We would like to thank the Reviewers both for their interest in our work and for carefully reading our manuscript; we greatly appreciate the insightful comments as they may contribute to increase the manuscript robustness and, in general, to improve its quality and readability. In the following, we supply a point by point reply to the general and specific comments raised by the Reviewers.

Reviewer 1:

R1-C1: There are a few areas where the grammar or writing could be polished although the sense is always clear
Answer: In revising the manuscript we will take care of polishing the grammar and writing.

R1-C2: Figure 1: It would be interesting to see the maximum flood depths plotted on the map, so as to get a sense of how the flood varied across the flood zone.
Answer: We agree, so we will modify Figure 1 by showing the flood depth map for the 2002 event.

R1-C3: Do the authors have any information on the flood map and how closely it matched the flood events that occurred. This is obviously a potentially large source of uncertainty in the results.
Answer: This information is described in the paper quoted in the manuscript (Scorzini et al. 2018), however when revising the manuscript, we will include some additional details on the hydraulic modelling of the event. For the 2002 flood, information on observed water depths was available in more than 260 points within the inundated area, deriving from indications provided by municipal technicians and by citizens in the damage compensation forms, as well as from interpretation of photographs taken during or immediately after the event. These data were used in the validation of the 2D hydraulic model; the resulting average absolute differences between observed and calculated water depths within the inundated area range from 0.2 to 0.4 m, depending on the zone. This is surely a possible source of uncertainty, however, as also known from the literature, resulting differences could be considered to provide relatively small impacts on the overall damage estimation. Moreover, as the same hazard data were used for the implementation of all damage models, this would be a common source of uncertainty for the application of all our models that should not condition the results presented in the paper.

R1-C4: In Table 1, where there is a set of discrete responses, is it possible to see what these choices are (e.g. level of maintenance: Low, medium, high, etc).
Answer: In revising the manuscript, we will include in Table 1 the missing description of the parameter “level of maintenance”, by including the three possible choices (low, medium, high).

R1-C5: Table 2 and associated text. Is it possible to give more detail on the adjustments needed to each model to get them to work. This could be due to my misunderstanding, but for example, taking the model of Jonkman et al, two variables are needed (h and FA). However, is the replacement value also needed?
Answer: Yes, in the revised version of the manuscript we will include in Step 2 of the “Methodology” some details on the adjustments needed for the models to work.
We will also make Table 2 more clear by adding the parameter “economic value” (the type of economic evaluation is already shown in the fifth column of Table 2) among the explicative variables for all the damage models (except for CEPRI, which is an absolute damage model).

R1-C6: Figure 3 - could the authors explain why there are buildings with no damage? Are these the buildings discussed in Section 3.4?
Answer: No, they are not. The zero damages are due to the specific assumptions behind the damage models (e.g. 0.25 m water depth threshold for damage occurrence in Arrighi et al.). We will add an explanation for this fact in the revised manuscript.

R1-C7: Line 442 - there is some text missing "given that both datasets show comparable values. in, as"
Answer: Yes, thank you. It was a typo and we will fix it in the revised manuscript.

R1-C8: Line 486 - I find it highly interesting that there is an expectation that many people do not claim through their insurance - this is worthy of a new line of inquiry in its own right (although not within the paper as it is out of scope).
Answer: Thank you for the comment.

R1-C9: Could the authors say more about what is needed to apply models not developed for Italy to work - for example, does the research suggest simple steps that could increase the performance of models developed in the Netherlands or Germany?
Answer: We think this aspect was already addressed in the original manuscript, although probably not stated very explicitly: to apply a damage model to a different country, it is important to verify the comparability between the original and the investigated physical (in terms of hazard and building features) and compensation context, as well as the availability and coherence of the input data. In the revised manuscript, we will include a summarizing statement on this point in the take-home messages reported in the conclusions of the paper.

