ATLAS PUB Note

ATLAS-PHYS-PUB-2020-028

10th December 2020

Recommendations for the Modeling of Smooth Backgrounds

The ATLAS Collaboration

In data analyses that exploit a distribution of the data, backgrounds are often modeled using a continuous description of the distribution shape. This technique is used in particular for situations involving narrow signal peaks and wide sidebands of regular backgrounds, for example in the study of Higgs boson decays or in searches for narrow resonances.

This note reviews the main techniques in use within ATLAS for smooth background modeling: closed-form functions, Gaussian processes and Functional decomposition. In all cases, the chosen model must provide a sufficiently accurate description of the background distributions. Systematic uncertainties should also be included to accounts for possible residual mismodeling effects.

Criteria used to select appropriate models and methods to define the corresponding modeling uncertainties are described, and recommendations applicable to various analysis configurations are provided.

[2020-12-10] Fixed typographical errors in the definition of the F-test and associated F-distribution.

ATL-PHYS-PUB-2020-028
10/12/2020

1 General considerations

In physics analyses, signal and background components can be separated by making use of differences in their distributions for a given observable, for instance an invariant mass. In these *shape analyses*, a fit to the data is performed to a model built from the probability distribution functions (PDF) of the observable(s) for each component. These PDF must accurately reproduce the true distribution in each case, in order for the fit results to be reliable.

In many cases, the PDF can be modeled using histograms, or *templates*, obtained from Monte-Carlo (MC) simulation or control region (CR) data. They can also be derived from the data using more complex procedures such as the "ABCD" method (see e.g. [1] for a recent ATLAS application). However in some cases the histogram description is insufficient:

- **Low template statistics**: The statistical uncertainties on the template yields, due to limited MC or CR event yields, must be propagated to the fit results as a systematic uncertainty, separate from the statistical uncertainty due to the limited size of the fitted dataset. If the size of the systematic is comparable to that of the statistical uncertainty, this can lead to a significant degradation of analysis sensitivity. To avoid this, templates must be typically constructed from a sample (either MC or a CR) that is at least several times larger than the analysis dataset, as discussed in Section 5. This can be technically difficult for analyses with large irreducible backgrounds, such as resonance searches in the dijet [2], dilepton [3], diphoton [4] and $t\bar{t}$ [5] mass spectra, or $H \to \gamma\gamma$ [6] and $H \to \mu\mu$ [7] measurements. In this case, using a smooth description instead of a histogram can help alleviate the effects of limited template statistics. The challenge is to provide a description that is as unbiased as possible, i.e. as close as possible to the shape that would have been obtained if infinite template statistics were available.
- **Poor knowledge of the background template shape**: in some cases neither MC nor any CR can be used to provide a sufficiently accurate description of the template shape. This can occur, again, in analyses with large irreducible backgrounds: the large background yields provide sensitivity even at low values of the signal over background (S/B) ratios, but only in cases for which the uncertainties on the background template shape are comparably small. These uncertainties include for instance theory uncertainties, or uncertainties in the extrapolation from a CR to the signal region (SR). In many cases, they can be included as systematic variations of the template shape [8]. However for $S/B \sim 1\%$ or smaller, as in the $H \to \gamma\gamma$ and $H \to \mu\mu$ analyses, uncertainties can be difficult to obtain a corresponding level of precision. A possible solution is then to use a smooth description of the background that is based on the available template, but also flexible enough to also describe a variety of similar shapes, among which the data shape can be found. This cannot be achieved with certainty since the data shape is a priori unknown, but one reasonably assume that this is achieved if the parameterization can describe well the nominal template shape as well as its known variations. The obtained description is then robust against possible differences between the true background shape and the data shape, at the cost of a decrease in signal sensitivity due to the added flexibility. The challenge in this case is to find a description that provides a low bias in the fit results (which requires sufficient flexibility in the shape), while minimizing the decrease in signal sensitivity (which is mitigated by using more rigid shapes).

The smooth model can be implemented in several possible ways; the ones commonly used within ATLAS are listed in Section 2, and criteria used to select an appropriate model for a given analysis are described in Section 3. All forms typically include free parameters, the values of which are obtained in a fit to data. For analyses making use of profile likelihood techniques, the profiling has the effect of automatically propagating these uncertainties to the measurement of the parameter(s) of interest. For other situations (*e.g.* cut-based analyses), the uncertainties in the model parameters must be propagated manually. In both cases, this only covers *within-model* uncertainty, obtained under the assumption that the model provides a perfect description of the data. These must be supplemented by *modeling* uncertainties, which account for a possible mismatch between the model and the data. The determination of these uncertainties is described in Section 4. Finally, the recommendations are summarized in Section 6.

2 Smooth modeling techniques

Smooth modeling techniques are generally applied in a setting where signal and background components are separated using the shape of a discriminating variable, often an invariant mass. The techniques are usually applied to one-dimensional distributions only; while they are also generally applicable to higher-dimensional problems, validating the resulting models becomes rapidly more complex as the number of dimensions increases and their use is therefore not widespread in high-energy physics.

The method relies generally on differences between the signal and background distributions of the observable. An important parameter is the observable range over which the analysis is performed. Wider ranges generally provide more information and therefore greater sensitivity, but at the expense of more difficulties for the modeling of the signal and background distributions. A general guideline is to include the widest possible observable range, while excluding problematic regions (for instance "turn-on" regions with rapid changes in the analysis efficiency, or regions near production thresholds) unless they present a physical interest that overrides the modeling difficulties.

A frequent specific situation is the case of a signal corresponding to a narrow resonance observed as a peaking distribution in the corresponding invariant mass, over a background component with variations on longer mass scales. In this case, the analysis range should generally be chosen as follows:

- The invariant mass range covered by the analysis should extend both below and above the signal peak, so that the background contribution below the peak can be reliable interpolated.
- The sidebands on either side should be wide enough so that the uncertainty associated with the subtraction of this background does not contribute significantly to the total uncertainty. In practice this depends on characteristics of the analysis such as the background level and the signal-over-background ratio.

In all cases, the modeling of the background component is often a critical ingredient, especially for low S/B where small changes in the background model leads to large in the signal fit results. Several techniques can be used to implement this modeling, and the ones mainly used within ATLAS are listed in the following sections. While the examples listed in this document are mainly focused on the narrow-resonance search scenario, they are generally applicable to other cases as well.

2.1 Closed-form function

This class of model represents the template shape using a PDF described as a simple function $f(m)$ of the observable m, expressed in closed form using usual mathematical functions.

This is also known as *functional form* modeling, since the model typically specifies only the form of the function, while the parameters can float freely in the fit. The forms most commonly used are:

- **Polynomials** : these can be defined as

$$f(m) = a_0 + a_1 m + a_2 m^2 + \ldots$$

  but any polynomial basis can be used. For instance *Bernstein polynomials*, which use $B_{k,n}(x) = x^k (1-x)^{n-k}$ $(1 \leq k \leq n)$ as basis monomials, have the property that any positive function can be approximated to arbitrary precision by polynomials with *positive* coefficients in this basis [9]. One can therefore enforce the positivity of the PDF by restricting all $a_i > 0$, without loss of generality. The variable x of the Bernstein polynomials is defined as $x = (m - m_{\min})/(m_{\max} - m_{\min})$ where $[m_{\min}; m_{\max}]$ is the fit range.

- **Power laws** are defined as

$$f(m) = a_0\, m^b$$

or more complex forms such as the PDF-inspired

$$f(m) = a_0\,(1-x^d)^c\,x^{b_0+b_1\log x+b_2\log^2 x+\ldots}, \tag{1}$$

for cases where m is an invariant mass, and $x = m/\sqrt{s}$ is related to the longitudinal momentum fraction used in the parameterization of parton distribution functions.

- **Exponentials** : descriptions generally include single exponentials, sums of several terms such as

$$f(m) = a_0\left(e^{-b_0 m} + a_1\,e^{-b_1 m} + \ldots\right)$$

or exponentials of higher-order polynomials,

$$f(m) = a_0\,e^{b_1 m + b_2 m^2 + \ldots}.$$

For all the above, it should be noted that in the *RooFit* framework [10], PDF normalization is handled "behind the scenes" by the framework itself and should therefore not be included in the PDF expression, which are always normalized to unity[1]. In the expressions above, the normalization terms (all denoted as a_0) are therefore always fixed to the appropriate value to achieve normalization to unity and are not associated with a free fit parameter[2].

An example of a smooth parameterization using a closed-form function is shown in Figure 1.

