Rebuttal

The Authors appreciate the useful comments of the reviewers and Editors. In the revised manuscript all the reviewers‘ comments have been addressed, the typos amended, text proof read and grammar checked. Changes from the previous version have been highlighted in red colour.

We hope they will find the new manuscript enhanced in quality and readability compared to the previous version.

Reviewer #1:

Reviewer, issue 1: *Bayesian updating has been used for fatigue crack evaluation and detection in Ref [1,2]. Moreover, the idea of combining XFEM for crack location identification has been reported in Ref [3]. Authors should show significant and clear improvement and difference comparing with existing literatures.*

[1] Model selection, updating, and averaging for probabilistic fatigue damage prognosis
[2] A probabilistic crack size quantification method using in-situ Lamb wave test and Bayesian updating
[3] Nondestructive identification of multiple flaws using XFEM and a topologically adapting artificial bee colony algorithm

More comprehensive literature review should be included in the introduction part to explicitly show the novelty of the proposed method.

Authors Reply: References [1, 2, 3] are now included within the manuscript introduction. The main novelty, differences and similarity of the proposed method are pointed out. In the revised manuscript (starting from line 49) the references are reviewed and limitations addressed:

‘‘Authors in Ref.[1] proposed a novel Bayesian updating approach for fatigue damage prognosis employing the so-called reversible jump Markov chain Monte Carlo. The framework can account for uncertainties and two simple crack growth model were analysed. However, computational time issues typical of these type of frameworks were not explicitly discussed. Similarly, the Authors in [2] proposed a Bayesian updating method for crack size quantification and using Lamb wave signals. The method was effective for damage prediction but problems of efficiency are not mentioned. H. Sun et al. [3] proposed an updating framework for multi flaws identification, based on extended finite element method and adapting artificial bee colony algorithm. A parametric study of the noise uncertainty was also proposed. The computational time was an issue and the author briefly discuss a hypothetical solution which consists in run the analysis in parallel on a compute cluster. .... The vast majority of the reviewed works did not account for efficiency in the computations at the same time providing an indicator of the imprecision surrounding the analysis. Furthermore, none of the reviewed papers assessed the robustness of the Bayesian updating procedure when different likelihood functional expressions were employed.‘‘

Reviewer, issue 2: *What is the advantage for using ANN model, comparing with more commonly used regression method?*

Authors Reply: Neural networks have been used because are, in principle, universal approximators capable of dealing with nonlinearities automatically and are easy to apply to a variety of problems (two in our case).

In the revised paper (lines 100-105):
‘‘ANNs have been used in this work because are flexible and, in principle, universal approximating functions. This makes their use and implementation relatively easy and for a variety of different applications. Compared to traditional regression methods, they allow nonlinearity to be captured automatically and are generally quite fast to implement. Different traditional regression methods could have been employed if adequately trained to reproduce the model input output relation.‘‘

Reviewer, issue 3: *Results (e.g. Fig. 14, 15, 18 ,19) shown that false detection is a major problem. Authors should give some suggestions and the possible solutions for this problem.*

Authors Reply: Thank you for this comment. Fig. 13, 14 shown a false detection problem, whilst Figs. 18-19 are rather missed detections. Specifically, Fig.18 and 19 summarise the result the analysis on increasing level of measurement noise and Young’s modulus uncertainty. In both cases the real problem is a miss-detection if we move toward a higher level of uncertainty, in fact, the confidence intervals tend to become non-informative (i.e. the percentile intervals for all the positions tends towards [0, max crack length]). This is indeed a problem which can perhaps be solved by providing better and more data (see the new section 8.3, convergence analysis) and by effective and efficient noise filtering systems.

For what concerns the false detections, hypothetical solutions to this issue are to collect more experimental data, improve the fidelity of the original FEM and the surrogate. Finally, improve the quality and quantity of data and test different likelihoods and different prognostic indicators can help in achieving better detections (i.e. better distinguish between different cracks positions and lengths).