Reviewer 2

R2-C1: The paper is limited by there being only one validation set. This validation set has problems as is correctly mentioned in section 3.4. In earlier similar studies (e.g. Jongman et al., 2012) that used multiple validation studies it was common to see that some models performed well on one validation set and bad on another. I think this limitation of the study should be mentioned a bit more clearly, also to point out for future studies of this kind that it may be a good idea to have multiple validation sets (such as in Jongman et al., 2012). Because of this and the next point, the main value of the paper is in a comparison between the models rather than an absolute value judgement of the models.
Answer: We certainly agree with the Reviewer that it is always desirable to have more validation datasets and that some models can work well in a case and worse in others, as shown in the study of Jongman et al. (2012). Unfortunately, most of the times it is even difficult to have only one dataset, given the well-known paucity of ex-post damage data. In any case, we will clarify the importance of having multiple validation datasets in the revised version of the manuscript (see also reply to comment R2-C5). Regarding the point on absolute value judgement of the models, we would like to stress that the main aim of our paper was not to identify the “best” damage model or to make any kind of “absolute” ranking, but rather to provide potential users of damage models with general considerations on the spatial transferability of the modelling tools and reliability of loss estimates. Besides, we would like to stress the additional innovative aspect of our study, e.g. in comparison with Jongman and others, which is the blind validation test providing more objective insights, than when modelers know the results they are aiming at. These points will be better stressed in the introduction and in the discussion/conclusion sections of the revised version of the manuscript (see also response to comment R2-C2).

R2-C2: The second weakness is also not really highlighted and that is that the number of compared models is too small to draw any strong conclusions. Obviously, it wasn’t feasible to select more models but some of the good model performances shown in this paper are pure coincidence, I’m sure about that. A good or bad performance of a model should therefore be seen as a single data point (sample) that could be just be a coincidence. This should be taken into account when drawing conclusions. I think this goes well in the conclusions section but in the results and in the discussion section some of the speculations should be done more cautious. Maybe go through these sections and reevaluate some of the reasoning based on the idea that there is a high level of coincidence in the results. Also be clearer about this limitation and the general idea that some observations could be simply a coincidence, add a few sentences about that somewhere.
Answer: The Reviewer claims that one weakness of our paper is related to the use of a limited number (9) of damage models. It is true that this is not a huge number, however it is line with other studies testing damage models that can be found in the literature (e.g. Jongman et al. (2012) compared 7 models). Most importantly, as also pointed out in the manuscript, we selected only those models that were mastered by the authors, in order to avoid any possible bias in their application (see also response to comment R2-C6).

The second point raised by the Reviewer is that the good/bad performances of the models are due to a coincidence, based on his/her own belief. We based our discussion and conclusions on the results emerging from the empirical analysis carried out in our paper. Some of the outcomes were not surprising and are corroborated by previous studies: for instance, (i) the better performances provided by local models rather than imported ones; (ii) models providing good results have proven to perform well also in other validation case studies for other events in Italy (e.g. Arrighi et al. worked well also for the 2010 flood in Veneto Region (Scorzini and Frank 2017); the same applies to INSYDE and Carisi et al., which were tested in other Italian flood events (Amadio et al. 2019); similar considerations can be made for the model of Dutta et al. which was already found to not properly work in Italian cases (Scorzini and Frank 2017)); (iii) multi-variable models can provide worse performances than simpler ones if they are applied in contexts different from the original one, either in terms of physical features or availability of the input data. However, we are aware, that these results are associated with uncertainties and the general picture might be different when the models would be applied in different case study areas. Thus, we will thoroughly check the results and discussion section to avoid overconfident statements.
In addition, as mentioned in the reply to the previous comment, our paper was not aimed at identifying the "best" damage model, but rather to provide potential users of damage models with general considerations on the spatial transferability of damage models and reliability of loss estimates. This point will be better stressed in the introduction of the revised version of the manuscript.

R2-C3: The statement in the discussion about the value of multi-variable models against simple models (line 526) cannot be stated like this. Multi-variable models can only be transferred when there is some overlap in the context between the training and validation data (see Wagenaar et al. 2008). When there is no overlap the transfer obviously wouldn't work no matter how many variables you add to the data. A multi-variable model should always be compared to a single-variable model based on the same data and not to a single-variable model from a different region to then conclude that a multivariable model isn't useful. Maybe this could be done with a very large number of models but not based on the tiny number tested in this paper (especially because this observation also contradicts common sense and earlier findings elsewhere).
Answer: We fully agree with the Reviewer and our paper corroborates this point, given that the results indicated that multi-variable models applied in contexts different from the original one could perform worse than simple models and this should be considered as a "caveat" for models' users.

R2-C4: The paper title, abstract and introduction puts a lot of emphasis on it being a blind-test. All properly carried out model validations studies don't use any knowledge from the validation data for the model development (standard practice). The constant emphasize on that in this study is therefore a bit misplaced I think. Did earlier studies not follow this approach? I believe they did. Maybe they didn't advertise it this strongly, or maybe this study was more strict or systematic on that but does that really add anything special?
Answer: We agree that all model validation studies keep the validation data separate from the data used for model development, but this is not the point here. Commonly, model validation studies are carried out in a way, that the modellers know the validation data and thus the result they are aiming at. Thus, model applications can be tuned to get as close to the desired result as possible. The adoption of our blind approach prevents any possibility of "tuning" the input variables of the damage models, especially the parameters related to the more qualitative vulnerability features. We will stress this point in the revised version of the manuscript.