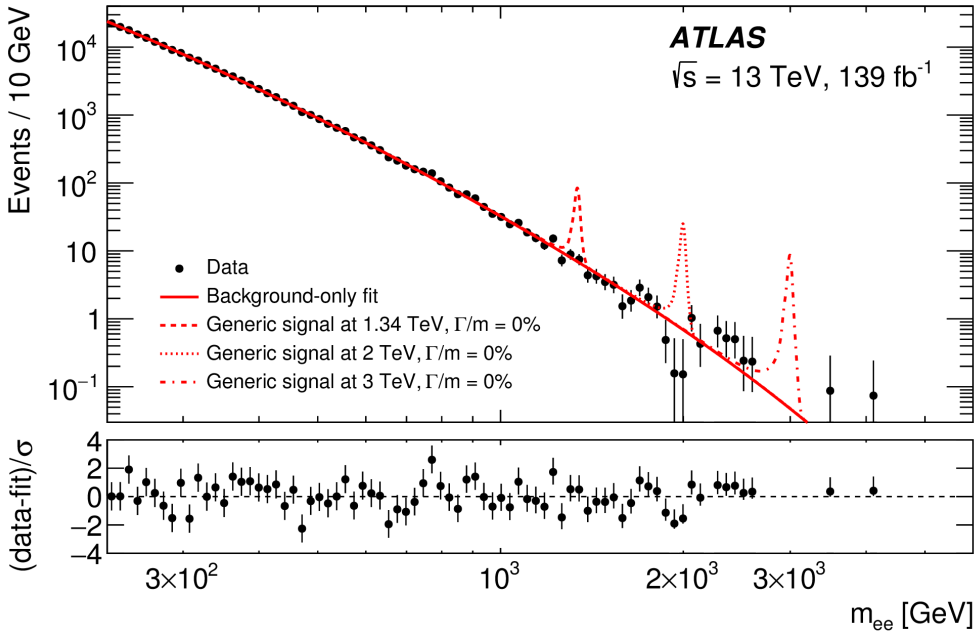

Figure 1: Example of a fit of a mass spectrum distribution to a closed-form function, taken from [3], using a function based on Equation (1) in the text.

These forms typically have several free parameters. There is no general principle specifying whether a parameter should be left free in the fits, or fixed to some value (as is done implicitly when *e.g.* using a given polynomial order). The amount of flexibility that is given to the function depends on a priori physics knowledge on the range of shapes that should accurately modeled. However in practice the functional form cannot usually be defined from

[1] For a fit with a signal and a background component, the total PDF will have the form $f(m) = N_S f_S(m) + N_B f_B(m)$ with explicit normalization terms N_S and N_B respectively for signal and background, and corresponding PDF f_S and f_B which are in both cases normalized to unity.

[2] These normalization parameters should however be included when counting the number of degrees of freedom associated with a given functional form.

first principles, and the choice of function is rather driven by technical considerations of simplicity of expression, least number of free parameters, fit stability and avoidance of numerical issues.

The method usually works well for simple background shapes, in particular smoothly falling mass spectra. More complex shapes, for instance involving efficiency thresholds, peaking backgrounds or other shape changes can also be accommodated, but this usually requires more complex expressions with more parameters, which can lead to excessively flexible shapes or to fit instabilities.

In general, the method also suffers from a degree of arbitrariness in the flexibility of the model: this depends in large part on the choice of functional form, which cannot be tuned in a continuous manner and may have a different effect on different parts of the spectrum. Some tuning can however be performed by selecting which parameters are left free.

***Pros*:**

- Simple to implement
- CPU-efficient: extremely quick to evaluate, suitable for MINUIT [11] fits with many iterations
- Generally good results for regular background shapes

***Cons*:**

- Difficult to implement for complex background shapes
- Limited ability to tune the flexibility in the parameterization (only through changes in the functional form or the number of parameters)
- Generally no physics or first-principle motivation for the chosen functional form

Sliding-window fits (SWiFT): The background model can be defined on the full search range, but one can also perform fits on smaller mass intervals. For narrow-resonance searches, the fit range can be defined as a *sliding window* that extends above and below the peak position for each resonance mass hypothesis. The reduction in the analysis range leads to easier modeling, and can help in cases where no suitable function is found to model the full search range. The narrower sidebands however also lead to weaker constraints on the background shape and thus lower signal sensitivity, although this effect is generally negligible for analyses with narrow signal peaks and wide sidebands.

An example is shown in Figure 2.

2.2 Gaussian process regression (GPR)

In GPR modeling [14], the template PDF is represented by a histogram and used to model binned data. The bin contents $\nu_1 \dots \nu_N$ that define the templates are constrained by the Gaussian PDF

$$G(\nu_i; r_i, C) = \frac{1}{\sqrt{(2\pi)^N |C|}} \exp\left[-\frac{1}{2}\sum_{ij=1}^{N}(r_i - \nu_i)C_{ij}^{-1}(r_j - \nu_j)\right], \tag{2}$$

with a covariance matrix C and a mean defined by the reference values $r_1 \dots r_N$. The elements of C can be obtained as $C_{ij} = K(m_i, m_j)$, where $m_1 \dots m_N$ are the bin centers, and the *kernel* function $K(m, m')$ is a smooth function of the observable.

The constraint in Equation (2) can represent either:

- The information coming from an auxiliary measurement that constrains the shape of the background component: *e.g.* Gaussian measurements in a CR, providing the central values $r_1 \dots r_N$ along with corresponding uncertainties encoded in the kernel.

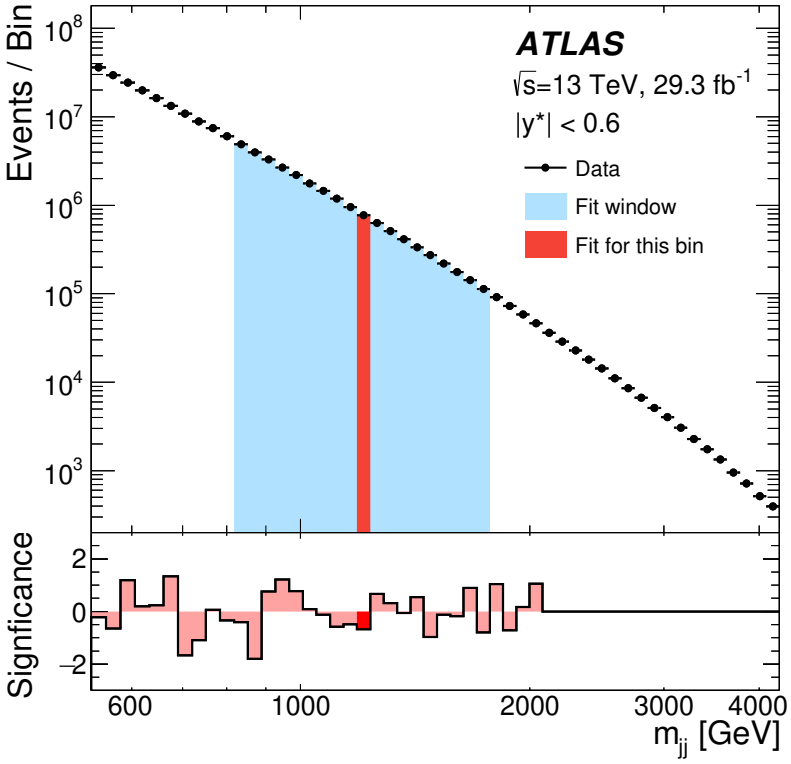

Figure 2: Example of a fit of a mass spectrum distribution to a closed-form function over a sliding-window interval, taken from [12]. The fit is performed for a signal centered on the bin shown in red, and the corresponding fit window is indicated in blue. The bottom panel shows the statistical significance of the deviations between the model prediction and the data in each measurement bin [13].

- Prior information on the background shape, for example from MC simulation. In this case the kernel is a measure of the associated simulation uncertainties
- A regularization condition enforcing a particular smoothness requirement on the background shape. In this case the kernel can take an *ad hoc* form.

The latter case is used in particular in searches for narrow resonances, in which the typical scale of variation of the background shape is much larger than for the signal. The description can then make use of a kernel with an intermediate scale, small enough to accommodate the variations of the background but large enough to avoid also absorbing a possible signal in the background model. This can be implemented using the *Radial basis function* (RBF) or *squared exponential* kernel,

$$K(m, m') = \exp\left(-\frac{(m - m')^2}{2\ell^2}\right)$$

where ℓ is the kernel scale. As noted above, this scale should ideally be larger than the signal (*e.g.* the mass resolution, for a narrow resonance), but smaller than the typical scale of the variations of the background shape. This generalizes to the *Gibbs kernel*,

$$K(m, m') = \frac{2\ell(m)\ell(m')}{\ell(m)^2 + \ell(m')^2} \exp\left(-\frac{(m - m')^2}{\ell(m)^2 + \ell(m')^2}\right)$$

where the observable scale $\ell(m)$ is an arbitrary function of the observable. A linear form $\ell(m) = \ell_0 + \beta m$ is often used: this can for instance account for a narrow resonance with a constant relative mass resolution, so that the width of the signal peak increases linearly with m.

The Gaussian mean represents prior knowledge of the shape (in a Bayesian context) or knowledge from an auxiliary experiment (in a frequentist setting). The final parameterization is often found to depend only weakly on the mean, for reasonable kernel choices.

The statistical treatment can follow either Bayesian or frequentist techniques. In a Bayesian setting, Equation (2) is considered as the prior, and a posterior is computed from Bayes' theorem [15]. In a frequentist setting, Equation (2) is considered as a constraint term that is added to the likelihood, and the PDF parameters $\nu_1 \ldots \nu_N$ are obtained from a fit to the data. In both cases, data likelihood can be constructed either as a product of Poisson terms or as their Gaussian approximation. In the Gaussian case, the results can be computed in closed form using linear least-squares (LLS) techniques [14], rather than non-linear minimization in MINUIT. This is convenient since the number of PDF parameters is equal to the number of measurement bins, which can be quite large in realistic situations. In analyses in which the data likelihood is not Gaussian, a hybrid method can be used where LLS is used to update the PDF parameters ν_i at each MINUIT iteration, while the other likelihood parameters are handled by MINUIT directly.

