Chapter 5

Signal and background modelling

The statistical interpretation of the data requires models of the expected $m_{\gamma\gamma}$ distribution for the signal and background processes in each category. These models must account for the sources of systematic uncertainty which affect the analysis. For the signal processes, the $m_{\gamma\gamma}$ distribution takes the form of a resonant peak around m_{H}, the width of which is entirely dominated by detector resolution. The construction of the signal model is described in Section 5.1. By contrast, the $m_{\gamma\gamma}$ distribution for background events is a falling non-resonant continuum. The method by which the background model is produced from data is described in Section 5.2. Finally, the handling of the systematic uncertainties is described in Section 5.3.

5.1 Signal modelling

5.1.1 Parametrisation of the signal $m_{\gamma\gamma}$ distributions

The shape of the $m_{\gamma\gamma}$ distribution is parametrised separately for each event category and production process, where the VH events are further split into WH and ZH samples. Since the vertex choice affects the shape of the $m_{\gamma\gamma}$ distribution, as was mentioned in Section 3.4, the modelling is performed separately for cases where the right vertex (RV) or wrong vertex (WV) was selected.

The signal $m_{\gamma\gamma}$ distributions are all parametrised using a DCB function [58] summed with an additional Gaussian function sharing the same mean. We refer to this functional form hereafter as a DCB+1G function. The DCB shape consists of a Gaussian function core and two tails which are described by power laws. The Gaussian core models the

width of the distribution, while the power laws cover the extended non-Gaussian tails caused by systematic under- or over-estimation of the value of $m_{\gamma\gamma}$. The DCB shape has five parameters: the width and mean of the Gaussian core, the cross-over points to the power laws on each side, and the order of the two power laws. The DCB+1G function therefore has a total of seven parameters. This is in contrast with the functional form used for signal modelling in previous studies of $H \to \gamma\gamma$ at CMS [63–65], which was a sum of up to five Gaussian functions, or a total of up to fourteen parameters.

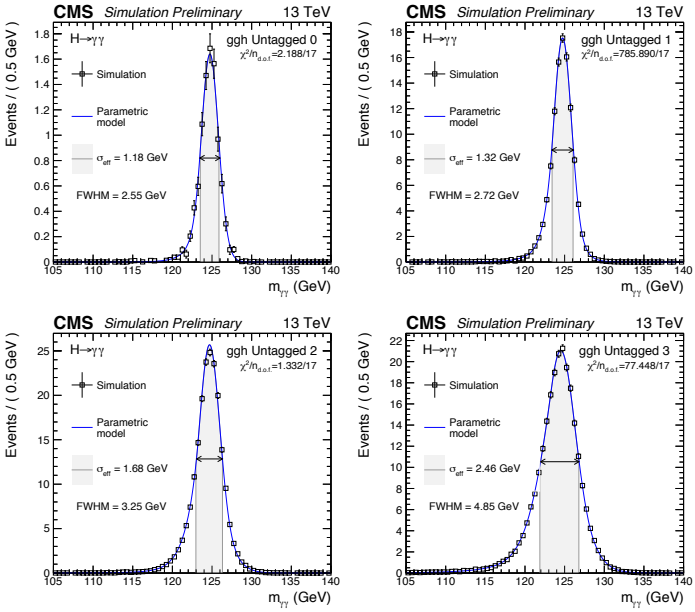
The values of the parameters of the functional form are determined using a NLL minimisation to fit the simulated $m_{\gamma\gamma}$ distributions. The choice of the DCB+1G function is motivated by the fact that it gives closer agreement than the simple sum of Gaussian functions, with substantially fewer degrees of freedom. The DCB shape is also better motivated physically than an arbitrary sum of Gaussian functions.

The parametrisations of the $m_{\gamma\gamma}$ distributions for the ggH process (with $m_{\mathrm{H}} = 125\,\mathrm{GeV}$) in the inclusive categories are shown in Figure 5.1 for both the DCB+1G and sum of Gaussians functional forms. The improvement in the agreement between the distribution and the parametrisation can be quantified by comparing the χ^2 values and the number of degrees of freedom. The DCB+1G typically gives a better agreement and this is also true for the other processes and categories.

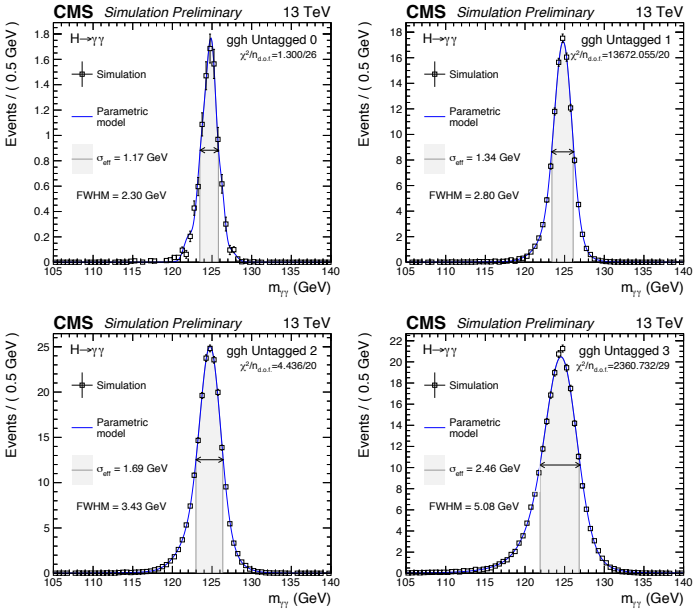
5.1.2 Dependence of model on m_{H}

Since the mass of the Higgs boson is not exactly known, the parametrisations from simulated signal samples assuming different m_{H} values are combined to form a single parametric model. There are two methods to perform this combination. The first option is to perform the fitting procedure described above for each m_{H} case. The individual parameters of the functional form can then be linearly interpolated from one m_{H} case to the next to produce the parametric model. This is the approach taken in [63–65].

The second option, which is used in this analysis, instead performs a simultaneous fit of all the different m_{H} samples, where the individual parameters of the functional form are themselves polynomials of m_{H}. The floating parameters in the fit are then the coefficients of these polynomials. The advantage of this method, which is referred to as simultaneous signal fitting (SSF), is a reduction in the total number of parameters used to determine the full model. Furthermore, the SSF method guarantees a sensible parametric model, whereas the linear interpolation method can lead to discontinuous



(a) Functional form: DCB+1G



(b) Functional form: Sum of Gaussians

Figure 5.1: The shape of the simulated $m_{\gamma\gamma}$ distribution (m_{H} =125 GeV) for the ggH process for the inclusive categories when parametrised with (a) DCB+1G and (b) sum of Gaussians, where the RV and WV contributions have been summed according to their relative yield. The plots show the agreement between the simulation and the parametrisation expressed as the χ^2, alongside the total number of degrees of freedom in the parametrisation.

or unphysical models since the individual m_{H} cases are parametrised separately and therefore are not guaranteed to be consistent.

