Dear Referee #1,

We would like to thank you for your constructive comments. We agree with most of the suggestions and, therefore, we will modify the manuscript to take on board your comments. In the following, we recall these comments and we reply to each of the comments in turn (outlined by "**<Authors' reply>**").

Please note that the line numbers of changes are indicated and correspond to the revised manuscript with marked changes.

Referee #1:

1. General comments

The authors present in this manuscript a methodology to derive low probabilities of failure for a nuclear plant, based on a simplified numerical model, by fitting a statistical distribution to the response. The paper propose several non-linear models to link the response to the different covariates and some model selection to derive the best estimation of failing probability, called here Fragility Curve.

The paper well expose the models used, however some of them could be better explained, and the results when the covariates uncertainties are taken into account are well presented. In comparison, the description of the construction of the database is less clear to me, and as is would be difficult to reproduce.

The paper is well written, with relevant references and good quality figures. The methods used, if not the newest, have not been already used in the domain, as far as I know. The application is sensible and realistic. The problem addressed is worth being published.

2. Specific comments

2.1 Statistical methods

The description of variable selection method is rather crude and could be better explained. For example, the double-penalty procedure is not presented, and would better serve the paper than the description of the GEV distributions.

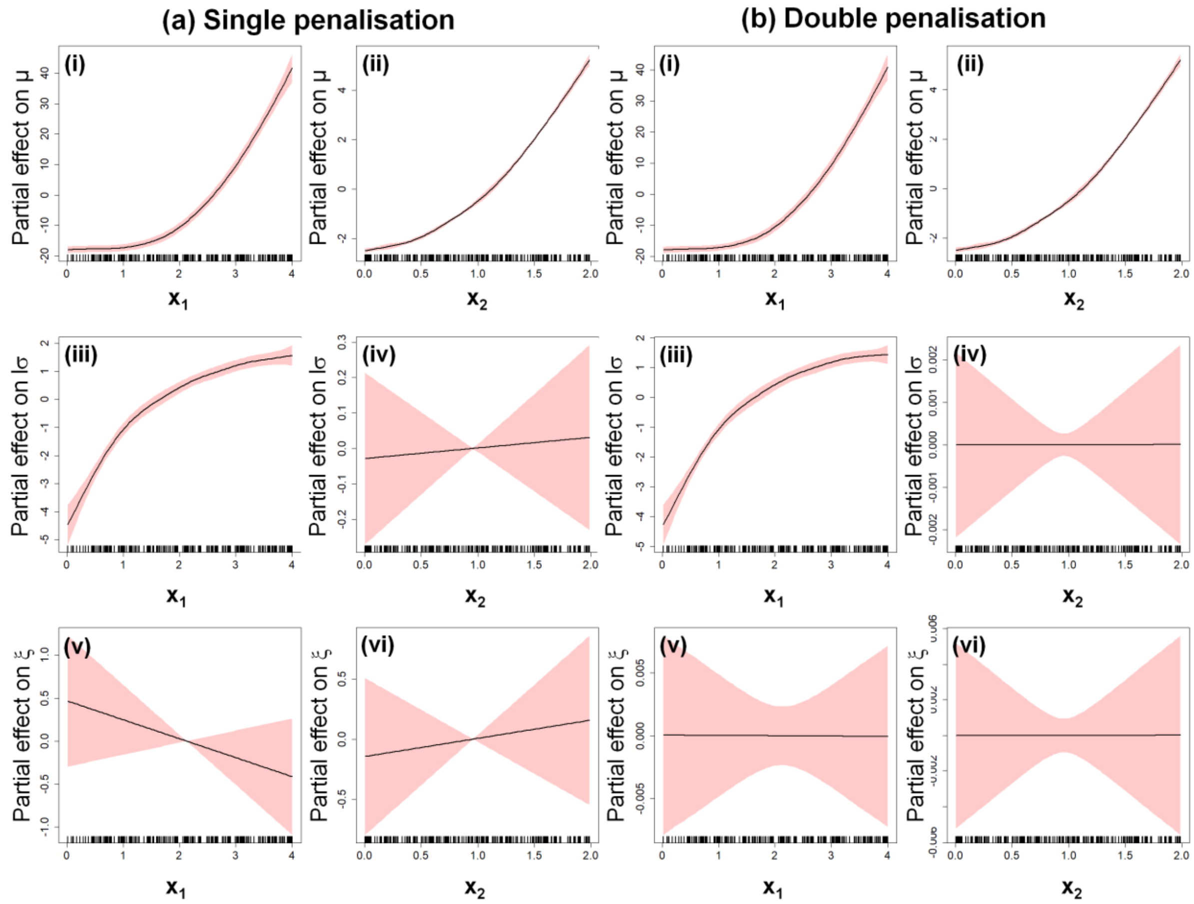
<Authors' reply> We agree with referee #1 and have elaborated on this aspect by providing a complementary appendix. Besides, Sect. 2.4 has been completed with an application on a synthetic test case as follows:

"To exemplify how the procedure works, we apply it on the following synthetic case.