The kernel parameters (usually referred to as *hyperparameters*) can be obtained by maximization of the marginal likelihood [15] or by optimizing over other measures of analysis performance (for instance signal sensitivity or fit bias).

An example fit using a GPR model is shown in Figure 3.

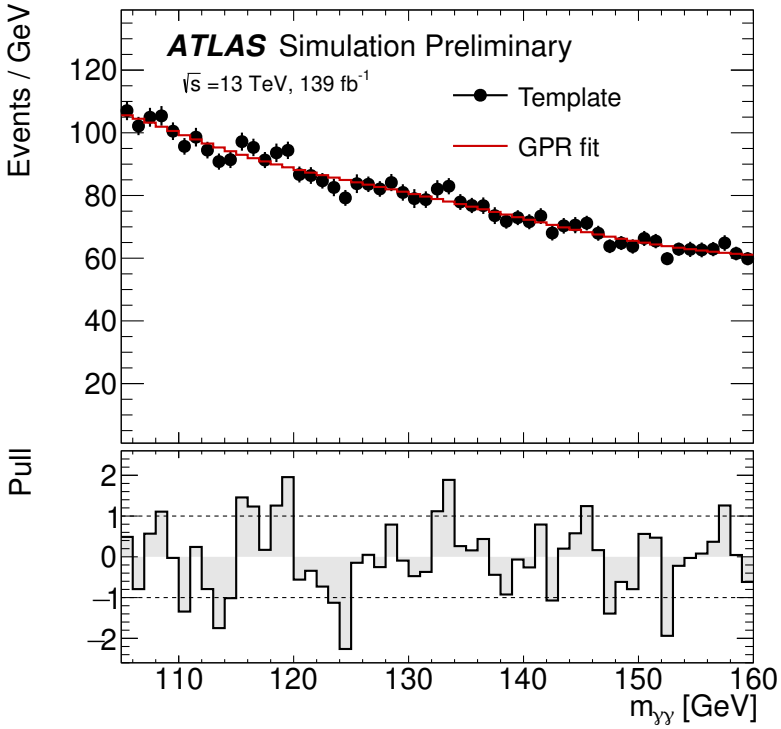

Figure 3: Example fit of a mass spectrum distribution to a GPR model, in one of the event categories of the $H \to \gamma\gamma$ analysis of Ref. [16]. The fit is here performed to a model including only a background component, in order to obtain a smoothed version of the template (See Section 3.1.1). The fit uses a Gibbs kernel with a linear dependence of the kernel scale ℓ on the observable. The bottom panel shows the difference between the fit prediction and the template yield in each bin, divided by the corresponding uncertainty.

***Pros*:**

- Fine control on the flexibility of the parameterization over the observable range, through the choice of kernel

- CPU-efficient, relies on linear algebra for most of the computations
- Generally limited dependence on the kernel and Gaussian mean, for choices that are reasonably well-suited to the problem.

***Cons*:**

- Only for binned likelihoods
- Arbitrariness in the choice of kernel and mean
- Possible unintended modeling effects, such as smoothing out features with length scales smaller than supported by the kernel
- Non-Gaussian aspects of the model (*e.g.* non-Gaussian systematics) need to be handled separately

2.3 Functional decomposition (FD)

Functional decomposition [17] is a technique in which the template shape is parameterized using a series expansion, in a similar spirit as a Fourier series. The basis functions are however based on real exponential functions: starting from $F_n(z) = e^{-nz}$, one builds linear combinations $E_n(z)$ that are orthogonal (under the L_2 norm). A few examples of the $E_n(z)$ are shown in Figure 4.

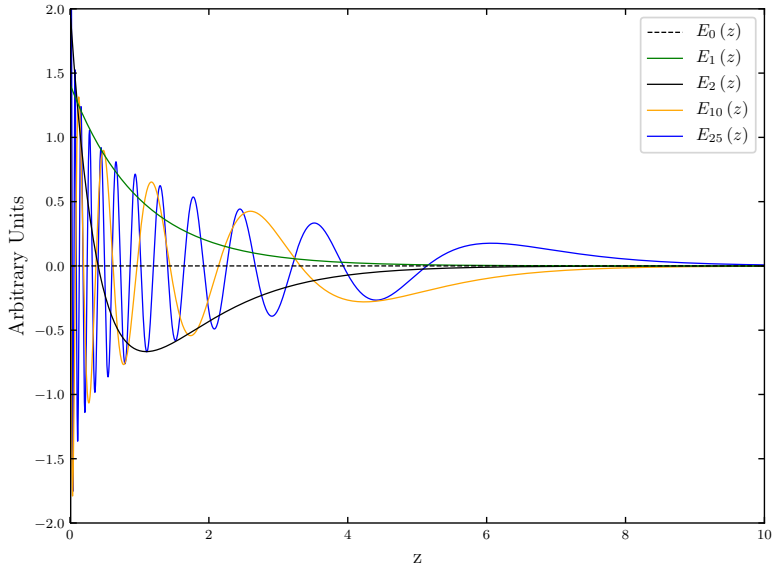

Figure 4: Examples of basis functions $E_n(z)$ used in the FD expansion (figure taken from Ref. [17])

The parameter z is a normalized version of the observable, correcting for the offset and scale of the fit range. It is defined as

$$z(m) = \left(\frac{m - m_0}{\lambda}\right)^{\alpha},$$

where m_0, λ and α are hyperparameters of the method. Any spectrum can then be decomposed as a linear combination of the $E_n(z)$,

$$f(m) = \sum_{n=1}^{N} f_n E_n(z(m))$$

with coefficients f_n that are computed from simple algebra [17]. The expansion can approximate the template shape with arbitrary precision for a sufficiently large expansion order N. The value of N is a hyperparameter of the method.

The λ and α are determined by maximizing the marginal likelihood, similarly to the GPR case described in the previous section. This is usually performed by scanning over hyperparameter values, accounting for the fact that some regions of parameter space may exhibit rapid variations of the marginal likelihood and therefore require a fine scan grid.

Fits with multiple components can be performed by expanding each component (possibly with different values of N). The fits can be performed in a hybrid manner, using MINUIT for non-FD parameters, while FD coefficients are minimized in closed form at each MINUIT iteration (similarly to the procedure for GPR described in the previous section).

An example fit using a FD model is shown in Figure 5.

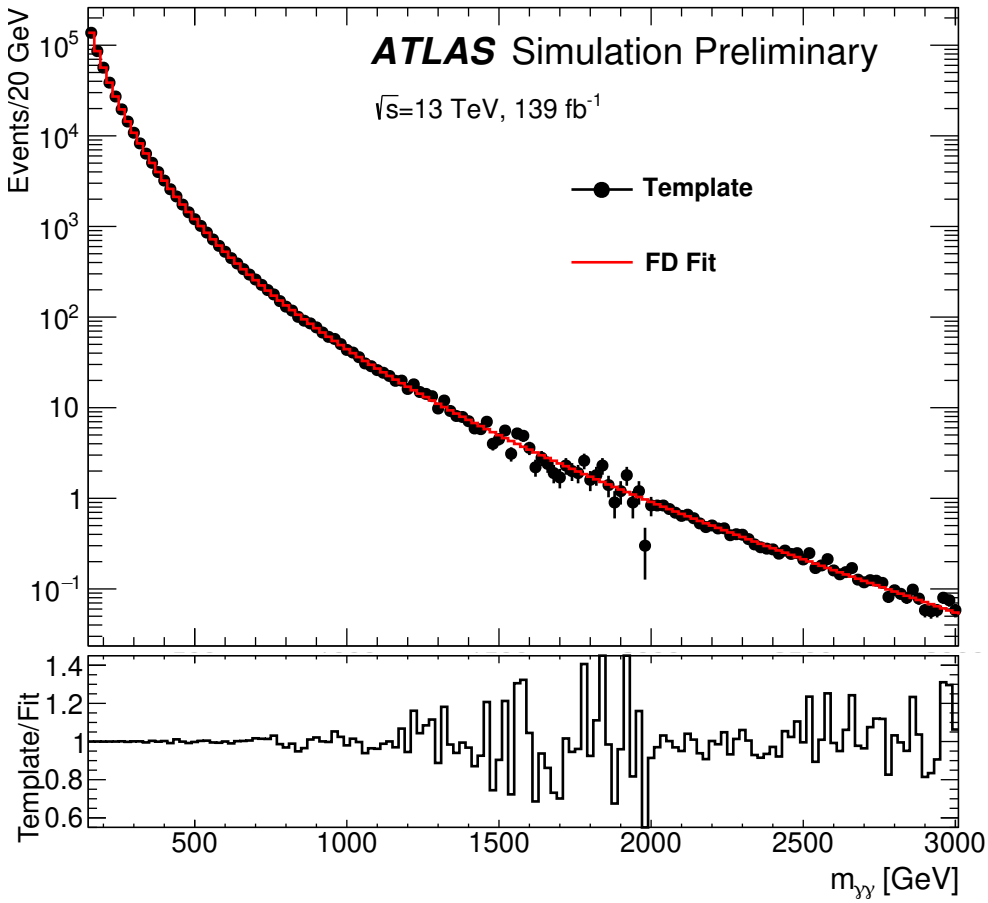

Figure 5: Example of a fit to a FD model, performed on the distribution of the diphoton invariant mass $m_{\gamma\gamma}$ in the search for high-mass resonances decaying to two photons described in Ref. [18]. The fit model includes only a background component, in order to obtain a smoothed version of the template in the same way as in Figure 3. The bottom panel shows the ratio of the template yield to the fit prediction in each bin.