The SSF method is applied separately for each process, category and RV or WV case, using seven different m_{H} cases: 120, 123, 124,125, 126, 127 and 130 GeV. The description of the parameters of the DCB+1G function uses polynomials of order 1. Polynomials of order 0, 1 and 2 were all tested, but it was found that there was no substantial improvement in the agreement when using an order greater than 1. This was determined to be because the floating parameters were not sufficiently constrained by the simulated $m_{\gamma\gamma}$ distributions to bring any meaningful improvement.

The signal models for the RV and WV contributions are combined according to their relative event content. The fraction of events where the selected vertex was within 1 cm in the z-direction from the true vertex is evaluated for each m_{H} sample, and then parametrised as a first order polynomial to get a smooth dependence on m_{H}.

The dependence of the full parametric signal model on m_{H} for the ggH process in each of the analysis categories is shown in Figure 5.2.

5.1.3 Normalisation of signal models

The signal models are normalised to match the number of event predicted by the SM after detector acceptance (A) and selection efficiency (ϵ) are taken into account. The expected number of events in each sample (ignoring A and ϵ) is first computed as the product of the relevant process cross-section, the $\mathrm{H} \to \gamma\gamma$ branching fraction and the integrated luminosity of the data sample. The cross-sections and branching fraction, as well as their dependence on m_{H}, are obtained from the LHCHXSWG recommendation [24]. After the simulated samples have been reconstructed and the events selected and categorised, the number of remaining events is related to the expected number by $\epsilon \times A$, which is parametrised in m_{H} using a polynomial fit. The final normalisation of the signal models for an arbitrary value of m_{H} can then be obtained from the parametrised $\epsilon \times A$ multiplied by the total number of events exected by the SM. Figure 5.3 shows the overall $\epsilon \times A$ for all categories combined as a function of m_{H}.

The normalised signal models for each process are summed together to obtain the expected signal $m_{\gamma\gamma}$ distribution for a given analysis category or for all categories combined, as shown in Figures 5.4, 5.5 and 5.6.

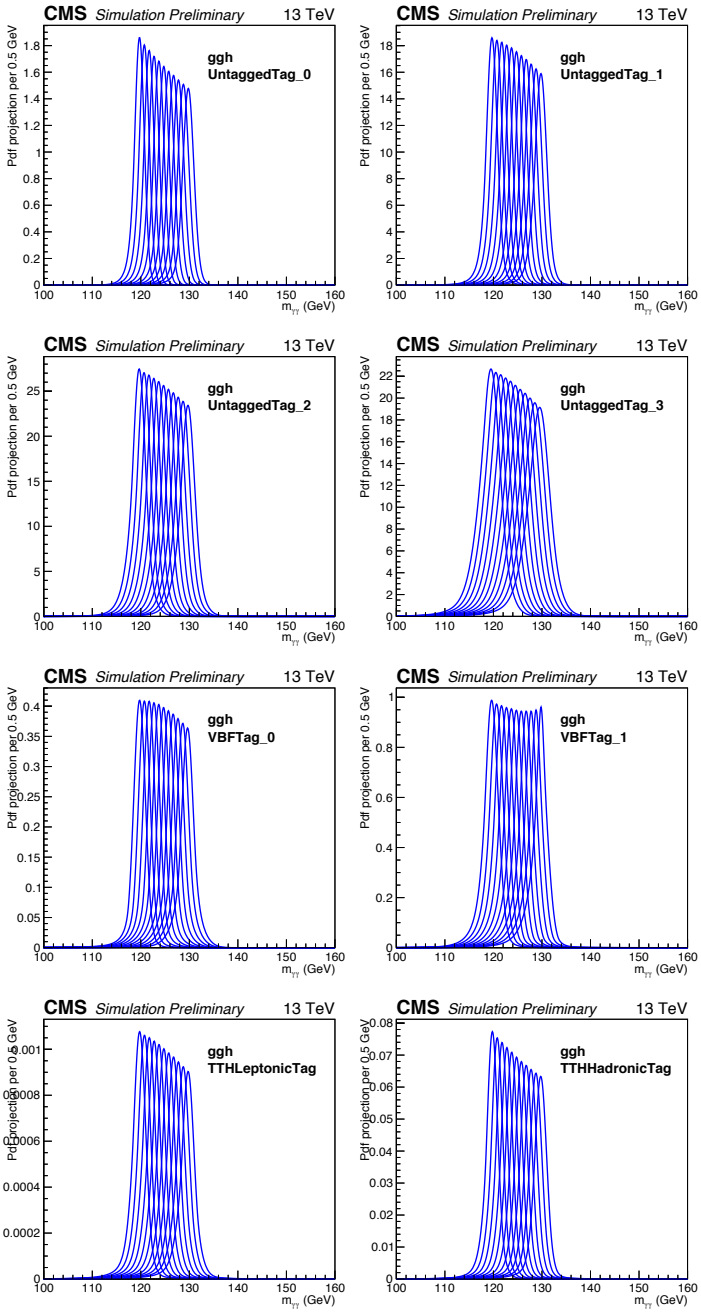


Figure 5.2: The m_{H}-dependence of the signal models for the ggH process for each of the analysis categories is shown. Each curve shows the signal model for a given value of m_{H}. The contributions from the RV and WV components of each model were summed together according to their relative size, linearly interpolated between the samples for different m_{H}.

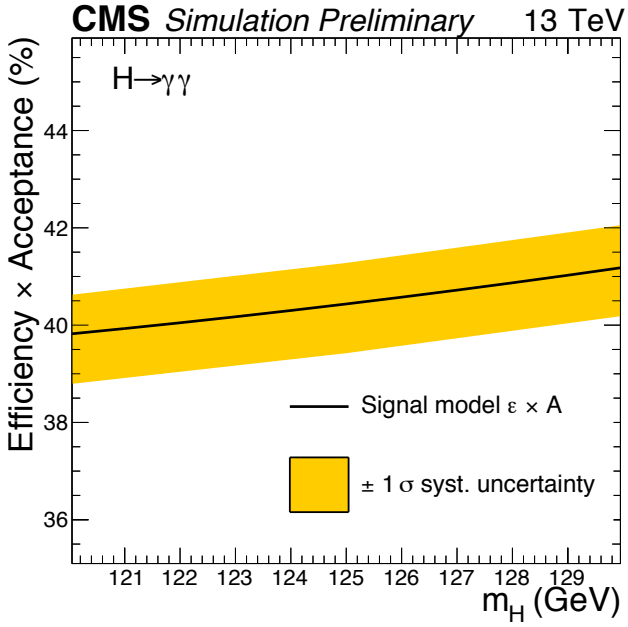


Figure 5.3: The $\epsilon \times A$ of all categories combined shown as a function of m_{H}. The ornage band shows the effect of the systematic uncertainties associated with trigger efficiency, photon identification and selection, photon energy scale and resolution and vertex identification.