Consider a nonstationary GEV distribution whose parameters are influenced by two covariates x_1 and x_2 (see Eq. 6) as follows:

$$f_{\mu}(x) = {x_1}^3 + 2.{x_2}^2 + 1$$
$$f_{l\sigma}(x) = {x_1}^2 \quad , \qquad (6)$$
$$f_{\xi}(x) = -0.1$$

A total of 200 random samples are generated by drawing x_1 and x_2 from a uniform distribution on [0; 4] and [0; 2] respectively. Fig. 2a provides the partial effects for the synthetic test case using the single penalisation approach. The non-linear relationships are clearly identified for μ (Fig. 2a-i,ii) and for $l\sigma$ (Fig. 2a-ii). Yet, the single penalisation approach fails to identify properly the absence of influence of x_2 on $l\sigma$ and of both covariates on ξ (Fig. 2a-iv,v,vi) since the resulting partial effects still present a linear trend (though with large uncertainty bands). Fig. 2b provides the partial effects using the double penalisation approach. Clearly, the penalisation achieves to identify the absence of influence (Fig. 2b-iv,v,vi) as well as the nonlinear partial effects for μ (Fig. 2b-i,ii) and for lσ (Fig. 2b-ii)".



New Figure 2: Partial effect for the synthetic test case using the single penalisation approach (a) and the double penalisation approach (b).

Moreover, it could be interesting to compare the results with a dedicated variable selection algorithm such as boosting for example (e.g. with gamboostLSS package). As is, it is difficult to understand how the selection if done and in particular how variables are excluded from the figures 10 and 11.

<Authors' reply> The current version of gamboostLSS package does not consider the GEV distribution and adding these new functionalities to this specific package is out of the scope of the current study. We however agree that mentioning alternative fitting (and variable selection) approaches should be added to the manuscript as future lines of research.

Besides, to bring additional elements to referee #1 (and out of curiosity), we applied the gamboostLSS procedure by randomly generating 200 observations from a nonstationary Gumbel distribution considering the following relationships:

$$f_{\mu}(x) = {x_1}^3 + 2.{x_2}^2 + 1$$

$$f_{l\sigma}(x) = {x_1}^2$$

The following figure provides the comparison between the partial effect for lσ derived from

(a) the double-penalisation-based fit as proposed in the present work;

(b) the boosting-based fit (using a 5-fold cross validation procedure combined with the noncyclical algorithm by Thomas et al. (2018) for selecting the stopping boosting cut-off).

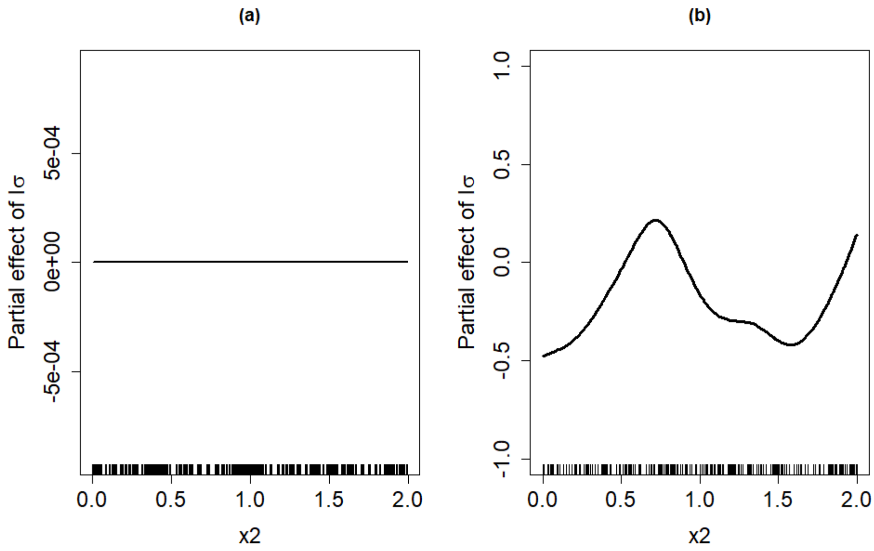


Figure. Partial effect for the synthetic Gumbel test case using the double penalisation approach (a) and the gamboostLSS approach (b).