***Pros*:**

- Control on the flexibility of the parameterization through the expansion order
- CPU-efficient, relies on algebra for most of the computations

***Cons*:**

- Possible unintended modeling effects, such as smoothing out features with shorter length scales than supported by the expansion
- Non-Gaussian aspects of the model (*e.g.* non-Gaussian systematics) need to be handled separately

3 Choice of model

For each modeling technique described in the previous section, one can define arbitrarily complex and flexible models:

- For functions, the flexibility is determined by by the functional form and the number of free parameters.
- For GPR, by the choice of kernel (its form and the associated hyperparameters) and of Gaussian mean.
- For FD, by the order N of the expansion, as well as the other hyperparameters listed above.

The appropriate amount of flexibility depends on the analysis setting. However general criteria on which this choice can be based are reviewed in the following sections.

3.1 Fit bias

The fit bias or *spurious signal* corresponds to the difference between the median fitted signal yield and the expected signal yield, when fitting a template with a known signal yield using a smooth background model. More flexible models generally provide better fits to data, and therefore reduce the bias in signal extraction. However excessive flexibility could lead to the background model being able to absorb part of the signal, which would increase the bias.

The appropriate level of flexibility should therefore be estimated in each analysis, in part based on bias computations on one or more well-chosen templates. The choice and construction of an appropriate template is discussed in Section 3.1.1, and the computation of the bias itself is discussed in Section 3.1.2; Section 3.1.3 discusses an application to templates with various levels of injected signal, used to check the linearity of the signal extraction.

3.1.1 Construction of the template

The template(s) used for the bias test should have a shape that is as close as possible to that of data, but with an event count that is large enough so that statistical uncertainties in the template contents are negligible. As discussed in Section 5 below, this generally requires event yield of $O(10-100)$ times larger than the data yield. In practice, this can be obtained (roughly in order of decreasing preference) from

- **A high-statistics CR**, in cases where one can be found with the same shape as that of the signal region.
- **A high-statistics MC sample**: reaching the required statistical uncertainties typically requires either fast simulation or particle-level simulation. For the latter, reconstructed quantities can be obtained by smearing and reweighting the sample using transfer functions determined from smaller samples produced with full simulation [3].

- **A smoothed MC sample or CR**: if the only template available has too low statistics to apply the bias test directly, one can first apply a smoothing procedure to the template to remove statistical fluctuations. This can be done by fitting the low-statistics sample to one of the models listed in Section 2 above: GPR, FD or a fit to a closed-form function. In all cases, the smooth model is then used to define the high-statistics template, instead of the original sample. The choice of technique depends on the one used for the analysis itself: using the same technique in both cases (*e.g.* the same function for both smoothing and fitting) can lead to a circular test with a bias that vanishes by definition. For analyses using closed-form function modeling, the smoothing should be performed using either a more complex function (*e.g.* a function of the same family but a higher number of parameters), a GPR or FD, taking care to choose a GPR mean that is not closely related to the chosen function. For analyses using GPR or FD for modeling, the smoothing can be performed using a closed-form function that provides a good fit to the low-statistics template and is not closely related to the mean of the Gaussian process.

The templates can be defined to describe background only, or a mixture of background and signal. The choice usually corresponds to the expected scenario used in the analysis: for instance background-only templates for searches, and templates corresponding to the SM expectation for measurements. Templates with a signal admixture are also used for the injection tests described in Section 3.1.3. For cases where a signal component is required, it can usually be added a posteriori to a background-only template determined as above.

If the shape of the template has associated systematic uncertainties, separate templates should also be built for each variation (*e.g.* $\pm 1\sigma$ for each independent source of uncertainty) using the same method.

3.1.2 Spurious signal estimation

The template produced as described above can be used to estimate the fit bias associated with the background model, in two main ways:

- The template can define a smooth dataset with very small statistical fluctuations, similar to an Asimov dataset [19]. In this case, the template should be normalized to the same total yield as the real dataset, in order for the signal uncertainty values (see below) to be representative of those in data.
- The template can be used to generate pseudo-experiments (*toys*) drawn from the template. The event count for each toy should be drawn from a Poisson distribution with a mean equal to the data yield, and the toy events should be drawn from the smooth template.

The modeling bias is estimated by fitting either the template itself (in the same way as one would use an Asimov dataset) or the pseudo-experiments, to a model with both signal and background components ($S + B$ fit). If the distribution of events is expected to be Gaussian, then the use of the template is preferred since it is usually less CPU-intensive : typically several 1000 pseudo-experiments should be performed for reliable results, while in the case of the template a single dataset is processed, as for an Asimov dataset. However if non-Gaussian effects are expected, then toys should be used since these effects may not be properly modeled using only the template. The pseudo-experiment method also allows to perform the fit stability checks described in Section 3.3.3 below, which can save time later.

In both cases, the fit should be performed in the same configuration as the one used in data, including all systematics and free parameters. Since the true amount of signal injected in the template is known, the fit bias or *spurious signal* can be computed as

$$S_{\text{spur}} = S_{\text{fit}} - S_{\text{template}},$$

the difference between the fitted signal yield and the one that was injected in the template. As discussed in Section 3.1.1, the value of S_{template} depends on the type of analysis performed – for instance one would use $S_{\text{template}} = 0$ for a search.

If the procedure is performed on the template itself (similarly to an Asimov dataset), then S_{spur} is evaluated from a single fit to the template. If pseudo-experiments are used, the median of the distribution of the per-experiment S_{spur} is used instead (and also denoted as S_{spur} in the following for simplicity).

In both cases, the uncertainty σ_{fit} on S_{fit} should also be recorded. If the results are obtained from the template, the parabolic fit uncertainty should be used, while for pseudo-experiments one can use the RMS of the distribution of the S_{fit}. For a resonance search performed over a mass range, the spurious signal and its uncertainty should be evaluated as a function of the resonance mass m_X (both for full-range and sliding-window fits). If other signal model parameters affect the spurious signal (*e.g.* the resonance width), separate spurious signal values should also be obtained for each tested model hypothesis.

The points at which the spurious signal is evaluated should be spaced sufficiently close to each other to capture all the variations of S_{spur} as a function of the model parameters. For instance in the case of variations with the resonance mass, the points should form a grid with a spacing comparable to the width of the signal peak. For the case of a resonance width, or other cases in which the spurious signal typically varies monotonically with the model parameter, testing only the extreme values is sufficient.

An example spurious signal computation as a function of mass is shown in Figure 6 .

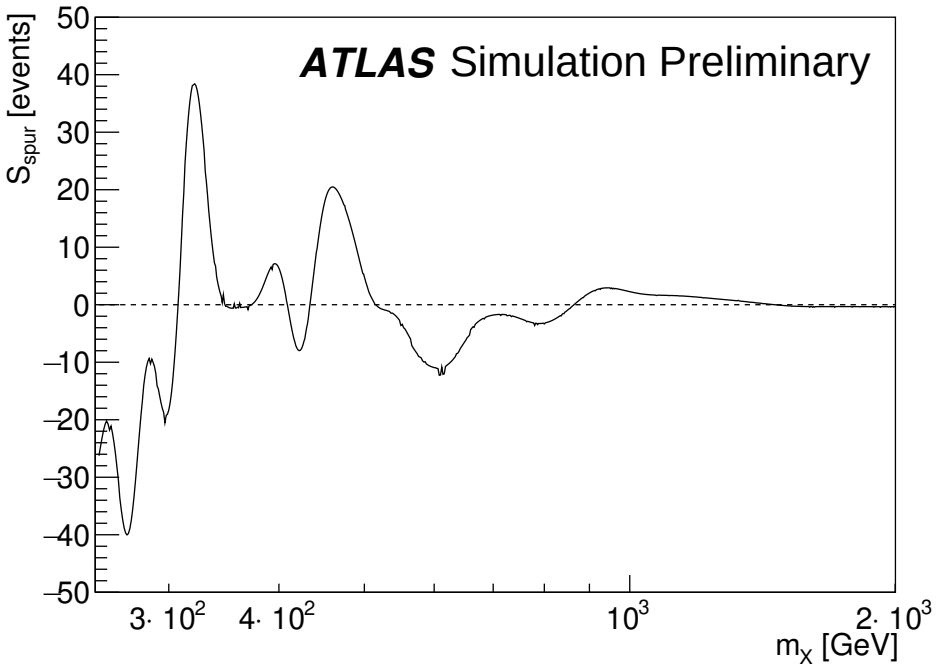

Figure 6: Spurious signal values S_{spur} as a function of the resonance mass m_X, from the search for high-mass dilepton resonances of Refs. [3].

If the shape of the template has associated systematic uncertainties, separate spurious signal values $S^{\pm}_{\text{spur}}$ should also be defined for each template variation.