5.2 Background modelling

5.2.1 The discrete profiling method

The background model is produced from data using the discrete profiling method [66], which treats the choice of functional form to describe the background as a discrete nuisance parameter in the final NLL fit to the data.

The underlying functional form of a given background distribution is typically not known. In some cases, several different families of function could in principle be chosen to parametrise the background distribution, all giving acceptable agreement with the data. Even within a given family of functions, it is not always clear which order to choose. Furthermore, the uncertainty associated with making a particular choice must be accounted for. The discrete profiling method avoids an arbitrary choice and provides a natural way to handle the uncertainty due to the background parametrisation.

When making a measurement of a parameter of interest using a NLL minimisation, nuisance parameters representing systematic uncertainties are typically profiled: they

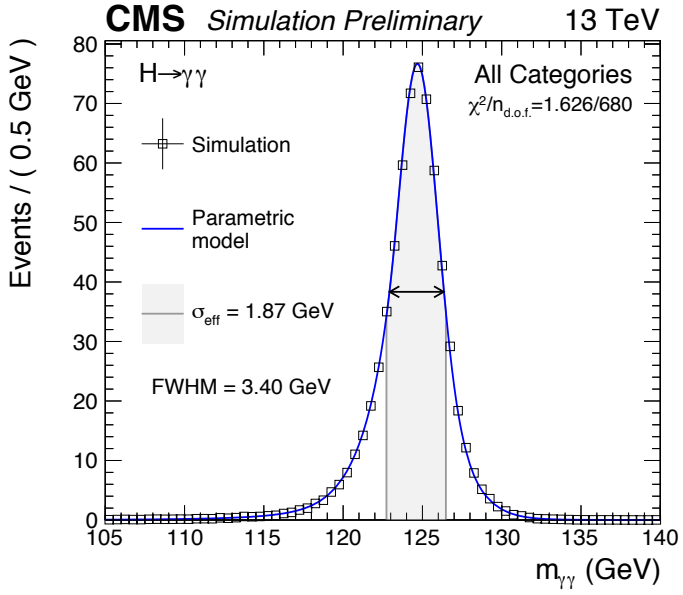


Figure 5.4: The signal model for all analysis categories combined for $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, obtained by summing the contributions from each production process according to the $\epsilon \times A$. The σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) and the FWHM (the width of the distribution at half of the maximum value) are also shown.

are allowed to float during the minimisation, but their final value is not of interest. The additional freedom produces a wider NLL curve, representing the additional uncertainty attributed to the nuisances. If the value of one of the nuisance parameters is instead fixed at the best-fit value, the width of the resulting NLL curve will be narrower but still with its minimum at the same place as the full profiled NLL curve. This width represents the uncertainty of the measurement without the effect of the fixed nuisance parameter. If the procedure is repeated for different fixed values of the nuisance parameter, different NLL curves, not necessarily at the minimum, will be produced. In the limit that many different values of the fixed nuisance parameter are sampled, the minimum envelope of all the fixed-nuisance NLL curves will converge to the full profiled NLL curve. The uncertainty can then be obtained from the envelope as for a usual NLL curve. This property is true also for discrete nuisance parameters. This procedure is illustrated in Figure 5.7. When parametrising the background distribution, since the chosen functional form itself is not of interest, it can be treated as a discrete nuisance parameter. The method described above can then be used to account for the uncertainty associated with the parametrisation of the background. The fact that different functional forms can have

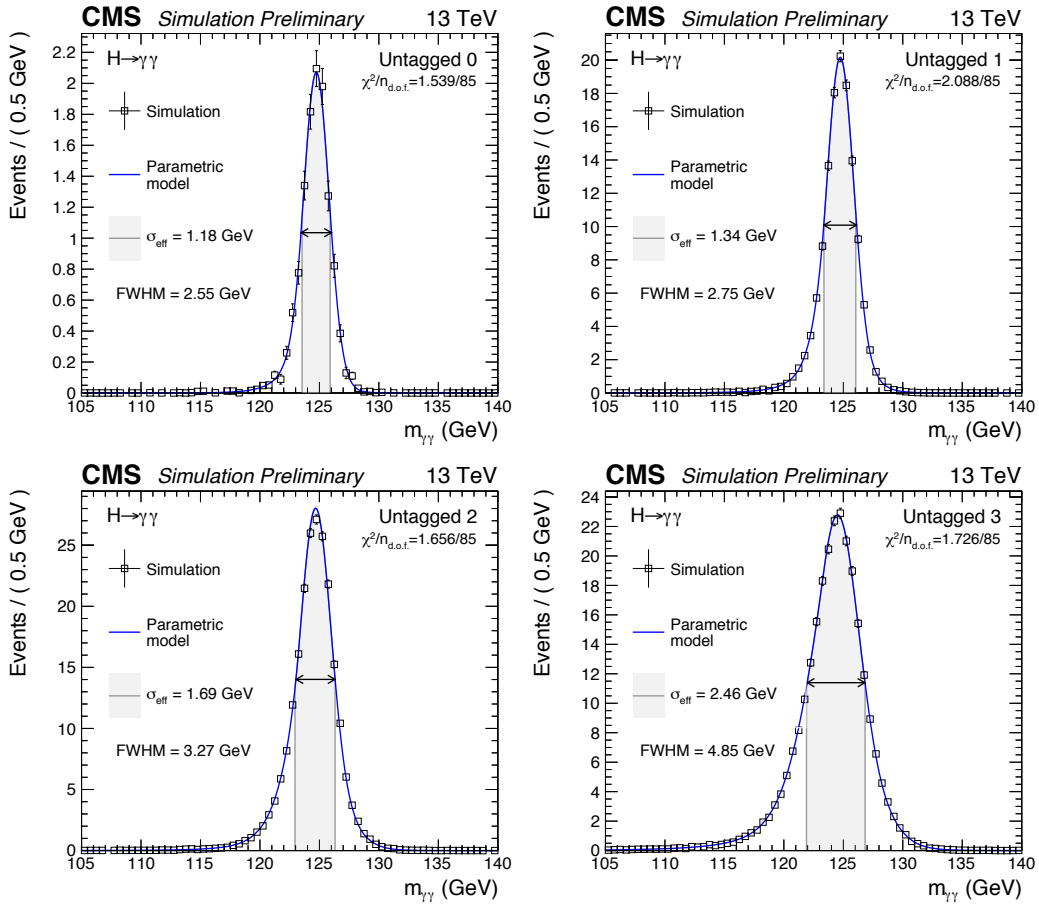


Figure 5.5: The signal models for the inclusive analysis categories for $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, obtained by summing the contributions from each production process according to their $\epsilon \times A$. The σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) and the FWHM (the width of the distribution at half of the maximum value) are also shown.

different numbers of degrees of freedom is taken into account by adding a penalty term to the NLL proportional to the number of parameters in the functional form.