This discussion has been summarised in the revised manuscript (see lines from 626 and section 8.3)

Reviewer #2:

Reviewer, issue 1: *I first collect some typos and minor points. Although they do not seem to affect reading, it reflects that the paper is not written rigorously and carefully. Hope the author can check it more carefully.*

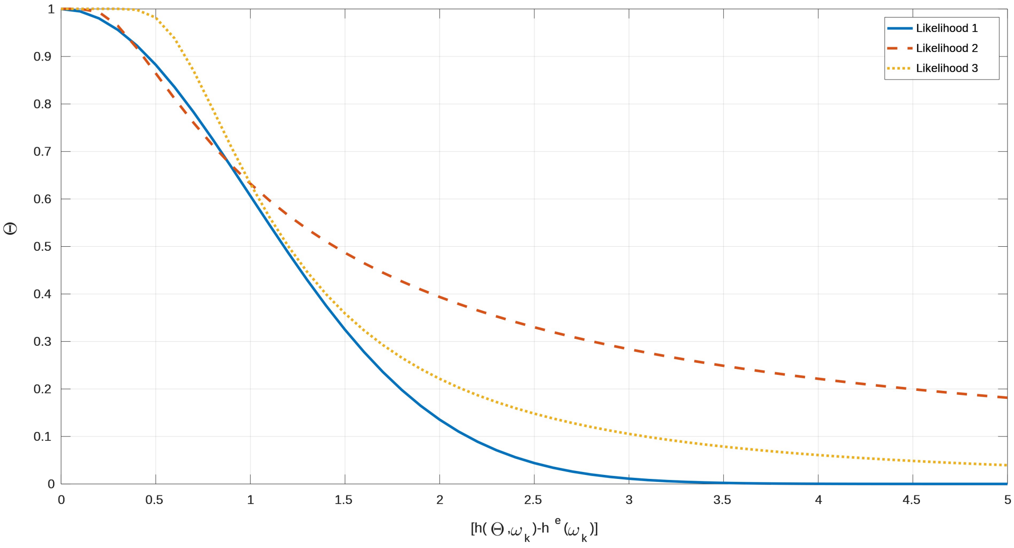
Authors Reply: Typos have been amended and text poof read.

Reviewer, issue 2: *the authors use the numerical likelihoods to compare the experimental frequency response with the simulations frequency response. This arises a question that whether the proposed three likelihoods are novel or not. How did the authors construct the likelihoods, are they generalized for most of Bayesian updating procedure? If the likelihoods are obtained from the former studies, is there any theoretical performance evaluation for these three likelihoods?*

Authors Reply: Thank you for this comment, perhaps there was a lack of clarity in the text. The 3 proposed likelihoods were newly defined by the authors to simply test the robustness of the procedure when data was embedded in the analysis differently.

For clarity, a new paragraph and figure have been included (lines 218-227)

‘The different likelihoods mathematical expressions are proposed on an empirical basis and used to test the detection robustness when the experimental data is encoded differently within the procedure. For clarity, the likelihood in equations 5, 6 and 7 are displayed in Figure 1 by solid, dashed and dotted lines, respectively. It can be observed that likelihood 1 decreases more rapidly than likelihood 2 and 3 for an increasing discrepancy between model and experiment. This means (from an intuitive point of view) that likelihood 1 will provide as likely only model that well-explain the data, i.e. which provide a small $[h({\theta},\omega_k)-h^e(\omega_k)]$. On the other hand, likelihood 3 will indicate as plausible also models resulting in higher discrepancies between simulated vibrational responses and the experimental data.data.‘



Reviewer, issue 3: *Secondly, in the Bayesian procedure for on-line damage detection, the authors compute the likelihood function using surrogated ANN model instead of high-fidelity model and compared with experimental response. However, it may be necessary to consider the difference between FEM model and experiments. Although the surrogate model is accurate, it is constructed only based on the FEM model. The error is likely caused by the difference between FEM and experiments. Did the author take part of experimental data into construction or verification of the surrogate model? Specially, the crack detection is not easy to simulate even if the XFEM techniques are well employed. The authors also need to ensure the likelihood difference are not from the error between surrogate model and experiments.*

Authors Reply: This is indeed an important point which we discussed in sections 8.1 and 8.2. Mainly due to budget constraints, we didn‘t have access to real experimental data (it would have been necessary to obtain vibrational responses from many healthy and cracked car suspension arms). We also agree with the reviewer, simulate cracks is not easy. Due to the deterministic nature of the XFEM, we did not expect it to perfectly reproduce the behaviour of a device crossed by random crack or fractures. It is for these reasons that we decided to account for random noises of different intensity and for uncertainty in Young`s modulus.