3.1.3 Signal injection tests

The spurious signal measurement described in the previous section ensures that the fit bias is small for the particular signal hypothesis corresponding to the expected analysis scenario. It is also useful to check that the fit bias remains small for other signal yields, to ensure that the fit provides reliable results in other possible analysis outcomes.

For searches, in which the expected scenario corresponds to background alone, injection tests should be performed for various levels of injected signal, mainly focusing on signal levels similar to the expected 95% CL upper limit.

For measurements, signal yields in a range centered on the expected yield and with a width of a few times its expected uncertainty should be considered. In both cases, a few test values within the specified range can be considered, so that the linearity of the extraction procedure can be verified.

As for the nominal spurious signal measurement, the signal injection procedure can be performed either using a single "Asimov" fit on the template, or on toys generated from the template shape.

3.2 Signal sensitivity

More flexible models generally lead to lower signal sensitivity, since this flexibility decreases the difference between background and signal shapes and therefore the separation power of the shape analysis.

The sensitivity of different models can be generally compared using the value of σ_{fit} computed as described in the previous section. This corresponds to the uncertainty on the signal yield in a fit to a template (or a set of pseudo-experiments) corresponding to the expected data. This expectation should itself correspond to the nature of the analysis: i.e. a background-only template should be used for searches, and a mixture of signal and background in the expected proportions for a measurement. Note that σ_{fit} is a well-defined quantity even when no signal is present in the template and can therefore be used for searches. Other related measures such as the expected signal significance can also be used instead.

In simple cases the sensitivity can also be estimated more simply – for instance for functions of the same family, the number of free parameters can be used as a proxy. This is due to the fact that a higher number of free parameters generally provides more flexibility to the background description, and therefore decreases the sensitivity to the signal. Within a given family of functions, the signal sensitivity therefore decreases with the function order as new free parameters are added.

3.3 Model selection criteria

3.3.1 Spurious signal criteria

Criterion relative to the uncertainty on the signal yield

The absolute spurious signal $|S_{\text{spur}}|$ should ideally be small. There is some arbitrariness as to how small is small enough, but the choice should be driven by the following considerations:

- The uncertainty $|S_{\text{spur}}|$ is generally implemented as an additive systematic term in the expression for the signal yield (see Section 4 below) associated with a Gaussian-constrained nuisance parameter. For Gaussian measurements, this leads to a total uncertainty on the signal yield $\sigma_{\text{tot}} \approx \sqrt{\sigma_{\text{fit}}^2 + S_{\text{spur}}^2}$. Avoiding a large increase in the uncertainty therefore requires $S_{\text{spur}} \lesssim 0.5\sigma_{\text{fit}}$. The magnitude of S_{spur} is affected by the statistical fluctuations of the template from which it is computed and can therefore be sizable even in the absence of a true modeling bias. The dataset used to build the template must therefore be large enough to ensure that the template statistical uncertainties do not contribute to σ_{tot} at a level similar to σ_{fit}. This issue is discussed further in Section 5 below.
- While the additive systematic term can partially compensate for modeling biases, it cannot fully absorb a bias of size S_{spur} since the Gaussian constraint assigns a penalty to cases where the nuisance parameter associated with the modeling is non-zero. True biases should therefore remain at a level much below that σ_{fit}, to avoid potential undercoverage.

The spurious signal criterion is generally applied as

$$S_{\text{spur}} < (20\%\text{–}50\%)\,\sigma_{\text{fit}}.$$

For Gaussian measurements, a 20% threshold corresponds to an increase in the measurement uncertainty by a generally negligible factor of $\sqrt{1+0.2^2} \approx 1.02$, and should represent a very safe choice. Thresholds value of up to 50% have also been considered and should represent safe choices, albeit leading to larger systematic impact (*e.g.* $\sqrt{1+0.5^2} \approx 1.12$ for a 50% threshold). However larger thresholds are less sensitive to statistical fluctuations in the template and therefore have lower requirements on the template size. A threshold value of 30% should represent a good compromise between these two requirements.

Values larger than 50% should be used with caution since as noted above, true biases of this size cannot in general be fully covered by the modeling uncertainties described in Section 4. The use of thresholds above 50% should be supported by modeling studies in which the background shape is modified to include a true bias of this size, and its effect on the signal yield is probed in the presence of the modeling systematics.

An alternative to increasing the spurious signal threshold is to simplify the modeling by using a narrower fit range, or a sliding-window technique as described in Section 2.1.

For models with parameters affecting the spurious signal computation, the criterion should be applied at all tested model hypotheses: in particular, for mass-dependent searches the criterion should be verified for all m_X in the search range. For measurements at a fixed mass (*e.g.* Higgs boson measurements), one may require the criterion to be verified in a range around the resonance mass. This protects against possible underestimations of the spurious signal due to accidental features of the template, as well as possible fluctuations in the signal peak position.

As a side note, the spurious signal may also be used to define the range of a model parameter covered in the analysis, as the range for which the spurious signal criterion is verified. This applies in particular to parameters, such as a resonance width, which induce monotonic increases in the spurious signal values.

If there are systematic uncertainties associated with the shape of the template, the spurious signal criterion should be verified on all the systematic variations of the template (constructed as described in Section 3.1.1).

An example of the application of these criteria is shown in Figure 7.

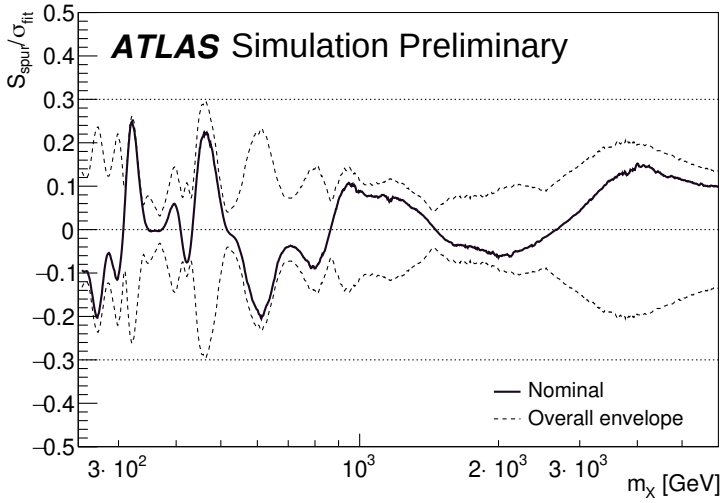

Figure 7: Value of $S_{\text{spur}}/\sigma_{\text{fit}}$ as a function of the resonance mass m_X, taken from the high-mass dilepton search of Refs. [3]. The dotted lines correspond to the criterion $S_{\text{spur}} < 30\%\,\sigma_{\text{fit}}$. The solid line shows the values of $S_{\text{spur}}/\sigma_{\text{fit}}$ obtained in the nominal template, while the dotted line shows the symmetrized envelope of the curves obtained for all the systematic variations of the template. The fact that the envelope is fully contained within the dotted lines shows that the spurious signal criterion is verified for all values of m_X.

Criterion relative to the expected signal yield

In some cases, an expected signal is present in the analysis: this is the case for measurements, for which a signal component is part of the nominal expectation, and for the signal injection tests described in Section 3.1.3, where a non-zero signal is also considered.

For these cases, an expected signal is included in the template and the spurious signal is therefore defined as $S_{\text{spur}} = |S_{\text{fit}} - S_{\text{template}}|$ (see Section 3.1.2). This spurious signal is then required to be small relative to the expected signal S_{template}. The criterion

$$S_{\text{spur}} < 10\%\, S_{\text{template}}$$

has been used in Higgs boson measurements and similar bounds should be generally applicable. The criterion should be applied to the nominal template, if signal is included in the nominal expectation, and also for each injected signal hypothesis when performing injection tests.

Limited template statistics

In cases where the template statistics are insufficient (see Section 5), template statistical fluctuations may give large contributions to $|S_{\text{spur}}|$.

For templates with weighted events (e.g. templates built from MC simulation, or from a CR that has been reweighted to the same event yield as in the signal region), the size of the fluctuations can be estimated in RooFit by fitting the template with the `SumW2Error(True)` option: this provides a fit uncertainty $\sigma_{\text{fit}}^{\text{template}}$ accounting for the individual weights of the template events, rather than the total weighted bin contents. Fluctuations of order $\sigma_{\text{fit}}^{\text{template}}$ are expected for $|S_{\text{spur}}|$, so that one should have $\sigma_{\text{fit}}^{\text{template}} \ll \sigma_{\text{fit}}$ for reliable spurious signal estimations. One can also check that the measured fluctuations in $|S_{\text{spur}}|$ are at least of the order of $\sigma_{\text{fit}}^{\text{template}}$, since significantly smaller values could indicate an underestimation of $|S_{\text{spur}}|$.

Given the above, fluctuations of order $\sigma_{\text{fit}}^{\text{template}}$ cannot be taken as evidence for low "true" $|S_{\text{spur}}|$, since a sizable true $|S_{\text{spur}}|$ may be hidden in the statistical fluctuations. However it can be clear in some cases that a large $|S_{\text{spur}}|$ may be due to a template fluctuation, for instance in the case of a narrow peak in $|S_{\text{spur}}|$ which is incompatible with a true mismodeling of the template. In these cases, a *relaxed* version of the bias criterion may be applied in the form

$$|S_{\text{spur}} - 20\%\sigma_{\text{fit}}| < n_\sigma \sigma_{\text{fit}}^{\text{template}},$$

where the value of $n_\sigma \sim 1 - 2$ can be tuned to account for the features of the template.