A complete set of all analytic functions should be considered to obtain an exact result using the discrete profiling method. In practice, it is only necessary to include the subset of all analytic function which give a good description of the data. In this analysis, the following four families of functions are considered:

- Sums of exponentials:

$$f_N(x) = \sum_{i=1}^{N} p_{2i} e^{p_{2i+1} x},$$

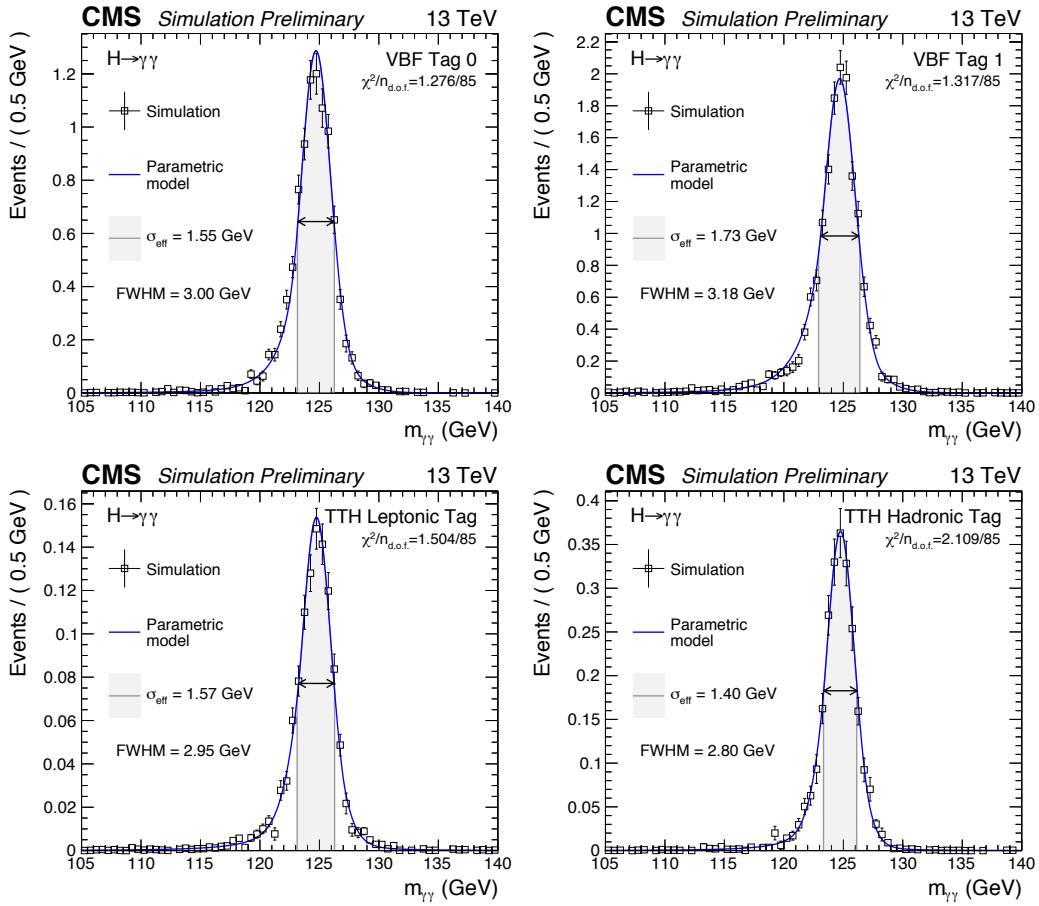


Figure 5.6: The signal models for the VBF-tagged and ttH-tagged analysis categories for $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, obtained by summing the contributions from each production process according to their $\epsilon \times A$. The σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) and the FWHM (the width of the distribution at half of the maximum value) are also shown.

- Sums of polynomials (in the Bernstein basis):

$$f_N(x) = \sum_{i=0}^{N} p_i b_{(i,N)}, \text{ where } b_{(i,N)} := \binom{N}{i} x^i (1-x)^{N-i},$$

- Laurent series:

$$f_N(x) = \sum_{i=1}^{N} p_i x^{-4+\sum_{j=1}^{i}(-1)^j(j-1)},$$

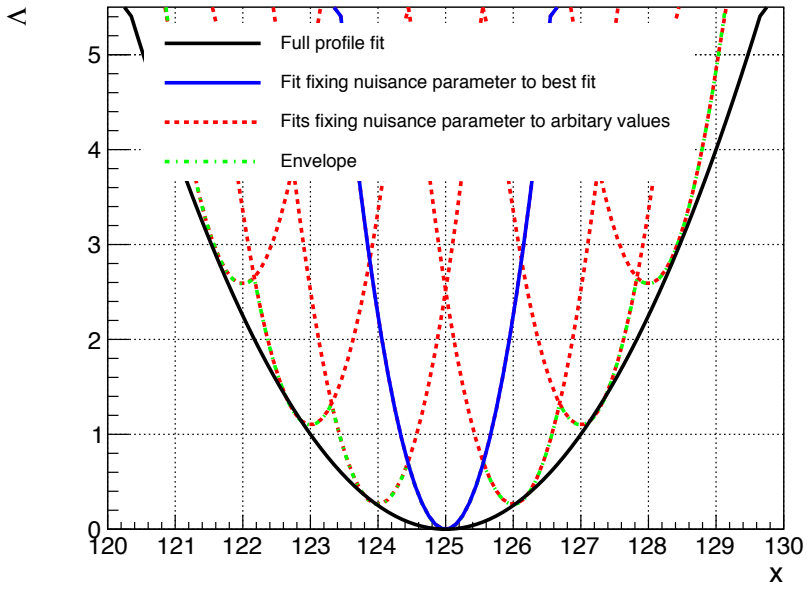


Figure 5.7: An illustration of the construction of the envelope to estimate the effect of a nuisance parameter. The NLL (denoted as Λ) curve obtained when performing a likelihood scan of parameter of interest x if the nuisance parameter is profiled is shown in black. The NLL curve obtained by fixing the nuisance to the best-fit value is shown in blue. The NLL curves for various fixed values of the nuisance other than the best-fit are shown in red. The minimum envelope of these curves, shown in green, approximates the original NLL curve obtained by profiling the nuisance parameter [66].