A positive aspect of the proposed framework is that it can be employed to repeat virtual experiments in a controlled environment. This is a necessary phase to rigorously test and assess the efficiency and robustness of the method.
Differently, Case study A employed real experimental measurements from an aluminium frame, which was used to further test the goodness of the approach when real experimental evidence was employed.
In the revised manuscript (from lines 576):

In the previous crack detection cases, the crack parameters (length and position) have been considered affected by epistemic uncertainty. The procedure detects the most plausible crack parameters of the XFEM accordingly to the experimental evidence. The procedure was efficient and effective, nevertheless, the employed XFEM is a deterministic solver and as such, it will unlikely behave like a real structure crossed by random cracks. Thus, to further test and prove the effectiveness of the proposed detection procedure, additional layers of uncertainty have been analysed. The crack detection updating case I and II have been performed by adding noises to the synthetic experimental FRFs. The analysis is then followed by an uncertainty propagation of the imprecisely known material proprieties of the cracked car component. This will serve to test the goodness of the framework for increasing discrepancy between XFEM and experimental evidence.

Reviewer, issue 4: *The authors point out that 103 samples are generated to construct the surrogated model, but the 103 samples do not well explore the parameter space. In fact, it is easy to generate well space-filled samples using Latin Hypercube Sampling or other variance reduction based sampling methods. Here the reviewer uses the matlab function lhsdesign to generate 103 sample in terms of the mass vertical positions (Fig.).*

Noticed that the 103 samples in this paper are actually obtained from the ref[17], but the authors have already known the bad space-fill characteristics of these samples, why not resampling? Compared with the 103 samples in this paper, we can found the surrogate model is highly unacceptable in the boundary of parameter space, for example, pm >30 or pm <10. From the viewpoint of reviewer, the second application is good enough to demonstrate and illustrate the applicability of the approach, hence the first example can be taken out from the paper unless the author wants to further revise and improve it. Otherwise, the current version seems not to be well prepared for readers.

Authors Reply: The reason why we included case study A was to test the framework when real-life experimental data were employed. In fact, we were able to run just virtual experiments on case study B, mainly due to budget limitations. We fully agree with the reviewer regarding the metamodel comment. In fact, we were expecting the ANN to perform poorly in the boundary region (the only samples on the boundaries were 5,5 and 35,35). A discussion on the issue was synthetically presented in the earlier manuscript and extended in the revised versions.

The main reason why we did not resample was to test the goodness of the method when the meta-model was (likely) not well-reproducing the data. As consequence, the attempted prediction close to the not-well-explored region of the parameter space was characterised by wider credibility intervals and they were less accurate. Another reason to not resample is the unavailability of the updated FE model but only the parameters-output samples.

Reviewer, issue 5: *Finally, the reviewer cares about the issue from limited experimental data. Both two examples use only limited experimental data to verify the likelihood and updating procedure. The reviewer understands the time cost of data collection, but is that too small, such as only five experimental data for the first example? Because the small data may cause the issue of imprecise probability, did the authors ever consider this problem?*

Alternatively, the conclusion for crack detection will change as experimental dataset size grows? As we know, the limited data may probably not cover the real response space such that the prediction may be only effective in some local regions.

Furthermore, the reviewer also suggests that the author can explore the convergence of cracks detection when more data are collected. That is also to say, the proposed approach can achieve that monitoring and prediction are being more accurate with the increasing of experimental data.

Reply: A brief introduction to imprecise probability is now proposed (lines 61-70 and 92-94). Problems of imprecise probability are indeed relevant here, in fact, the employed health indicators (e.g. Natural frequencies and peak FRFs) were extracted from just a single experiment.

With the goal of providing a measure of the imprecision surrounding the analysis, we provided an interval-valued indicator for the posteriors (the [5-th - 95-th] percentile credibility interval). The interval-valued indicators provide, at least, an idea of what is the level of data scarcity surrounding the analysis. If the updating is affected by imprecision (e.g. lack of data), the percentile interval will be wider and in extreme cases close to non-informative results (i.e. interval [0 maximum crack length]). This is confirmed in the new section 8.3 (pp 28) where the convergence of the procedure is discussed for increasing availability of data. Specifically, the posterior percentile intervals and mean value are assessed for increasing number of experiments and when an increasing number of health indicators are extracted from each experiment.