The use of the relaxed criterion and the value of n_σ should be motivated by the presence of a likely statistical fluctuation in the template that prevents the application of the baseline criterion. In particular, cases where the relaxed criterion passes but the baseline one does not should be checked in order to understand whether this can be ascribed to a likely statistical fluctuation or to other features of the template, since the latter may have a physical origin. Such checks may include alternative templates, or ad-hoc arguments such as whether the variations of S_{spur} around the offending values are compatible with those expected for a true background modeling bias.

For cases in which the baseline criterion fails due to a likely statistical fluctuation, an alternative approach to the relaxed criterion is to smooth the template as described in Section 3.1.1, since this may also help to remove features that are known to be incompatible with the true background shape.

3.3.2 Sensitivity criterion

Among the models passing the spurious signal criteria, the one providing the highest sensitivity to signal should be selected. As described in Section 3.2, this can be achieved by choosing the model giving the smallest σ_{fit}. In the case of functions of the same family (*e.g.* polynomials), one may also simply select the one with the least number of parameters.

3.3.3 Other criteria

Several other checks should be performed to ensure the model provides a good description of the data:

- The model should provide a good fit to the template defined in Section 3.1.1. For this purpose, a goodness-of-fit criterion should be applied in addition to the spurious signal requirements. The criterion $p(\chi^2) > 1\%$ has been applied in Higgs boson analysis, with the χ^2 computed using uncertainties reflecting the event yield in the data, rather than the samples used to build the template. A similar criterion should be generally applicable in other situations. In cases where the criterion cannot be met, a possible solution is to simplify the modeling problem by reducing the fit range or using a sliding-window fit technique as described in Section 2.1.
- Stability checks should be performed by fitting the model to toy datasets generated from the template defined in Section 3.1.1, in the expected signal hypothesis. These tests should ensure that fits perform as expected, with good convergence and sensible distributions for the best-fit values of all parameters. If the model is expected to be approximately Gaussian, it should be checked that these distributions are indeed Gaussian to a good approximation. In all cases, unexpected shapes or the presence of outliers should be investigated. At least a few hundred toy datasets should be used to ensure that problems occurring at percent-level rates are identified.
- Since some features in data may not be described by the MC, an F-test should be performed in data to validate the chosen model against an alternate model with higher complexity. In the case of closed-form functions, the alternate model should be a function of the same family, but with one or more additional free parameters (*e.g.* for polynomial functions, the alternate is obtained by adding more polynomial orders). For FD, one can similarly use a model with a larger expansion order N. For the GPR case, no prescription for this test is yet defined.

The F-test uses the test statistic

$$F = \frac{\frac{\chi^2_{\text{nom}} - \chi^2_{\text{alt}}}{n_{\text{alt}} - n_{\text{nom}}}}{\frac{\chi^2_{\text{alt}}}{n - n_{\text{alt}}}}$$

where χ^2_{nom} and χ^2_{alt} are the χ^2 values for fits to data using respectively the nominal and the alternate model, n_{nom} and n_{alt} are the number of free parameters in each model, and n is the number of bins used for the χ^2 computation.

The value of F is large if the alternate model provides a significantly better fit to data than the nominal model. In the asymptotic limit F follows the F-distribution $F(n_{\text{alt}} - n_{\text{nom}}, n - n_{\text{alt}})$, so that F can be converted to a p-value $p(F)$ against the alternate model (for instance using the `ROOT::Math::fdistribution_cdf_c` function of `ROOT`). An upper bound on $p(F)$ can therefore be applied as a test that the additional degrees of freedom do not significantly improve the description of data. Typically the nominal model is retained as long as $p(F) > 0.05$.

A similar test can be performed using the profile-likelihood ratio

$$\lambda_F = -2 \log \frac{L_{\text{alt}}}{L_{\text{nom}}}$$

where L_{nom} and L_{alt} are respectively the profile-likelihood values obtained in data using the nominal and the alternate model. As for the χ^2-based definition above, the nominal model is nested within the alternate model, which includes additional free parameters. These two likelihood values therefore correspond to the case in which these additional parameters are free in the fit (for L_{alt}) or where they are fixed to their nominal values (for L_{nom}). In the asymptotic approximation λ_F follows a χ^2 distribution with a number of degrees of freedom equal to the number of additional parameters. For the case of one additional parameter, the p-value against the alternate model can therefore be obtained as $p(F) = 2[1 - \Phi^{-1}(\sqrt{\lambda_F})]$ where Φ is the cumulative distribution function of the normal distribution. Again, the nominal model can be retained as long as this p-value is not too small. Typically the criterion used is again $p(F) > 0.05$.

The likelihood-based criterion is expected to be more robust against non-Gaussian behavior than the χ^2-based criterion, in spite of the reliance on asymptotic formulas. It is also simple to implement for analyses using likelihood-based models. It should therefore be preferred for likelihood-based analyses, and in particular when non-Gaussian effects are expected.

The F-test should ideally be performed in a *blind* manner: for instance wide bins can be used to hide possible narrow signal peaks. Alternatively, the procedure can be applied after the unblinding of the data, as long as one fully specifies beforehand the test criteria and the alternative functions to use in case of test failure.

The model used in the test should correspond to the expectation: for searches, a background-only model should be used, while measurements should also include a signal component corresponding to the expected signal yield.

An illustration of the application of the F-test is shown in Figure 8.

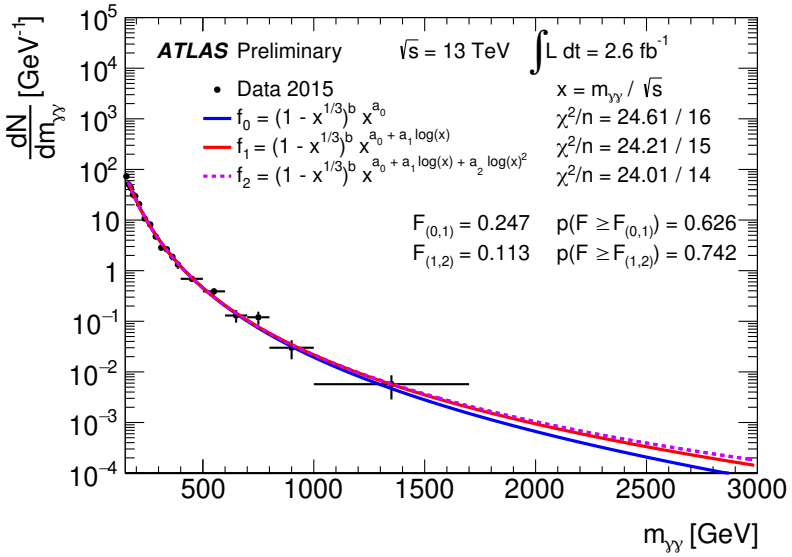

Figure 8: Example of an F-test applied in the high-mass $\gamma\gamma$ search [20]. The test compares the χ^2 values for 3 versions of the functional form defined by Equation (1) (see Section 2.1). The large values of $p(F)$ for each test indicate no significant preference for the additional degrees of freedom provided by f_1 over f_0, or f_2 over f_1.

In cases where these tests fail, possible solution can be to simplify the modeling by using a narrower fit range, or a sliding-window technique as described in Section 2.1.

4 Modeling uncertainties

Modeling uncertainties are defined under the conservative assumption that the spurious signals measured in the various templates considered are not necessarily an accurate estimation of the true biases in fits to data. Rather, the assumption is that the S_{spur} values can only be used as an estimate of the typical size of possible biases. In particular, no attempt is made to correct for these biases, and S_{spur} values are only used to quantify the size of the modeling uncertainties.

Similarly, the fact that S_{spur} is compatible with zero, within the uncertainties of its measurement (i.e. the quantity $\sigma_{\text{fit}}^{\text{template}}$ defined in Section 3.3.1), should not be interpreted as an indication that modeling uncertainties can be neglected. This situation can occur for instance in cases where the template used in the spurious signal computation has low statistics, and does not preclude the presence of true biases hidden within the statistical noise. The measured S_{spur} should therefore be used also in this case as a conservative estimate of the size of the modeling uncertainties.

In general, the spurious signal is computed on the nominal template as well as on templates corresponding to $\pm 1\sigma$ variations in independent sources of uncertainty (See Sections 3.1.1 and 3.1.2). In the following, we call S^0_{spur} the spurious signal computed on the nominal template, and $S^{\pm}_{\text{spur}}$ the spurious signal computed in each varied template.