- Sums of power-law functions:

$$f_N(x) = \sum_{i=1}^{N} p_{2i} x^{-p_{2i+1}},$$

where for all k, the p_k are a set of parameters, and N represents the order of a particular function in the family.

The maximum order of the candidate function from each family is obtained separately for each analysis category using the following procedure. Starting with the lowest-order function in the family, the parameters of the candidate function are varied to minimize the NLL with respect to the $m_{\gamma\gamma}$ distribution. A penalty of 0.5 times the number of parameters in the functional form is added to the value of the NLL to account for differences in the number of degrees of freedom. The same procedure is applied to the function of next-highest order in the family. In the limit of large sample

size, the difference in the NLL between functions of successive orders N and $N+1$, $2\Delta NLL_{N+1} = 2(NLL_{N+1} - NLL_N)$, is distributed as a χ^2 with M degrees of freedom where, M is the difference in the number of free parameters in the order-$(N+1)$ and order-N functions. A p-value is then calculated as:

$$\text{p-value} = \text{prob}(2\Delta NLL > 2\Delta NLL_{N+1}|\chi^2(M)).$$

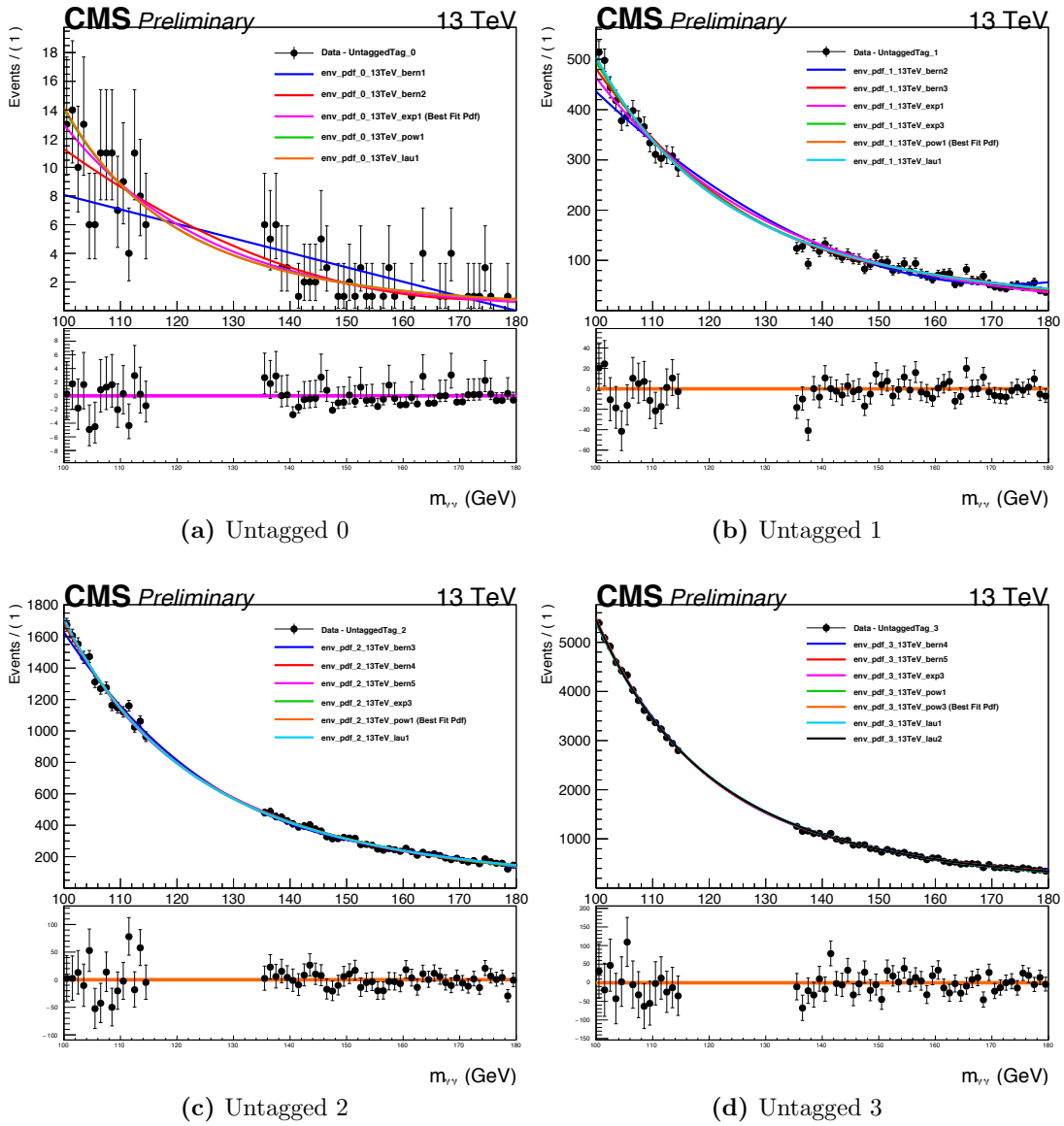
If the p-value is less than a predetermined threshold, chosen as 0.05, the higher-order function is supported by the data, otherwise the higher-order function is assumed too flexible given the data. In the former case, the next-highest order function is then considered, an so on. Otherwise, the procedure terminates having found the highest-order suitable function. An additional constraint is applied to remove low order functions which do not fit the data well. The remaining functions from each of the four families are added to the final set of candidate functions to be used in the discrete profiling, which are shown for each category in Figures 5.8 and 5.9.

A series of tests demonstrated that this method provides good coverage of the uncertainty associated with the choice of the function and provides an unbiased estimate of the signal strength. These tests are described in detail in [66].

The expected number of background events per GeV around 125 GeV is shown in Table 5.1 alongside the expected number of signal events for each category, which is broken down by the contribution of each production mode. The σ_{eff} (the width of the smallest widnow containing 68.3% of the distribution) and σ_{HM} (the width of the distribution at half of the maximum value, divided by 2.35) are also shown. The expected number of background events is quoted for a $\pm 1\,\sigma_{\text{eff}}$ window around 125 GeV, using the best-fit background function.