Both approaches achieve to identify the negligible influence of x_2 on lσ, but the magnitude of the influence remains small-to-moderate for gamboostLSS and highly dependent on selection of the stopping boosting cut-off. On this aspect, the double-penalisation procedure appears to be more robust. This should however be confirmed in a more extensive benchmark exercise that could be a line for future research of the present work.

Reference

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., & Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. Statistics and Computing, 28(3), 673-687.

I also have some concerns about the model selection : since here the authors are not interested in predicting new values, are AIC and BIC the best selection criteria to use ? In particular, for an explanatory model, the QQ plots can be a better tool and may leeds to different conclusion. For example, in the case of parametric uncertainty, I would go for the Gumbel model (figure 9). Could the authors precise why the use AIC and BIC in this case and how could they go further ?

<Authors' reply> The use of AIC,BIC criteria is guided by best practices in the domain of nonstationary extreme value analysis (e.g., Kim et al., 2017; Salas and Obeysekera, 2014), and more particularly recommended for choosing among various fragility models (e.g. Lallemant et al., 2015); see also an application of these criteria in the domain of nuclear safety by Zentner (2017). We agree however with referee #1 that further explanations should be given regarding model selection based on information criteria, because the perspectives differ when using AIC or BIC:

- On the one hand, AIC-based analysis considers a model to be a probabilistic attempt to approach the infinitely complex data-generating truth – but only approaching not representing (Höge et al. 2018: Table 2). This means this type of analysis aims at addressing which model will best predict the next sample, i.e. it provides a measure the predictive accuracy of the different models (Aho et al., 2014: Table 2);

- On the other hand, the purpose of BIC-based analysis considers each model as a probabilistic attempt to truly represent the infinitely complex data-generating truth assuing that the true model exists and is among the candidate models (Höge et al. 2018: Table 2). This perspective is different from the one of AIC and focuses in an approximation of the marginal probability of the data (here lEDP) given the model (Aho et al., 2014: Table 2); hence giving some insights to address which model generated the data, i.e. it measures goodness of fit.

Testing both criteria, AIC or BIC, thus provides both visions on the problem of model selection. This will be more clearly described in a new sub-sect. 2.2 entitled “Model Selection.”

Regarding the comment on QQ plot, we agree that it can be used to validate model with the goal of explaining the observations. Yet, we only partly agree with referee #1 about its role for model selection. In current practices, QQ plots are rather used for model checking and not model selection, i.e. to control the model fit by examining the residuals once the model has been selected, and to identify why the model is adequate (or not). Its effectiveness has clearly been shown in statistical literature, but when restricted to visual inspection (see e.g. Loy et al., 2016). Plotting the relative differences between the theoretical quantiles and the ones given by the Gaussian and the GEV models may be envisaged to improve the identification of the discrepancies. In addition, we propose to complement this diagnostic by the analysis of the PP plot as well. This will enable us to better emphasize the goodness of fit for large quantile levels. See below an example of presentation.