The modeling uncertainties are then defined as follows:

- For analyses with no free signal shape parameters, the baseline modeling uncertainty is $S^0_{\text{modeling}} = |S^0_{\text{spur}}|$.
- For analyses performed over a range of values of a signal shape parameter, for instance a resonance mass m_X, the baseline uncertainty $S^0_{\text{modeling}}(m_X)$ is defined as the envelope of $|S^0_{\text{spur}}(m_X)|$ over m_X. The envelope is defined by a smooth function passing through the extreme values of $|S^0_{\text{spur}}(m_X)|$. This avoids the issue of uncertainties that vanish near zero-crossings of $S^0_{\text{spur}}(m_X)$, since the positions of these crossings depend on the precise shape of the template and are not expected to be exactly reproduced in data. An example of the computation of the envelope is shown in Figure 9. In some cases, a few spurious signal values may be excluded from the envelope computation, if they are deemed to originate from statistical fluctuations rather than a true bias: this may apply for instance to spurious signal variations with a length scale that is considered to be incompatible with originating from a true modeling bias. However this procedure should be applied with caution since there is generally no way to know whether the spurious signal can be attributed entirely to fluctuations, or at least in part to a true modeling bias.

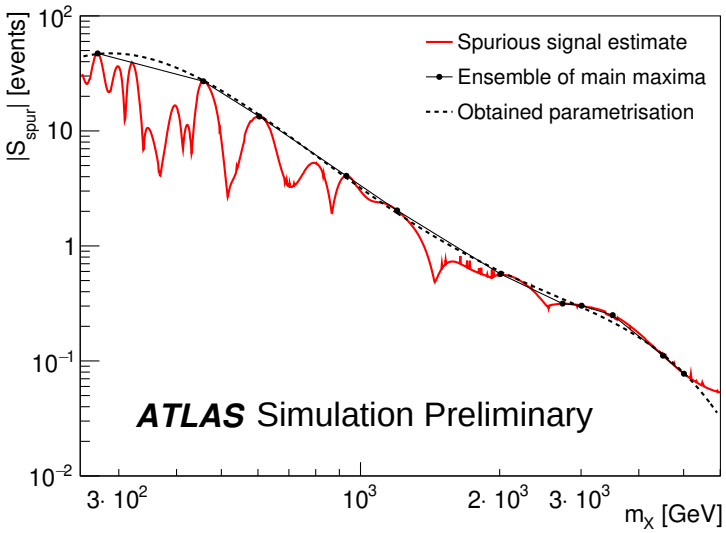

Figure 9: Example of the computation of the modeling systematic uncertainty for the high-mass dilepton search of Refs. [3]. The plot shows the absolute spurious signal $|S_{\text{spur}}(m_X)|$ (solid red line) and the envelope used to define the modeling systematic uncertainty $S_{\text{modeling}}(m_X)$ (dashed line).

In both cases, S^0_{modeling} should be at least as large as σ_{fit}, the uncertainty on S_{spur} due to statistical fluctuations in the template (see Section 3.1.2) : if this is not the case, then the values $|S^0_{\text{spur}}|$ used in the computation of S^0_{modeling} may be artificially low and lead to an underestimated uncertainty.

The value of S^0_{modeling} is considered as a systematic uncertainty on the signal yield, to cover potential modeling biases. In likelihood analyses, the systematic is implemented as an additional signal contribution; for instance for a signal yield parameterized with a cross-section σ_{signal}, the total signal yield is written as

$$N_{\text{signal}}(m_X) = \sigma_{\text{signal}}\,\mathcal{L}\,(\mathcal{A}\times\epsilon) + S^0_{\text{modeling}}(m_X)\,\theta_{\text{modeling}}$$

where the first term corresponds to the true signal yield, and the second the spurious signal defining the modeling uncertainty; $\mathcal{L}$ and $\mathcal{A}\times\epsilon$ are respectively the integrated luminosity and acceptance times efficiency factors, and

θ_{modeling} the nuisance parameter (NP) associated with the uncertainty. An external Gaussian constraint with mean 0 and width 1 is applied on θ_{modeling} following (see Ref. [21] for a description of a full analysis implementation). Since the uncertainty is applied directly on the yield, correlations between different m_X values need not be considered (different resonance masses correspond to different fits). In non-likelihood analyses, the modeling uncertainty can simply be added in quadrature to the uncertainty on the signal yield.

Separate systematic uncertainties $\delta S^{\pm}_{\text{modeling}} = |S^{\pm}_{\text{modeling}} - S^{0}_{\text{modeling}}|$ should also be considered to account for the differences between the S^{0}_{modeling} obtained in the nominal template and the $S^{\pm}_{\text{modeling}}$ obtained in templates with independent systematic variations applied (with the same method of computation as described above, but using $S^{\pm}_{\text{spur}}$ instead of S^{0}_{spur}). The uncertainties can be included in several ways:

1. For Gaussian measurements, each $\delta S^{\pm}_{\text{modeling}}$ is added in quadrature to the nominal S^{0}_{modeling}. The total uncertainty is then included as a single additive term in the formula above, with a single NP.

2. For cases where the number of independent variations is small, the total uncertainty can be computed as the envelope of S_{modeling} over all systematic variations (as well as over m_X, if applicable). An example of this computation is shown in Figure 10. The total uncertainty is then included as a single additive term in the formula above, with a single NP. The method should be applied with caution if the combined effect of the variations can be expected to be significantly larger than their envelope, for instance if the variations are strongly correlated. The method is also not applicable for a large number of independent variations, which can have a combined effect that is much larger than the envelope of their individual impacts. In these cases, the methods listed below should be preferred instead.

3. The total uncertainty can also be obtained from spurious signal values computed from an ensemble of toys in which the systematic variations are randomly sampled. This follows the same procedure as the pseudo-experiment technique described in Section 3.1.2, except for the fact that each toy is drawn not from the nominal template, but from a template with variations applied. This template is randomly selected for each toy from the ensemble of systematic variations. This technique is equivalent to summation in quadrature in Gaussian cases, but also allows the inclusion of non-Gaussian effects as well as possible correlations between the uncertainties.

4. For each variation, a separate additive term is included in the signal yield, each with a separate NP. If the positive and negative variations are not identical ($\delta S^{+}_{\text{modeling}} \neq \delta S^{-}_{\text{modeling}}$), an interpolation procedure should be applied between the uncertainties assigned to positive and negative NP values.

The techniques are listed roughly in order of increasing complexity. In cases where the $\delta S^{\pm}_{\text{modeling}}$ are small relative to S^{0}_{modeling}, Method 1 (addition in quadrature) is recommended, with Method 2 (envelope-based estimation) a possible alternative. In cases where the $\delta S^{\pm}_{\text{modeling}}$ are large, very asymmetric, or the associated uncertainties are expected to be significantly non-Gaussian, then Method 4 (individual NPs) should be used preferentially as this provides the finest description of the uncertainties. However Method 3 can be used as an alternative in cases where Method 4 is not feasible – for instance in the case of a large number of NPs, or unreliable uncertainty values due to limited template statistics.

As noted in Section 3.3.1, the modeling uncertainty described here may not fully cover a true bias of the size indicated by the spurious signal. If such a bias were present, it could be accounted for by a 1σ shift in the value of the NP associated with the uncertainty; however due to the Gaussian nature of the constraint, this would incur a likelihood penalty, so that the best-fit value of the NP would generally not amount to a full 1σ shift. This is one of the reasons for restricting the spurious signal to less than 50% of the statistical uncertainty, so that potential undercoverage is limited when all sources of uncertainty are accounted for. Conversely, the uncertainty may over-cover if the spurious signal value is dominated by statistical fluctuations in the template, rather than a true modeling bias. As noted above, fluctuations do not preclude the existence of a true bias, but may lead to an uncertainty value that is significantly larger than needed to cover the bias itself. The use of an envelope to define

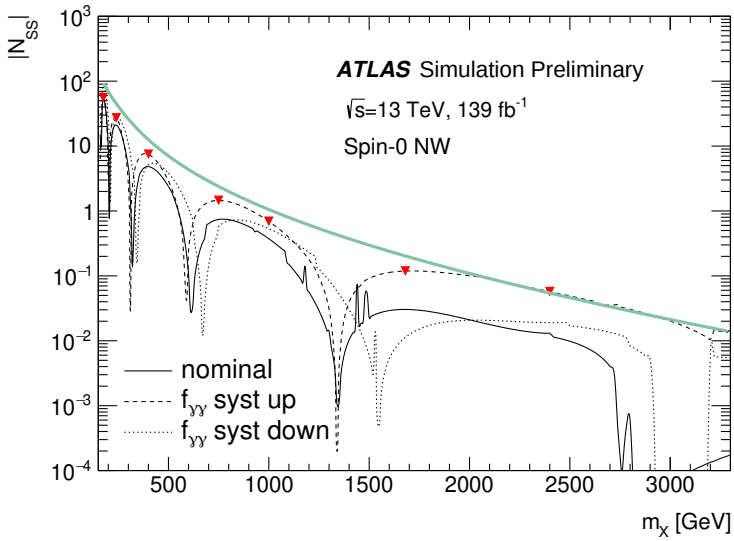

Figure 10: Example of the computation of the modeling systematic uncertainty for the high-mass $\gamma\gamma$ search [18]. The plot shows the absolute spurious signal $|S_{\text{spur}}(m_X)|$ computed for the nominal background shape (solid black line) and its two systematic variations (dashed and dotted lines). The parameterization used to define the modeling systematic uncertainty $S_{\text{modeling}}(m_X)$, which approximates the envelope of the three spurious signal curves, appears as a solid green line.

the uncertainty as a function of a parameter may also lead to an overestimation of the uncertainty for some values of the parameter.