R2-C5: This study is very similar to the paper Jongman et al., 2012. It might be interesting to add a paragraph in the discussion section to compare the results of the papers. Of course, this paper is comparing much newer and different models. Yet the general observation that the results are very different from each other and that its difficult to make sense of that is the same among the studies.
Answer: In the revised version of the manuscript we will include some discussion on the mentioned study, by quoting it especially regarding the importance of having multiple validation sets (see also response to comment R2-C2), although we do not fully agree with the Reviewer on the similarity with

the paper of Jongman et al. 2012 (at least for the main objectives). In fact, our study aims at providing potential users with general considerations on the spatial transferability of the modelling tools and reliability of loss estimates, with a specific focus on micro-scale damage models for the residential sectors, while the study of Jongman et al. is focused on strengths and weaknesses in existing modelling approaches (working at different spatial scales and for different exposed sectors) towards the development of a harmonized European approach, which implies an adjustment of modelling tools that was not instead performed in our study, where models have been implemented in their original formulation.

R2-C6: The large group of authors suggests that an expert on each of the compared models was included in the paper. This is however clearly not the case for the Jonkman et al. model and hence the paper makes claims about this model that aren't true. Apart from fixing these mistakes (which I pointed out below), I think it should be clarified somewhere which experts worked on which models and whether the expert personally developed the model or interpreted information from literature (much more error prone).
Answer: The authors were either developers of the models (Arrighi et al., Carisi et al., CEPRI, FLEMO-ps, Insyde) or frequent users with a certain knowledge of the models (Dutta et al., Fuchs et al. and Jonkman et al. are commonly used in Switzerland by the group of authors from the University of Bern). In revising the manuscript, we will include an additional comment, in the introduction section, on the importance of having the contribution in the study of model's developers/experts, as this prevents any possible bias in the results that could arise from an incorrect application of the models (for example, a non-expert may experience a misunderstanding of any of the input variables which would affect the final results).

Minor comments:

R2-C7: Line 46: This is grey literature and I can't find it on the internet. Perhaps add one of the very many peer-reviewed journal articles that could also be used to support this statement and are often much older than 2007.
Answer: We will include the following additional reference in the revised manuscript: Merz, B., Kreibich, H., Thieken, A., & Schmidtke, R. (2004). Estimation uncertainty of direct monetary flood damage to buildings. Natural Hazards and Earth System Sciences, 4: 153-163.

R2-C8: Line 101: I have never seen the term "low-variable" model, maybe consider a different word? Do you mean "single-variable" here?
Answer: We decided to introduce a new term, as we think that "single-variable" is used incorrectly since damage models always consider at least two variables (footprint area and water depth). The meaning of the term will be explained in the revised text.

R2-C9: In section 2.3/table 2. Some basic but important information in this section seems missing for some of the models, such as the origin of the model (country/region) (for Carisi et al. – mono) and the intended flood type (for many of them) and the year the model was made (maybe can be retraced from reference list but not easy for reader). Maybe its good to add this information to table 2 or at least make it clear in the section. Also maybe try to make the different texts a bit more uniform (i.e. present same information in same order).
Answer: Thank you for the suggestion. We will include missing information in the revised version of Table 2.

R2-C10: Line 245. I don't see the significance of this model using a mathematical function. I also know much older models doing that and think its an irrelevant characteristic. Some other damage functions may also follow a mathematical function but just communicate it different.
Answer: The Reviewer is right. Then, we will revise the sentence regarding Dutta et al.'s model by deleting "with a mathematical function".

R2-C11: Table 2: I don't think its correct to classify the Jonkman et al., 2008 model as empirical. It was inspired by many different sources of empirical data but no systematic empirical method was applied and the model is basically expert judgement considering a few empirical data points.
Answer: Thank you very much for the specification. We checked the paper again. According to the paper, the function was developed based on empirical data combined with existing literature and expert judgment as we have written in the manuscript. To make it more clear, we will change the sentence in L279-281 to "The model by Jonkman et al. (2008) it is a simple relative damage model considering water depth, building area (of all floors) and building (replacement) value as explicative variables, developed on the basis of empirical flood damage data of the past in the Netherlands in combination with existing literature and expert judgment." Accordingly, we will also revise Table 2 for the Jonkman et al. model by changing "empirical" with "mixed".