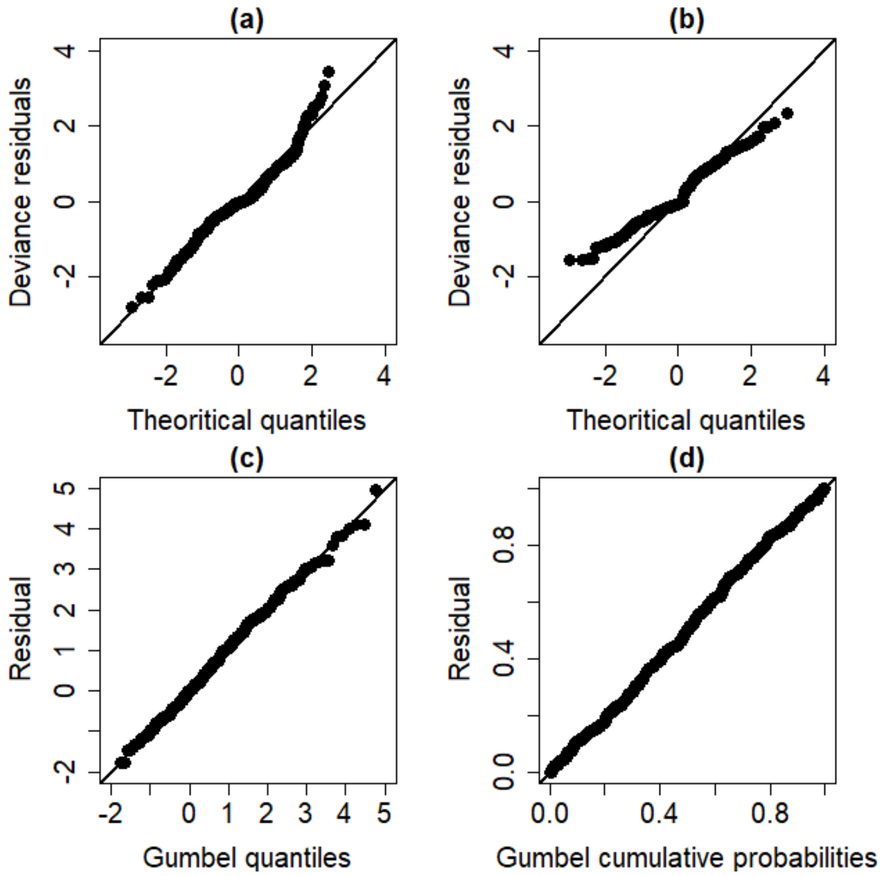


Figure: Diagnostic plots to check the validity of the considered model: (a) QQ plot for the deviance residuals for the NOsmo2 model; (b) QQ plot for the deviance residuals for the

GEVsmo2 model with epistemic uncertainty; (c) QQ plot on Gumbel scale; (d) PP plot on Gumbel scale.

It should also be underlined that using QQ(or PP) plots only account for one part of the problem of model selection, i.e. goodness of fit. It does not account for the complexity of the considered model contrary to information criteria like BIC (or AIC). The advantages of information criteria is to include the first aspect (or predictive capability when using AIC) but also a correction related to the complexity of the model; here provided by the number of model parameters. Bayesian information criterion generally penalizes more complex models more strongly than does the AIC.

Regarding the comment on the selection of the Gumbel model, we agree with referee #1. This aspect was outlined in the original version of the manuscript as follows: "The estimated shape parameter reaches here a constant value of 0.07 (+/-0.05), hence indicating a behaviour close to the Gumbel domain". This result could be seen as an additional element supporting the flexibility of the proposed approach based on GEV, which encompasses the Gumbel distribution as a particular case.

<u>**References**</u>

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. Ecology, 95(3), 631-636.

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. Water Resources Research, 54(3), 1688-1715.

Kim, H., Kim, S., Shin, H., & Heo, J. H. (2017). Appropriate model selection methods for nonstationary generalized extreme value models. Journal of Hydrology, 547, 557-574.

Lallemant, D., Kiremidjian, A., & Burton, H. (2015). Statistical procedures for developing earthquake damage fragility curves. Earthquake Engineering & Structural Dynamics, 44(9), 1373-1389.

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of Q–Q Plots: The power of our eyes!. The American Statistician, 70(2), 202-214.

Salas, J. D., & Obeysekera, J. (2014). Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. Journal of Hydrologic Engineering, 19(3), 554-568.

Zentner, I. (2017). A general framework for the estimation of analytical fragility functions based on multivariate probability distributions. Structural Safety, 64, 54-61.

<u>***2.2 Application case***</u>

The selection of the ground-motion records if not described precisely enough from my point of view, for example the scaling levels are not stated. E.g. return levels for with quantity ? The records are non-linear and non-stationary in time, so how the spectrum is computed and scaled? The computational time for running seems to be omitted, it might be interesting to give an idea if a more important database could be generated.

<Authors' reply> We thank the reviewer for his/her interest in the inner workings of the ground-motion selection procedure: we agree that the conditional spectrum approach used here is currently not sufficiently described. Therefore, we propose to replace the initial text in lines 212-222 by the following text:

"Thanks to the consideration of reference earthquake scenarios at various return periods, the scaling of a set of natural records is carried out to some extent, while preserving the consistency of the associated response spectra. The steps of this procedure hold as follows:

- *Choice of a conditioning period*: the spectral acceleration (SA) at T^* = 0.38s (fundamental mode of the structure) is selected as the ground-motion parameter upon which the records are conditioned and scaled.
- *Definition of seismic hazard levels*: six hazard levels are arbitrarily defined, and the associated annual probabilities of exceedance are quantified with the OpenQuake engine (www.globalquakemodel.org), using the SHARE seismic source catalogue (Woessner et al., 2013), for an arbitrary site in Southern Europe. The GMPE from Boore et al. (2014) is used to generate the ground motions, assuming soil conditions corresponding to $V_{s,30}$ = 800 m/s at the considered site. Data associated with the mean hazard curve are summarized in Table 2.