5 Event yield requirements for template production

A template, produced from MC or a CR as described in Section 3.1.1, is used in several of the tests described above. It should contain a large enough number of events so that the results of the tests are not obscured by statistical fluctuations. This is particularly important for spurious signal tests: if the template fluctuations have a shape that is similar to that of the signal, they will provide a spurious signal contribution. If this makes the spurious signal so large that it fails the criteria of Section 3.3, this can lead to the incorrect rejection of the background model.

This issue cannot usually be solved by increasing the number of free background parameters, since a background model that could accommodate them would typically also absorb a true signal component; this motivates the use of the relaxed selection criteria presented in Section 3.3.1, which are designed to be robust against the presence of such fluctuations. However relaxed criteria also lead to lower sensitivity to a true mismodeling of the background, and it is generally preferable to avoid the issue by using larger samples of simulation or CR data to generate the templates. In this section, we estimate how large these samples should be in order to apply the non-relaxed spurious signal criteria.

The impact of limited template statistics on the fitted spurious signal is estimated by $\sigma_{\text{fit}}^{\text{template}}$, computed as described in Section 3.1.2. For spurious signal criteria of the form $S_{\text{spur}} < \alpha\,\sigma_{\text{fit}}$ (with typically $\alpha \sim 20\%$–50%), one needs $\sigma_{\text{fit}}^{\text{template}}$ to be smaller than $\alpha\,\sigma_{\text{fit}}$ by a significant factor, so that the template fluctuations of size $\sigma_{\text{fit}}^{\text{template}}$ do not saturate the spurious signal bound. One can therefore require $\sigma_{\text{fit}}^{\text{template}} < \alpha\,\sigma_{\text{fit}}/K$ with a factor K chosen to be $K \sim 2-3$ to enforce this safety margin. Since the uncertainties scale as the inverse square-root of the event yields, this in turn leads to the requirement

$$\frac{N_{\text{template}}}{N_{\text{data}}} = \frac{K^2}{\alpha^2}$$

on the ratio of the template statistics N_{template} to the data statistics N_{data}. For $K = 2$ and $\alpha = 20\%$, this amounts to $N_{\text{template}} = 100\, N_{\text{data}}$, while a looser spurious signal requirement of $\alpha = 50\%$ leads to $N_{\text{template}} = 16\, N_{\text{data}}$.

These values are only valid when the search is performed at a single point in the model parameter space. For a search over *e.g.* a mass range, the spurious signal criterion should be verified everywhere and this leads to a *look-elsewhere effect*, with a probability to have a large fluctuation *somewhere* in the spectrum which is higher than the naive estimates used above. This can be mitigated by using larger values of K in analyses with wider search ranges, but this increases further the required value of N_{template}.

In cases where the statistics requirements are difficult to achieve, the techniques described in Section 3.1.1 can be used: for instance the use of fast or parameterized MC instead of full MC, or applying smoothing algorithms to the template before performing the spurious signal computation.

6 Summary of recommendations

The procedure to select an appropriate smooth model and assign a modeling uncertainty can be summarized as follows:

- The spurious signal S_{spur} should be computed by fitting a $S + B$ model to a template with a known signal yield, and computing the difference between the fitted and the expected signal yields. The associated fit uncertainty σ_{fit} should also be computed. The template should be defined as in Sections 3.1.1 and 5. If the fit behavior is Gaussian, the template can be either fitted directly (similarly to an Asimov dataset); otherwise, it serves as a basis for the generation of an ensemble of toy datasets from which S_{spur} and σ_{fit} are obtained.
  - The model should verify $S_{\text{spur}} < 30\%\, \sigma_{\text{fit}}$ for all points in model parameter space. If this threshold cannot be met, its value can be relaxed up to $50\%\, \sigma_{\text{fit}}$, although this leads to a larger modeling uncertainty. If the shape of the template used to compute S_{spur} is affected by systematic uncertainties, the criterion should be passed for all possible systematic variations of the template.
- Among models passing the spurious signal criterion, the one with the highest sensitivity should be used. For models based on closed-form functions, one can simply chose the model with the smallest number of degrees of freedom.
- The fit of the model to the nominal template, scaled to a number of events matching those of the data, should have a χ^2 probability $p(\chi^2) > 1\%$.
- Fit stability should be checked on $O(100)$ toy datasets generated from a distribution corresponding to the expectation.
- For models based on closed-form functions or FD, an F-test should be performed in data, using as alternate a model of the same family with one more degree of freedom. The model with the extra degree of freedom should be used if the F-test probability verifies $p(F) < 0.05$. The likelihood implementation of the F-test described in Section 3.3.3 should be preferred for likelihood-based analyses.
- For measurements with no free parameters in the signal shape, the modeling uncertainty should be $S_{\text{modeling}} = |S_{\text{spur}}|$. For searches over a range of values for a signal shape parameter m_X, the uncertainty should be computed as the envelope of $|S_{\text{spur}}(m_X)|$ over m_X. The same should also be done for other model parameters on which S_{spur} may depend. For likelihood-based analyses, this uncertainty should be implemented as an extra additive term $S_{\text{modeling}}\, \theta_{\text{modeling}}$ in the signal yield; for other analyses, the uncertainty should be added in quadrature to the uncertainty on the event yield.
- Systematic uncertainties on the template shape should be accounted for by computing the spurious signal $S^{\pm}_{\text{modeling}}$ on each varied shape. The differences $\delta S^{\pm}_{\text{spur}}$ between these values and the spurious signal S^{0}_{spur} computed on the nominal template should be considered as additional uncertainties on the signal yield. Alternatively, these uncertainties can be included by taking an envelope over all systematic variations when

computing the modeling uncertainty S_{modeling} as described in the previous bullet, if the number of variations is not too large ($\lesssim 10$). The shape variations can also be included by sampling the systematic variations when computing the spurious signal on toy datasets as described in the first bullet.

References

[1] ATLAS Collaboration. *Measurements of $t\bar{t}$ differential cross-sections of highly boosted top quarks decaying to all-hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*, Phys. Rev. D **98** (2018) 012003, arXiv:1801.02052, .

[2] ATLAS Collaboration. *Search for new resonances in mass distributions of jet pairs using 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **03** (2020) 145, arXiv:1910.08447.

[3] ATLAS Collaboration. *Search for high-mass dilepton resonances using 139 fb^{-1} of pp collision data collected at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. B **796** (2019) 68, arXiv:1903.06248.

[4] ATLAS Collaboration. *Search for new phenomena in high-mass diphoton final states using 37 fb^{-1} of proton–proton collisions collected at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. B **775** (2017) 105, arXiv:1707.04147.

[5] ATLAS Collaboration. *Search for $t\bar{t}$ resonances in fully hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **10** (2020) 61, arXiv:2005.05138.

[6] ATLAS Collaboration. *Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **98** (2018) 052005, arXiv:1802.04146.

[7] ATLAS Collaboration. *A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector*, Submitted to Phys. Lett. B. arXiv:2007.07830.

[8] K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke. *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, Technical Report CERN-OPEN-2012-016, CERN, 2012.

[9] S. Bernstein. *Démonstration du Théorème de Weierstrass fondée sur le calcul des Probabilités*, Comm. Soc. Math. Kharkov **13** (1912) 1

[10] W. Verkerke and D. P. Kirkby. *The RooFit toolkit for data modeling*, eConf **C0303241** (2003) MOLT007 arXiv:physics/0306116.

[11] F. James and M. Roos. *Minuit - a system for function minimization and analysis of the parameter errors and correlations*, Computer Physics Communications **10** (1975) 343-367.

[12] ATLAS Collaboration. *Search for Low-Mass Dijet Resonances Using Trigger-Level Jets with the ATLAS Detector in pp Collisions at $\sqrt{s} = 13$ TeV*, Phys. Rev. Lett. **121** (2018) 081801, arXiv:1804.03496.

[13] ATLAS Collaboration. *Formulae for Estimating Significance*, ATL-PHYS-PUB-2020-018, 2020. URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-025.

[14] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X. URL: http://www.gaussianprocess.org/gpml.

[15] M. Frate, K. Cranmer, S. Kalia, A. Vandenberg-Rodes, and D. Whiteson. *Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes*, arXiv:1709.05681.

[16] ATLAS Collaboration. *Measurement of the properties of Higgs boson production at $\sqrt{s} = 13$ TeV in the $H \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment*, ATLAS-CONF-2020-026, 2020. URL: https://cds.cern.ch/record/2725727.

[17] R. Edgar, D. Amidei, C. Grud, and K. Sekhon. *Functional Decomposition: A new method for search and limit setting*, arXiv:1805.04536.

[18] ATLAS Collaboration. *Search for resonances decaying to photon pairs in 139 fb^{-1} of pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, ATLAS-CONF-2020-037, 2020. URL: https://cds.cern.ch/record/2727744.

[19] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. **C71** (2011) 1554, [Erratum: Eur. Phys. J.C73,2501(2013)], arXiv:1007.1727.

[20] ATLAS Collaboration. *Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, ATLAS-CONF-2015-081, 2015. URL: https://cds.cern.ch/record/2114853.

[21] ATLAS Collaboration. *Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, Phys. Rev. D **90** (2014) 112015, arXiv:1408.7084.