5.3 Systematic uncertainties

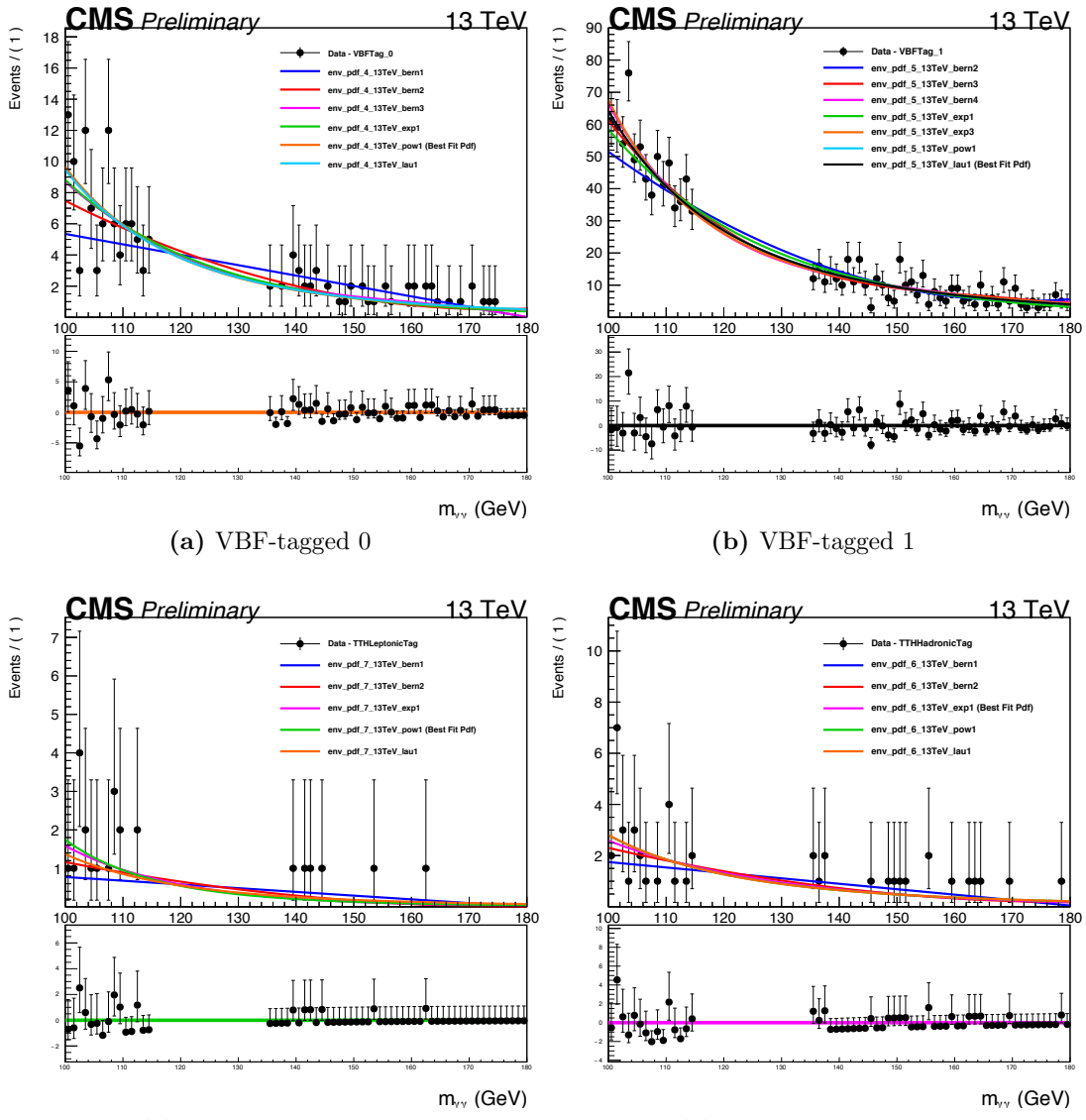
The systematic uncertainty impacting the background model is handled directly by the discrete profiling method described above. The uncertainties which affect the signal model are more numerous, and are implemented differently depending on their effect on the signal model.



(a) Untagged 0 **(b)** Untagged 1

(c) Untagged 2 **(d)** Untagged 3

Figure 5.8: The set of candidate functions chosen to parametrise the background using the discrete profiling method in the inclusive categories.



(a) VBF-tagged 0 **(b)** VBF-tagged 1

(c) leptonic ttH-tagged **(d)** hadronic ttH-tagged

Figure 5.9: The set of candidate functions chosen to parametrise the background using the discrete profiling method in the VBF-tagged and ttH-tagged categories.

Event Categories	SM 125GeV Higgs boson expected signal								Bkg
	Total	ggh	vbf	wh	zh	tth	σ_{eff}	σ_{HM}	(GeV^{-1})
Untagged Tag 0	11.92	79.10 %	7.60 %	7.11 %	3.59 %	2.60 %	1.18	1.03	4.98
Untagged Tag 1	128.78	85.98 %	7.38 %	3.70 %	2.12 %	0.82 %	1.35	1.20	199.14
Untagged Tag 2	220.12	91.11 %	5.01 %	2.18 %	1.23 %	0.47 %	1.70	1.47	670.44
Untagged Tag 3	258.50	92.35 %	4.23 %	1.89 %	1.06 %	0.47 %	2.44	2.17	1861.23
VBF Tag 0	9.35	29.47 %	69.97 %	0.29 %	0.07 %	0.20 %	1.60	1.33	3.09
VBF Tag 1	15.55	44.91 %	53.50 %	0.86 %	0.38 %	0.35 %	1.71	1.40	22.22
TTH Hadronic Tag	2.42	16.78 %	1.28 %	2.52 %	2.39 %	77.02 %	1.39	1.21	1.12
TTH Leptonic Tag	1.12	1.09 %	0.08 %	2.43 %	1.06 %	95.34 %	1.61	1.35	0.42
Total	647.77	87.93 %	7.29 %	2.40 %	1.35 %	1.03 %	1.88	1.52	2762.65

Table 5.1: The expected number of signal and background events per category. The σ_{eff} of the signal model is also provided as an estimate of the $m_{\gamma\gamma}$ resolution in that category. The expected number of background events is quoted per GeV in a $\pm 1\,\sigma_{\text{eff}}$ window around 125 GeV.

Systematic uncertainties which affect the shape of the $m_{\gamma\gamma}$ distribution are built directly into the signal models described in Section 5.1. Most systematics of this type impact individual photon energies, and therefore the invariant mass. The effect of the systematic variation is propagated to the mean, σ_{eff} and normalisation of the signal $m_{\gamma\gamma}$ distribution for each category and signal process. Corresponding nuisance parameters are then inserted which can modify the normalisation, mean and width of the DCB+1G parametrisations. An exception is the nuisance corresponding to the vertex efficiency, which instead modifies the relative mixing fraction of the RV and WV components of the model. The nuisances parameters are then Gaussian-constrained and allowed to be profiled in the NLL minimisation when making measurements of the parameters of interest. Nuisances of this type are referred to hereafter as *shape nuisances*.