New Table 2: Estimation of the seismic hazard distribution for the application site.

Scaling level	SA(0.38s) [g]	Annual Probability of Exceedance	Return Period
#1	0.185	4.87E-2	20 y
#2	0.617	4.99E-3	200 y
#3	0.836	2.50E-3	400 y
#4	1.492	5.00E-4	2,000 y
#5	2.673	5.00E-5	20,000 y
#6	3.882	5.00E-6	200,000 y

- *Disaggregation of the seismic sources and identification of the reference earthquakes*: the OpenQuake engine is used to perform a hazard disaggregation for each scaling level. A reference earthquake scenario may then be characterized through the variables [M_w; R_{jb}; ε] (i.e., magnitude, Joyner-Boore distance, error term of the ground-motion prediction equation), which are averaged from the disaggregation results (Bazzurro & Cornell, 1999). This disaggregation leads to the definition of a mean reference earthquake (MRE) for each scaling level.
- *Construction of the conditional spectra*: for each scaling level, the conditional mean spectrum is built by applying the GMPE to the identified MRE. For each period T_i, it is defined as follows (Lin et al., 2013):

$$\boldsymbol{\mu_{\ln SA(T_i)|\ln SA(T^*)} = \mu_{\ln SA}(M_w, R_{jb}, T_i) + \rho_{T_i,T^*} \cdot \varepsilon(T^*) \cdot \sigma_{\ln SA}(M_w, T_i)} \quad (7)$$

where $\mu_{lnSA}(M_w,R_{jb},T_i)$ is the mean output of the GMPE for the MRE considered, $\rho_{Ti,T*}$ is the cross-correlation coefficient between $SA(T_i)$ and $SA(T^*)$ (Baker & Jayaram, 2008), $\varepsilon(T^*)$ is the error term value at the target period T^* = 0.38s, and $\sigma_{lnSA}(M_w, T_i)$ is the standard deviation of the logarithm of $SA(T_i)$, as provided by the GMPE. The associated standard deviation is also evaluated, thanks to the following equation:

$$\boldsymbol{\mu_{\ln SA(T_i)|\ln SA(T^*)} = \mu_{\ln SA}(M_w, R_{jb}, T_i) + \rho_{T_i,T^*} \cdot \varepsilon(T^*) \cdot \sigma_{\ln SA}(M_w, T_i)} \quad (8)$$

The conditional mean spectrum and its associated standard deviation are finally assembled in order to construct the conditional spectrum at each scaling level. The conditional mean spectra may be compared with the uniform hazard spectra (UHS) that are estimated from the hazard curves at various periods. As stated in Lin et al. (2013), the SA value at the conditioning period corresponds to the UHS, which acts as an upper-bound envelope for the conditional mean spectrum.
- *Selection and scaling of the ground-motion records*: ground-motion records that are compatible with the target conditional response spectrum are selected, using the

algorithm by Jayaram et al. (2011): the distribution of the selected ground-motion spectra, once scaled with respect to the conditioning period, has to fit the median and standard deviation of the conditional spectrum that is built from Eq. 7 and 8. The final selection from the PEER database (PEER, 2013) consists of 30 records for each of the 6 scaling levels (i.e., 180 ground-motion records in total).

The non-linear dynamic analyses are performed on a high performance-computing cluster, enabling the launch of the multiple runs in parallel (e.g., a ground-motion of a duration of 20s is processed in around 3 or 4 hours). Here, the main limit with respect to the number of ground-motion records is not necessarily related to the computation cost, but more to the availability of natural ground motions that are able to fit the conditional spectra at the desired return periods."

References:

Baker, J.W., & Jayaram, N. (2008). Correlation of spectral acceleration values from NGA ground motion models. Earthquake Spectra, 24(1), 299-317.

Bazzurro, P., & Cornell, C.A. (1999). Disaggregation of seismic hazard. Bulletin of the Seismological Society of America, 89(2), 501-520.

Jayaram, N., Lin, T., & Baker, J.W. (2011). A computationally efficient ground-motion selection algorithm for matching a target response spectrum mean and variance. Earthquake Spectra, 27(3), 797-815.

PEER (2013). PEER NGA-West2 Database, Pacific Earthquake Engineering Research Center, https://ngawest2.berkeley.edu.

Woessner, J., Danciu, L., Kaestli, P., & Monelli, D. (2013). Database of seismogenic zones, Mmax, earthquake activity rates, ground motion attenuation relations and associated logic trees. FP7 SHARE Deliverable Report D6.6.

