\mu_{ggH+ttH}$ ), and $VH$ and $qqH$ modes ( $\mu_{VH+qqH}$ ), relative to the SM determined by scanning the likelihood (left) and the Feldman-Cousins procedure (right). Each colour indicates the result by combining all sub-channels in a particular decay mode. The yellow diamond at (1, 1) indicates the SM values. The contours are determined at the best fit mass, $m_H = 125.8$ GeV. . . . .	91
4.9. The 68% confidence contours extracted from data in the individual decay channels (coloured regions) and the full combination (solid line). The yellow square shows the SM value, while the fermiophobic and background only scenarios are indicated by the pink dot and red diamond respectively. . . . .	93
B.1. Total number of iterations in binning optimization scan as a function of the broad step size $P$ . The curve is shown for different numbers of final BDT boundaries. The minimum always occurs at the same value of $P$ as indicated by the green vertical line. . . . .	100
B.2. Increase in expected significance in the presence of a SM Higgs as the number of final BDT output bins is increased. The three curves show the improvement for different numbers of initial bins, $B$ . The red curve is representative of the result obtained from performing the optimization procedure in the 2011 analysis. . . . .	101
B.3. Systematic uncertainties on the $ggH$ signal model. The effect of $\pm 3\sigma$ variations derived in MC are shown with red dashed lines while the interpolated $\pm 3\sigma$ are shown with blue. . . . .	103

## List of Tables

3.1. Background MC used throughout the analysis with production cross-sections and corresponding equivalent integrated luminosity. . . . .	23
3.2. Signal efficiency for the preselection measured in data and MC using tag-and-probe in $Z \rightarrow e^+e^-$ 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. . . . .	30
3.3. Dijet selection criteria for the two identified jets to be considered likely associated to $qqH$ production. The leading and subleading $E_T$ jets are denoted $j^1$ and $j^2$ respectively. . . . .	35
3.4. 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. . . . .	54
3.5. Comparison of expected median upper limit and quantiles obtained using the asymptotic calculation of $CL_s$ and toys. The error quoted in the toys column is the statistical uncertainty from only generating 1000 toys at each value of $\mu$ . The comparison is made at three mass hypotheses in the range 120 to 140 GeV. . . . .	62
4.1. A realistic counting experiment across several channels. The number of observed events and that expected from signal and background processes are given per channel. Several sources of systematic are included which effect the expected rate of each signal or background process. Where a dash is entered, the systematic uncertainty has no effect on that process or channel. . . . .	77

---

4.2. Summary of analyses included in the ICHEP 2012 combination [12]. The column for $H$ prod indicates the production process targetted by the sub-channel. A label “un-tagged” indicates that the main contribution is from the $ggH$ production process. The final states for each channel are exclusive (no events lie in more than one sub-channel). The notations used here are: $jj$ indicating a dijet pair whether from a $W, Z$ decay or being consistent the vector-boson fusion process; $j_b$ denotes a jet which is identified as a $b$ -jet; $l$ is either a muon ( $\mu$ ) or electron ( $e$ ); OF and SF are dilepton pairs with opposite flavour ( $e\mu$ ) and same flavour ( $ee$ or $\mu\mu$ ) respectively. . . . .	82
4.3. Boson and fermion vertex scaling as a function of $\kappa_V$ and $\kappa_f$ for each production/decay included in the combination. Each cell represents the scaling factor applied to the production (row) decay (column) combination.	92
B.1. Additional energy resolution included in the $H \rightarrow \gamma\gamma$ signal model measured from comparison of $Z \rightarrow e^+e^-$ data and MC. The label “NOT GAP” indicates superclusters whose seed crystal is located more than 5 crystals away from an ECAL module boundary whereas the label “GAP” indicates superclusters for whose seed crystal is within 5 crystals of an ECAL module boundary [26]. . . . .	96
B.2. Relative energy scale difference in data and MC ( $\Delta P$ ) in the ECAL barrel, measured in $Z \rightarrow e^+e^-$ data. The first uncertainty given is statistical while the second is the systematic assigned to cover the difference in the $r_9$ distributions between electrons and photons [26]. . . . .	97
B.3. Relative energy scale difference in data and MC ( $\Delta P$ ) in the ECAL endcaps, measured in $Z \rightarrow e^+e^-$ data. The first uncertainty given is statistical while the second is the systematic assigned to cover the difference in the $r_9$ distributions between electrons and photons [26]. . . . .	98