Uncertainties on selection efficiencies typically do not affect the shape of the $m_{\gamma\gamma}$ distribution, but do change the final event count (or *yield*) of the signal model for each category. Uncertainties of this type are therefore implemented as nuisance parameters which impact the yield. Depending on the source of the systematic uncertainty, categories can be scaled differently, although all categories will simultaneously either increase or decrease. The corresponding nuisance parameters are profiled in the NLL minimisation with a log-normal [67] constraint, which cannot give rise to negative yields. Systematics of this type are referred to as *yield nuisances*, which are either symmetric or asymmetric. In the symmetric case, the upward and downward variations of the uncertainty are the same size. In the asymmetric case, the upward and downward variations are of different sizes.

Finally, certain sources of systematic uncertainty affect the categorisation of events. In this case, the systematic variations can cause events to move from one event category to another, or altogether out of the acceptance of the analysis. These systematic uncertainties are implemented as nuisance parameters which change the relative yield of event categories, and are referred to as *category migration nuisances*. They are implemented analogously to yield nuisances, except that the yield of certain categories will increase while the yield of others must decrease in turn. Furthermore, separate nuisances are implemented to account for different types of migration: for instance, between the individual VBF-tagged categories on the one hand, and then between all VBF-tagged and all Untagged categories on the other.

5.3.1 Theory uncertainties

Gluon fusion contamination of VBF-tagged and ttH-tagged categories

The theoretical prediction for the jet multiplicity in gluon fusion events is unreliable in cases with large number of jets. This leads to the introduction of an uncertainty on the contamination of ggH events in other analysis categories which use jets in their selections.

For VBF-tagged categories, the uncertainty is estimated using the Stewart-Tackmann procedure [68]. This results in a set of category migration nuisances: a migration between VBF-tagged categories (at most 39% on category yield) and a migration between Untagged and VBF-tagged categories (at most 10% on category yield);

For the ttH-tagged categories, this uncertainty is modelled by three yield nuisances:

- shower modelling uncertainty, which is estimated by comparing the jet multiplicity in data and simulation for $t\bar{t} \rightarrow$ jets events, where the tops quarks both decay to leptons, leading to variations of the order of 45% on the yield of the ttH-tagged categories;
- gluon splitting modelling, which is estimated from the ratio of cross-sections of ttbb and $tt + 2$ jet events in 13 TeV data;
- an additional nuisance included due to the small size of the simulated samples used in these studies.

$H \to \gamma\gamma$ branching ratio

The uncertainty on the $H \to \gamma\gamma$ branching fraction is taken directly from the LHCHXSWG recommendation [24]. It is implemented as a yield nuisance affecting all analysis categories, and the variation is of 2.08% on the category yield.

Parton Distribution Functions

The uncertainty on signal process cross-sections due to parton distribution function (PDF) uncertainties produces both an overall change in signal yield and category migrations.

One symmetric yield nuisance for each signal process is included, accounting for both PDF and the strong force coupling constant (α_s) uncertainties. Each varies the category yields according to the corresponding uncertainty in the process cross-section, as provided by the LHCHXSWG recommendation [24]. Namely, the ggH, VBF, WH, ZH, and ttH process cross-sections are scaled by 3.2%, 2.1%, 1.9%, 1.6% and 3.6% respectively.

The relative yield change is modelled using a series of category migration nuisances. The size of the variations is determined according to the PDF4LHC prescription [69], by re-weighting individual events according to the NNPDF30 PDF set [70]. The effect of the variations is normalised by their effect on the overall yield, such that they represent only migrations. The procedure results in 60 uncorrelated symmetric migration nuisances. The largest migrations are of the order of 2% of category yields, but are typically below 0.4%.

QCD scale

The uncertainty on the scale of the QCD interaction is parametrised in terms of the renormalisation ($\mu_{Renorm.}$) and factorisation ($\mu_{Fact.}$) scales. The asymmetric yield nuisances corresponding to variations of these parameters are taken direction from the LHCHXSWG recommendations for cross-sections [24]. The size of the effect for the ggH, VBF, WH, ZH and ttH processes is $+4.6\%/-6.7\%$, $+0.4\%/-0.3\%$, $+0.5\%/-0.7\%$, $+3.8\%/-3.0\%$ and $+5.8\%/-9.2\%$ respectively. Three additional asymmetric category migration nuisances are also included. The size of their effect is estimated at generator level, by varying the values of $\mu_{Renorm.}$ and $\mu_{Fact.}$ by factors of 2 (upward variation) or 0.5 (downward variation). The three category migration nuisances correspond to varying $\mu_{Renorm.}$ up or down while keeping $\mu_{Fact.}$ constant ; varying $\mu_{Fact.}$ up or down while

keeping $\mu_{Renorm.}$ constant ; and varying both $\mu_{Renorm.}$ and $\mu_{Fact.}$ up or down uniformly. In each case the migrations are found to be of the order of 5-10% of the category yield.

Strong force coupling constant

The effect of the uncertainty on the value of α_s is modelled as described in the PDF uncertainty section. The yield nuisances for the PDF uncertainties also incorporate the uncertainty on α_s. Three additional asymmetric category migration nuisances are included, the effect of which is determined according to the PDF4LHC prescription [69] normalised by the overall effect on the signal yield. The largest migrations are of the order of 3.7% of the category yield.

Underlying event and parton shower

The uncertainty on the *underlying event* is introduced because different models exist to describe the interactions of quarks and gluons in p-p collisions, each giving different predictions. In particular, this affects the modelling of jets. This uncertainty is taken into account as a set of category migration nuisances. The sizes of the variations are obtained from dedicated simulated samples where the parameters of the generators have been tuned differently. The category migrations evaluated are between: VBF-tagged 0 and VBF-tagged 1 categories (of the order of 7%); and all VBF-tagged categories and all Untagged categories (of the order of 9%).

The *parton shower* refers to the emission of QCD radiation from partons during p-p collisions. As for the underlying event, an uncertainty is introduced because different models and generators give different predictions, in particular in the reconstruction of jets. The resulting uncertainty is modelled analogously to the underlying event uncertainty. The category migrations evaluated are between: VBF-tagged 0 and VBF-tagged 1 categories (of the order of 7%); and all VBF-tagged categories and all Untagged categories (of the order of 9%).