2.3 Results

The models compared here do not include parametric models (polynomials, nonlinear...) and the selected models are the non-linear smooth models. One question is related to the ability of this models to extrapolate beyond the range of variation of the training set? It might be interesting to compare to classical parametric models (if any) of with some polynomials models to also investigate the extrapolation ability.

<Authors' reply> Regarding the problem of extrapolation ability, though of interest, we are not fully convinced that this is within the scope of our study.

- Considering the study without uncertainty, the selection of the ground-motion records is performed in order to cover a wide range of plausible earthquake scenarios (hence of PGA) in particular by considering an upper bound for PGA of 30 m/s^2 (i.e. ~3g) for return periods up to large values of 200,000 years (see reply to comment 2.2). Extrapolating outside this upper bound may not be considered physically realistic;

- Considering the study with uncertainty, the mechanical and geometrical parameters are here bounded and extrapolating outside the considered range may suffer from a lack of realism as well. In a more generic case, for instance with uncertainties represented by unbounded probability distributions, the problem could however appear. This aspect is now clearly outlined in the discussion section Sect. 5.

We however totally agree with referee #1 that the problem of extrapolation is more stringent when nonstationary is related to temporal covariates as outlined for instance by Salas and Obeysekera (2014).

Including a larger number of parametric models may be of interest, but the proposed models should remain realistic. Current practices in seismic vulnerability analysis mostly focus on simple linear models, because in most situations data suggest it and because they remain interpretable. Models of intermediate complexity like polynomial models of second order are rarely used. We choose not to include them in the current analysis.

In order to provide some elements to referee #1, the following figure shows that a second-order polynomial GEV model (denoted GEVpoly1,2 whether it is applied on μ or σ) is not identified by the AIC/BIC analysis as the "most appropriate" model (compared to the smooth GEV).

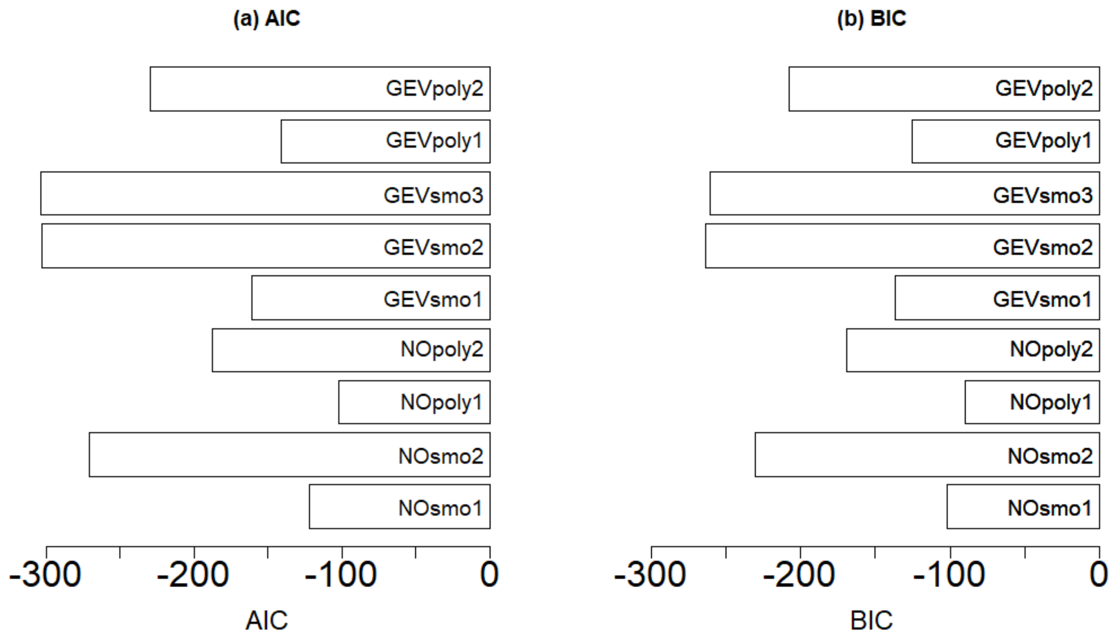


Figure: Model selection criteria (AIC (a) and BIC (b)) for the different models considering the derivation of a FC without parametric uncertainty.

In current practices, when non-linearity is suspected, more complex non-parametric models are generally preferred based for instance on neural networks (see e.g., Wang et al., 2018) or on kernel smoothing (see e.g. Mai et al., 2017) because they enable to derive from the data the non-linearity by avoiding to specify the form/shape of the non-linearity. This is also the advantage of the proposed approach. Comparison to these alternatives is here out of the scope of the present study and we choose to underline this perspective in the discussion section Sect. 5.