R2-C12: Section 3: I don't like this title very much why not just "results" this is confusing
Answer: Ok, we will change the title of Section 3 with "Results".

R2-C13: Line 293-294: Maybe clarify that you look at a subset of the models here and didn't adjust the whole building models to only use the ground floor value (unclear at first).
Answer: Thank you. We will better clarify this point in the revised version of the manuscript.

R2-C14: Line 298 and 299: Can you rephrase this sentence its currently confusing.
Answer: The Reviewer is right. Also in line with comment R3-C13, we propose to simplify the sentence as follows: "Total damage estimations differ by a maximum factor of 12.6, suggesting that the shape of the damage functions exacerbate the variability of models' outcomes due to exposure estimation".

R2-C15: Line 305: Could you rephrase this sentence, it's difficult to follow: "Individual damage estimates differ on average by a factor of 28, with the more frequent factor around 10."
Answer: Also in line with comment R3-C13, we propose to simplify the sentence as follows: "Individual damage estimates differ on average by a factor of 28".

R2-C16: Table 3: Could you split the 3th column, this is confusing and an uncommon form a presenting it. The title of 3.1 is a bit unclear. Could you rephrase it? It's a bit long and I didn't get that with blind mode you mean that its not compared to observations yet (which I expected). Maybe call it "comparison of the models". Or keep the title and clarify in the first sentence that the reader shouldn't expect the comparison to real observations yet.
Answer: In the revised manuscript we will modify Table 3 as suggested by the Reviewer, while we would like to keep the title of section 3.1, given that the meaning of "blind" has been described earlier in the paper; however, to enhance clarity, in the revised manuscript we will include an introduction statement explaining that the results presented in that section do not yet involve any comparison with observed damage data given that, at the stage described in section 3.1, the models are still applied in a "blind mode".

R2-C17: Figure 4: This table is interesting but requires some additional discussion. When you have two very different exposure values but the variation among the buildings solely depends on the size of the buildings the correlation between the two very different exposure values is still one. So this figure mostly says something about the characteristics that differ per building (if I understand the figure correctly). This doesn't become clear from the text.
Answer: Based on Reviewer's comment, we realized that we missed to describe in detail Figure 4, as it provides information on a building-by-building comparison, so the size of the buildings does not have effects on the results shown in the Figure. We will then clarify this point in the revised manuscript.

R2-C18: Section 3.2. I like the content of this section but I had to read it twice to fully understand it. Maybe the authors could try to clarify this section a bit more. I think especially the title and the first sentence don't make it very clear at first (its all correct but it's a bit of deciphering for a reader)

Answer: Ok, we will try to better clarify this section in the revised manuscript.

R2-C19: Table 4: Could you split the column again like in the previous table.
Answer: This will be fixed in the revised version of the manuscript.

R2-C20: Table 4: Could you consider including the model origin of all models in this table. That would be a very interesting reminder for the reader. Especially for readers who don't read the entire text this would be very useful.
Answer: Thank you for the suggestion. In the revised manuscript we will show in the first column of the Table (in parentheses) the origin of the different models.

R2-C21: Line 487: In case of the Netherlands this isn't true. The empirical data in this model is at best used relative. Absolute values are 100% synthetic.
Answer: Thank you for the clarification. We will not mention the Netherlands in this sentence.

R2-C22: Line 554: Micro-models are essential when measures are undertaken at micro level (e.g. for insurance or studies about elevating specific houses as is common in some countries). When aggregated damages are assessed they may indeed add less information. A second reason why micro-models are important is for at least for the location of buildings. The difference between a house flooding or not is sometimes a matter of just meters. Its therefore important to work with precise models on location even if the other building characteristics are all the same.
Answer: We agree with the Reviewer on the importance of micro-scale models, but, in our opinion, the main point that deserves some discussion is not when and why micro-scale models are useful/important (as this is also well known from the literature), but rather on their actual usefulness if they are not able to provide reliable results.