5.3.2 Photon energy uncertainties

Modelling of detector response in GEANT 4

The imperfect modelling of electromagnetic showers in the detector simulation software GEANT 4 is modelled with a shape nuisance, as it impacts the photon energy scale. The size of the variation was determined with a dedicated simulated sample where the parameters of shower modelling were modified, leading to a small uncertainty on the photon energy scale. The size of the effect is approximately 0.05% on the mean of the $m_{\gamma\gamma}$ distribution in each category.

Modelling of the material budget

The imperfect modelling of the amount of material between the vertex and the ECAL affects the simulation of the photon and electron showers. The uncertainty related to this effect is estimated with dedicated samples there the amount of simulated material is uniformly varied by $\pm 5\%$. It is treated as two separate shape nuisances, for EB and EE photons separately, which affect the mean of the $m_{\gamma\gamma}$ distribution by at most 0.1% (0.03%) and the σ_{eff} by at most 3.5% (7.6%) for the EB (EE) photons.

Non-linearity of detector response

The uncertainty associated with the fact that the ECAL response is not linear is estimated by comparing boosted $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ decays in data and simulation. Individual photon energies are affected by up to 0.2%. The effect is propagated as a shape nuisance, which varies the mean of the $m_{\gamma\gamma}$ distribution by 0.1% in each category.

Non-uniformity of the light collection

The uncertainty on the response of the ECAL crystals depending on their position in η is modelled separately as shape nuisances for photons in the barrel and in the endcaps. The size of the uncertainty is 0.07% on the photon energies. The effect is propagated to the $m_{\gamma\gamma}$ distribution of each category and applied as separate shape nuisances in the EE and EB, which vary the mean by up to 0.2% and the σ_{eff} by up to 5%.

Per-photon energy resolution

The uncertainty of the per-photon energy resolution is conservatively evaluated by scaling the output of the $BDT_{\gamma E}$ described in Section 3.3.5 by $\pm 5\%$. This uncertainty is propagated throughout the analysis and modelled as a yield nuisance. The size of the variation is at most 3% depending on the analysis category.

Photon energy scale and resolution

After the calibration of the ECAL described in Section 2.2.3, there are still some discrepancies in the photon energy scale and resolution between simulation and data. Since electrons and photons are both reconstructed as SCs, these discrepancies can be studied using $Z \to e^+e^-$ events where the electrons are reconstructed as photons. For SCs in eight R_9 and $|\eta|$ classes, the invariant mass distributions in data and simulation are both fitted with a Breit-Wigner (which models the natural shape of the Z-peak) convoluted with a Crystal Ball function (CB) function (which describes the ECAL resolution and losses due to unrecovered energy from bremsstrahlung). The natural width and pole mass of the Z boson are fixed to their accepted values [1] in this parametrisation.

The corrections to the photon energy scale are given by the relative differences between the best-fit means of the CB in data and simulation, divided by the Z boson pole mass, in each bin. The corrections to the photon energy resolution are applied by adding additional smearing terms to the width of the CB in quadrature. The additional scale and resolution corrections described above each have uncertainties, which are related to choices made for the $Z \to e^+e^-$ event selection and classification, as well as the difference between the final electron and photon energy regression BDTs. The uncertainties on the energy scale and resolution are quantified for each photon class, and propagated to the $m_{\gamma\gamma}$ distribution in each analysis category. This results in a shape nuisance for the photon energy scale in four classes (high and low R_9, each for EB and EE), and eight shape nuisances for the photon energy smearing (as for the photon energy scale, but parametrised also as constant and stochastic contributions).

The size of the systematic uncertainties are of the order of 0.15% to 0.5% depending on the photon class. The effect on the mean of the $m_{\gamma\gamma}$ distribution is at most 0.25%, while the effect on the σ_{eff} is at most 20%, depending on the analysis category.

Preselection

The efficiency of the photon preselection is quantified using the tag-and-probe method described in Section 3.3.3, which also provides systematic uncertainties for different photon classes. The systematic uncertainties are propagated to a yield nuisance, the effect of which is around 4% on the category yields.

Shower shape corrections

The uncertainty deriving from the imperfect modelling of shower shape variables is estimated using simulated samples with and without the corrections. This effect is of order 0.064% on the photon energy scale, and is implemented as four shape nuisances for photons in different η and R_9 classes, which vary the $m_{\gamma\gamma}$ distribution mean by at most 0.2% and σ_{eff} by at most 1.8%.

5.3.3 Per-event uncertainties

Integrated luminosity

The uncertainty on the value of the integrated luminosity of the data sample is modelled as a scale nuisance, the size of which is 6.2% on the yield of all signal processes.

Jet energy scale and resolution

The uncertainties on the jet energy scale are described by migrations nuisances: between VBF-tagged 0 and VBF-tagged 1 (at most 13.6% on the category yield); between all VBF and Untagged categories (at most 12%) ; and between ttH-tagged and Untagged categories (at most 15%).

The nuisances for the jet energy resolution are treated analogously, with migrations of at most 2.2% between VBF-tagged 0 and VBF-tagged 1 categories; at most 2.9% between VBF and Untagged categories ; and at most 12.8% between ttH-tagged and Untagged categories.

Lepton reconstruction and b-tagging efficiencies

The uncertainty on the reconstruction of electrons and muons is determined by considering the ration of leptons reconstructed in data and simulation. The result is implemented as a yield uncertainty for the ttH-tagged categories of the order of 1%.

Similarly, the uncertainty on the tagging of b jets is evaluated by varying the ratio between the measured b-tagging efficiency in data and simulation within their uncertainty. This is propagated to a yield nuisance, which has an effect of the order of 2% for ttH-tagged categories.

Rejection of jets from pileup

The uncertainty on pileup jet rejection using the selection of the σ_{RMS} (as described in Section 3.5.3) is described by a migrations nuisances between all VBF and Untagged categories (at most 3.5%).

A further uncertainty, relating to the number of pileup jets in simulation which are reconstructed as genuine jets from the hard scatter, is described by migrations nuisances: between VBF-tagged 0 and VBF-tagged 1 (at most 1.2% on the category yield); and between all VBF and Untagged categories (at most 1.3%).

Trigger efficiency

The uncertainty on the trigger efficiency is estimated using a tag-and-probe method as described in Section 3.2.2. The uncertainty is applied as a yield nuisance parameter of size 0.1%.

Vertex-finding efficiency

The uncertainty on the vertex-finding efficiency is determined by comparing the fraction of correctly identified vertices in $\mathrm{Z} \to \mu^{-}\mu^{+}$ between data and simulation. The uncertainty is modelled as a shape nuisance which alters the RV/WV mixing fraction of each signal model. The variation is of the order of 1.5% on the RV fraction.