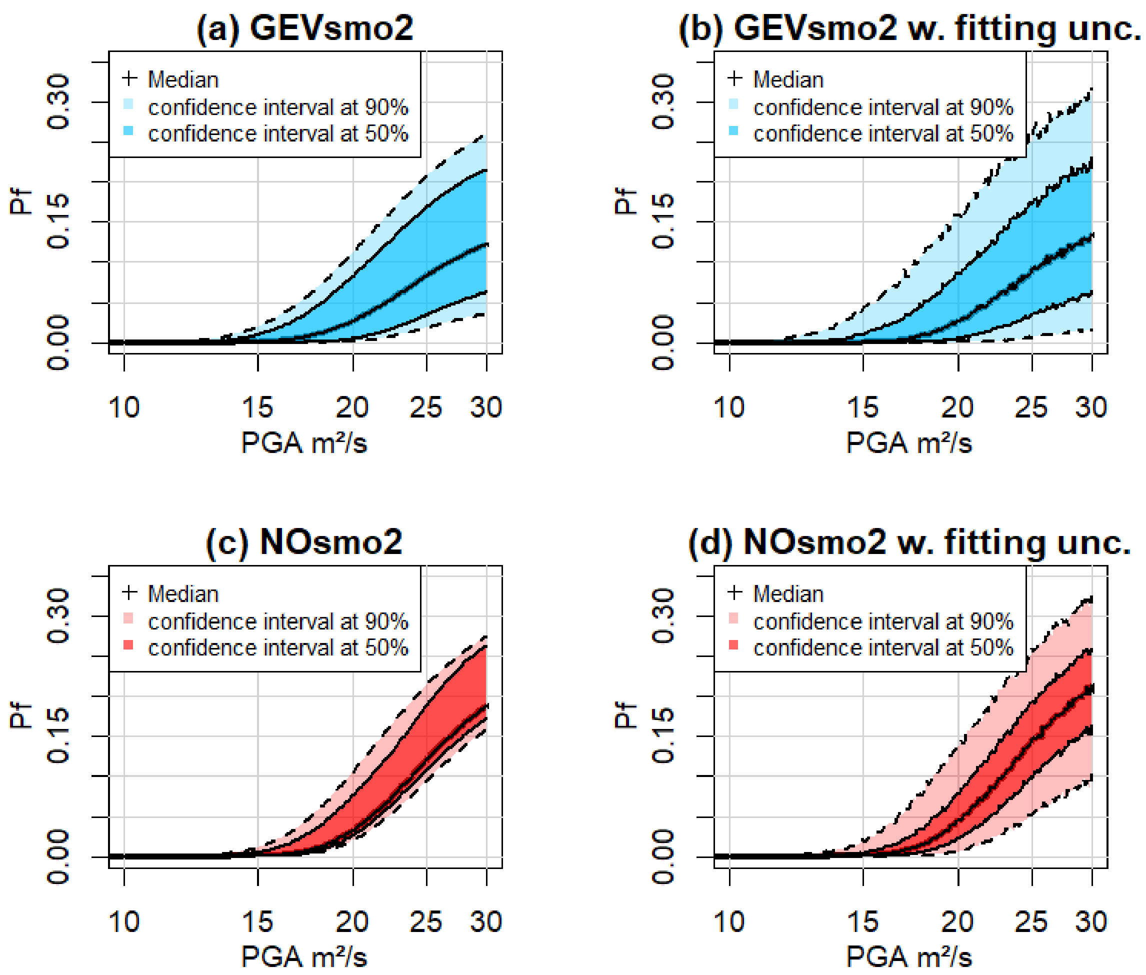
If my understanding is correct, the uncertainties in the estimation of the marginal effects are neglected in computing the fragility curves, that is the reason why the are no uncertainties on figure 7. However, in figures 12 and 13, uncertainties linked to the variability of the input variables are shown. As is, it difficult to know which source of uncertainties is the highest and a discussion on this point would add a great value to the paper.
<Authors' reply> We totally agree with the referee #1 and have completed the analysis by incorporating the uncertainties on the regression coefficients (Eq. 5). The following procedure is conducted to account for the mechanical and geometrical uncertainties:

- Step 5.1: the considered *IM* is fixed at a given value;

- Step 5.2: for the considered *IM* value, a large number (here chosen at *n*=1000) of samples of ***U*** are randomly generated;
- Step 5.3: for each of the randomly generated ***U*** samples, the failure probability is estimated for the considered *IM* value;
- Return to step 5.1.