Reviewer 3

R3-C1: In my opinion, a main drawback of the whole study lies in the approach to 'validate' the results. The authors claim to assess the models' reliability through a comparison with observed damages of a real flood event (chapter 3.3) and show respective results (for example Table 4 and 5). The following chap 3.4 is then dedicated to explaining why the models results differ so strongly from the observed results. Crucial reasons for this discrepancy found are then assigned to inconsistencies in the damage claims, that is, the validation data. This seems for me like an odd approach. If the damage claim dataset was meant to be used as a validation set, more emphasis should have been put on clarifying inconsistencies and maybe on further filtering the dataset down to a set of reliable data on damage of a reduced number of buildings
Answer: The reason why we didn't filter the dataset before performing the validation exercise, as the Reviewer is suggesting in his/her comment, is that we decided to apply a real "blind approach" also to the handling of the empirical data, avoiding any "adaptation of the observations to the model". The main reason behind this choice is the intention to underline a common problem we face in the scientific community, concerning the quality of damage data used for validation, and to warn about conclusions that can be derived from validation analyses: for instance, if a model does not fit well some empirical data, this does not mean that it is not a "good" model and vice versa.
Moreover, it is often the case that empirical data are used in validation analyses without any possible preliminary evaluation on their quality and significance, simply because no ancillary information is available. An example of this kind of data is represented by insurance data, which usually lack of useful information to obtain insights on the quality/significance of the damage databases (e.g. Denmark (Zhou et al. 2013), France (André et al. 2013), the Netherlands (Spekkers et al. 2013) and the US (Wing et al. 2020)). Then, the general question we would like to arise is the following: how one can be sure to derive solid conclusions on the results of a validation analysis if no information on the quality of used empirical data is available? We propose to include some comments on this point in the conclusions (after L580) of the revised version of the manuscript.

- *André, C., Monfort, D., Bouzit, M., and Vinchon, C. (2013). Contribution of insurance data to cost assessment of coastal flood damage to residential buildings: insights gained from Johanna (2008) and Xynthia (2010) storm events, Nat. Hazards Earth Syst. Sci., 13, 2003–2012.*

- *Spekkers, M.H., Kok, M., Clemens, F.H.L.R., and Ten Veldhuis, J.A.E. (2013). A statistical analysis of insurance damage claims related to rainfall extremes. Hydrology and Earth System Sciences, 17(3), 913-922.*
- *Wing, O.E., Pinter, N., Bates, P.D., and Kousky, C. (2020). New insights into US flood vulnerability revealed from flood insurance big data. Nature communications, 11(1), 1-10.*
- *Zhou, Q., Panduro, T.E., Thorsen, B.J., and Arnbjerg-Nielsen, K. (2013). Verification of flood damage modelling using insurance data. Water science and technology, 68(2), 425-432.*

R3-C2: The authors only declare having had some informal conversations with experts but the explanations about low damage values observed remain fuzzy. For the approach of this paper a more thorough survey of affected people would have been necessary.
Answer: Unfortunately, as many years have passed since the flood event, it was not possible to get in contact with all of the affected people (many of them have moved out). However, as indicated in the manuscript, we had the opportunity to have conversations with representatives of the Committee of Flooded Citizens in Lodi, who were able to give us descriptive information about occurred damages in large part of the town. Moreover, we want to stress that we did not only had conversations with people, but we also performed an analysis of the different damage components to have more insights on observed data (Section 3.4).

R3-C3: The statement in line 471ff “According to our interpretation, inconsistency between expected and declared damage can be attributed to the fact that what is declared by citizens does not correspond to the actual money required to replace or reconstruct the whole physical damage suffered by the building” cannot satisfy and actually triggers a the more philosophical question, whether the ‘damage’ targeted by the model represents the damage felt by the people affected.
Answer: We actually wanted to raise a modelling question, rather than a philosophical one. The analysis of the described post-event damage data suggested what is reported in the statement in L471; of course, as also highlighted in the text, it is our interpretation, but it fits well with the obtained results.

R3-C4: I suggest shifting those parts of the chapter 3.4 with the explanation about how the validation dataset was generated to an earlier part of the paper as background information about the approach (where also the various models are described).
Answer: We had long discussions among us on which could be the best choice for the structure of the paper and, finally, we felt that the proposed one was the most coherent with the adopted blind approach, so that the reader is not influenced from this information when reading previous results (see also reply to comments R3-C1 and R3-C9). This said, we would like to keep the mentioned parts of section 3.4. where they were in the original manuscript,

R3-C5: The detailed statistical comparison of the model results with the overall average does in my opinion not really add value to the result analysis (this relates particularly to Table 3 and partly to Table 4, it is more adequately visualized in Figure 5!). Firstly, because the number of models is so small. Secondly, and more important, since it is a bit like comparing apple and pears as you say yourself in the interpretation. The large differences in the model results derives from the different types of models. Therefore, I would not list in detail the variation of the models to the average but only describe and explain the differences of those models with similar approaches. The analysis could be done in a more qualitative way since numbers such as average or variation does not make so much sense when the overall number of models is so small.
Answer: The data reported in the Tables were intended to be used as a result of a sensitivity analysis and not of a detailed statistical analysis. Given that other Reviewers did not complain on this point, we would like to maintain the Tables as they are, better specifying the aim of our analysis so as not to create misunderstandings.