The result of the procedure corresponds to a set of *n* FCs from which we can derive the median FC as well as the uncertainty bands based on the pointwise confidence intervals at different levels. These uncertainty bands thus reflect the uncertainty on the mechanical/geometrical parameters. This procedure can be extended by accounted also the uncertainty on the fitting of the probabilistic model (e.g., GEV or Gaussian) by randomly generating the corresponding model parameters at step 5.2 (by assuming that they follow a multivariate Gaussian distribution).

In Sect. 4, we now discuss in more details two cases: (1) with uncertainty on the geometrical/mechanical parameters only (termed as "epistemic uncertainty"); (2) with the uncertainty on the fitting as well (i.e. uncertainty on the model parameters). See below a possible presentation of the results (new Figure 13). Such presentation will be used to compare the implications of both types of uncertainty.



New Figure 13. Fragility curve (relating the failure probability P_f to *PGA*) considering epistemic uncertainties only (left), and fitting uncertainty as well (right). (a,b) GEV-based FC;

(c,d) FC based on the normal assumption. The coloured bands are defined based on the pointwise confidence intervals derived from the set of FCs (see text for details).

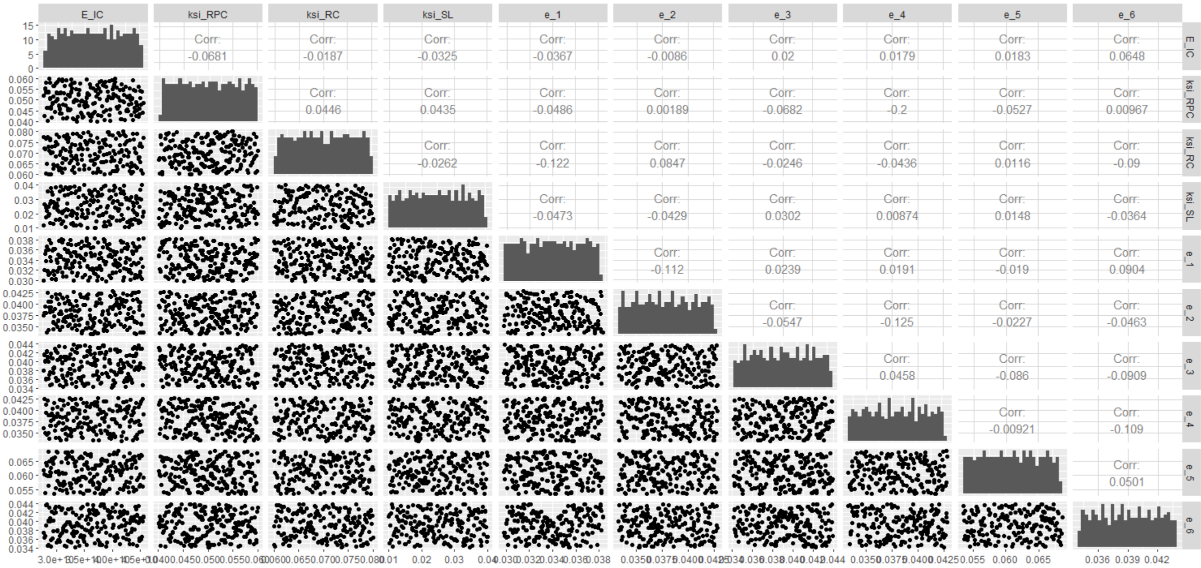
3. Technical remarcks

- Both formula, figures and tables should be centered to be easier to read; - in figures 10 and 11,

<Authors' reply> We apologize for this lack of readability but we followed the instructions provided in the word template of NHESS where the equations, tables and figures are formatted as "justified" (https://www.natural-hazards-and-earth-system-sciences.net/for_authors/manuscript_preparation.html).

- some variables seems to be evenly distributed and some other (e.g. E_IC, nXi_RC, e5 in figure 10) seems to be random : it seems that all of them should be uniform of the range of variation stated in Table 1 ?

<Authors' reply> We checked this aspect (see figure below) and it seems that it is only a visual effect related to the parametrisation of the ticks on Figure 10.



- The link functions are not stated precisely in table 2;

<Authors' reply> This is now specified.

March 9th, 2020

Jeremy Rohmer[1], Pierre Gehl[1], Marine Marcilhac-Fradin[2], Yves Guigueno[2], Nadia Rahni[2], Julien Clément[2]

[1]BRGM, 3 av. C. Guillemin, 45060 Orléans Cedex 2, France

[2]Institute for Radiological Protection and Nuclear Safety, Fontenay-Aux-Roses, 92262, France