R3-C6: Line 81: “Reality is hardly reproduced by observed data” – I would suggest not to use the term ‘reality’ since there is no univocal damage value as you prove later on yourself
Answer: This was a provocative statement. Indeed, we also specified that there is not a “univocal” reality and therefore the term should not be used or used with caution. We would like to keep the

sentence as it is, however if the Editor thinks it is too strong, we may consider a softening or the possibility to write “Reality”, in quotes.

R3-C7: Line 86f: “comparative studies over a broad range of test cases are essential for acquiring more confidence in the reliability of modelling tools” – after having read your paper I would not say that the test case increased the confidence in reliability, I would put the emphasis more towards understanding in detail how certain model results come about
Answer: This was a general statement in the Introduction section; however, according to Reviewer’s suggestion we propose to modify it as follows: “comparative studies over a broad range of test cases are essential for acquiring a thorough understanding of the performances of the modelling tools that could help in enhancing the confidence in their reliability”.

R3-C8: Line 101: “the focus of this study lies in this specific set of models” – what does this mean?
Answer: In the revised manuscript we will amend the sentence as follows: “This study focuses on micro-scale (i.e. individual item scale) direct damage assessment to residential buildings, in line with the larger availability of damage modelling approaches developed in Europe for this specific sector.”.

R3-C9: Line 149: “was not uniform, as only some of the owners justified costs for fixing the damage by means of invoices” – if not earlier, here the reader should suspect that the quality of the ‘validation’ dataset is questionable. I would propose to already link here to further explanations about this ‘observed’ damage data.
Answer: See also reply to comment R3-C1. The case presented in the paper is a “fortunate” one, given that we were able to make some considerations on the quality of the data, based on the availability of ancillary information. In line with the adopted blind approach, we do not agree on anticipating in the presentation of the case study some of the results that come out only after the unblinding of the observed damage data.

R3-C10: Line 186: consider to call it variation rather than ‘difference’
Answer: Thank you, we will consider the suggestion in the revised manuscript.

R3-C11: Line 195: unclear for me, until here I thought the observation data was derived from damage claims made after the flood event. How can you use ‘official claims’ to explain inconsistencies between estimations and observations? Or is the officially claimed data different from the claims mentioned earlier? Then you could have used the official ones for validation?
Answer: We thank the Reviewer for highlighting a possible source of misunderstanding. The same database of damage claims was used throughout the paper. In the revised manuscript, we will delete the word “official” in order to avoid confusion.

R3-C12: Line 207: does only this model use stage-damage curves? Or why is that here mentioned explicitly?
Answer: Clearly it is not the only model using stage-damage curves. To avoid confusion, in the revised manuscript the sentence in L205-208 will be rephrased as: “The model developed by Arrighi et al. (2018a, 2018b) is a relative synthetic model which expresses monetary damage as a function of water depth and recovery cost for buildings with and without basement. A zero-damage threshold is set for a water depth lower than 0.25 m for buildings without basement”.

R3-C13: I consider some of the explanations about calculation findings as being far too long because too obvious when the conclusion is that the differences can be traced back to the different model approaches (for example Lin 332 ff)
Answer: We will try to shorten the explanations about calculation findings in the revised version of the manuscript.

R3-C14: Line 453 ff: it is not clear of the percentages refer to the amount of building or the outlier value (I suppose the former but that needs to be clarified)

Answer: This was already made clear in the original manuscript, as it can be noted in L453 where it was stated “examining in detail the outlier claims”.

R3-C15: Line 581: “Consultations of experts with local knowledge can ensure the correct interpretation and use of observed damage data” – I would not agree with that, it may help but does not ensure. –
Answer: The Reviewer is right and then in the revised manuscript we will better specify that “Consultations of experts with local knowledge can help in the correct interpretation and use of observed damage data”

R3-C16: Line 95: “being them unknown” – unclear, pls consider reformulation
Answer: We will amend this sentence in the revised manuscript by changing the word “unknown” with “undisclosed”.

R3-C17: Line 442: something went wrong with “in, as.”
Answer: Yes, thank you. It was a typo and we will fix it in the revised manuscript.