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 took 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 have modified 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, in the revised manuscript, we have included the following details on the hydraulic modelling of the event (L138): "*In detail, with respect to the hazard, information on observed water depths was available for more than 260 points within the inundated area, deriving from indications provided by municipal technicians and by citizens in damage compensation requests, as well as from interpretation of photographs taken during or immediately after the flood. These data were used for the validation of the 2D hydraulic simulation of the event: the resulting average absolute differences between observed and calculated water depths within the inundated area ranged from 0.2 to 0.4 m, depending on the validation zone in which observed water depth data were aggregated (Scorzini et al., 2018). This is surely a possible source of uncertainty; however, reported differences could be considered to provide relatively small impacts on the damage estimation. Moreover, given that all tested damage model shared the same hazard data, this would be a common source of uncertainty that should not affect the overall results of the blind test*".

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 the revised version of the manuscript, we have included 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: The point on the adjustments needed for the models to work was already described in Step 2 of the "Methodology" (P7. L170-173), but in the revised version of the manuscript we have better exemplified this part by including the following text in the discussion (L 578): *"For example, the building categories (BT) assumed by CEPRI ("apartment"/ "single storey building"/ "multi-storeys buildings") are different than the Italian ones ("apartment"/ "detached"/ "semi-detached") so that a correspondence has to be defined, also on the bases of the number of floors (NF); specifically "apartment" is defined by BT ="apartment", "single-storey" is defined by BT = "detached" or "semi-detached" and NF = 1, "multiple-storeys" is defined by BT ="detached" or "semi-detached" and NF > 1. Correspondence among building categories was defined also for the implementation of FLEMO-ps, although in this case the task was quite straightforward, since the German building categories are almost coincident with the Italian ones (FLEMO-ps distinguishes between "Multi-family house" / "Semi-detached house" / "One-family home").*
Moreover, we have made 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 and INSYDE, which are absolute damage models).

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. and 0.01 m water depth threshold in Dutta et al. and Jonkman et al. to distinguish between flooding and surface water runoff). In the revised manuscript, we have included an explanation for these results in the caption of Figure 3.

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 have fixed 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. In the revised manuscript we have included the following summarizing statement in the take-home messages reported in the conclusions of the paper (L650): "*As a general recommendation, to select a damage model for an application in 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"*.

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); we added a sentence in the conclusions highlighting this aspect. 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 clarified the importance of having multiple validation datasets in the introduction and in the discussion sections of the new 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 modellers know the results they are aiming at. Although we think that these points were already explicated in the first version of the manuscript, we further emphasized them in the introduction of the revised version (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 pointed out in the original version of the manuscript and better explained in the new version, 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) (L91): "*Tested models have been chosen among those mastered by the authors; indeed, the authors were either developers of the models or experienced users with significant knowledge of them, in order to prevent any possible bias in the results that could arise from an incorrect application of the models (for example, a non-expert user may misunderstand the meaning of some input variables, which would affect the final estimation)*".

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, as now better explained in the revised version of the paper (see discussion): 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. From another point of view, results were critically analysed, by considering both the features of the models and the context of investigation, and never taken from granted, despite their agreement with our expectation. For example, in section 3.3 we comment: "*The table also indicates the better performances of the Italian/local models with Arrighi et al. showing the lowest difference. However, by looking at its features, it is possible to state that even this last model tends to overestimate damage. First, because it does not consider clean-up costs (like INSYDE and CEPRI), which are instead included in the observations. Second, because the lower value of the total damage with respect to other models is partly due to the effect of the zero damage threshold for water depths lower than 0.25 m (see Sect. 3.1)*"; in section 3.4 we comment: "*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 (…) This would explain why synthetic models overestimate observed damage, as they are usually based on full replacement/reconstruction costs. Likewise, it would explain why the model by Arrighi et al. performs better than others: indeed, the recovery value adopted by this model is defined as the average difference between the market value of new buildings and that of equivalent, older buildings requiring renovation. It is then sensible that this value reflects a balance between the two opposite extreme behaviours of buyers (which, in turn, depend on their financial resources): i.e. to completely renovate the building or to bring the building back to a minimum level of functioning. In our view, such behaviours can be compared with those of flooded owners*" and, again "*Moreover, declared monetary damage is strongly correlated to the expectations that citizens have to be reimbursed. This expectation is low in Italy (…). This would also explain why*

empirical models (derived from claims) developed in regions with high expectations and then high values of declared damage (like Germany or the Netherlands), overestimate the observed damage in this case study". Thus, after checking the paper again we think that no overconfident statements were included.
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 have been further 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 stressed this point in the revised version of the manuscript (see introduction, L101): "*possible biases are avoided as participants cannot be influenced by validation data, being them undisclosed in the implementation phase of the models, e.g. by trying to adjust or tune their models, especially regarding the more qualitative input parameters, in light of observed damages".*

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 conclusion of the revised version of the manuscript we added a 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 experienced users with significant 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 included 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 included 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 is explained in the revised text (see introduction, L57).

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 included 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 revised 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 changed the sentence to "*The model by Jonkman et al. (2008) it is a simple relative damage model considering water depthand building (replacement) value of all floors 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 also revised 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 changed 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 better clarified this point in the revised version of the manuscript.

R2-C14: Line 298 and 299: Can you rephrase this sentence its currently confusing.
Answer: In the revised version of the manuscript, the sentence was rephrased as following (L309): *“Total damage estimations differ among the modelling approaches by a maximum factor of 12.6, which is limited to 3.1 with respect to the mean value of total damage estimations, 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: we simplified the sentence as follows (L315): “*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 modified Table 3 as suggested by the Reviewer, while we preferred 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 included a 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 clarified this point in the revised manuscript (L338): “*Figures 2 and 3 further highlight a common trend in exposure and damage values supplied by the different models, also confirmed in Fig. 4 and 5, showing the Pearson’s correlation coefficients for individual (i.e. building by building) exposure and damage estimates. (…)”*

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: We tried to better clarify this section in the revised manuscript, by amending the title (now as “*Role of input variables in the determination of divergent models’ outcomes”*) and the first sentences of the section, as follow (L382): “*In order to explain the differences observed in the blind implementation, models were compared in terms of trends and variance of individual damage estimates, for classes of values of input variables, and by considering one variable at a time. The objectives of the analyses were to investigate whether the consideration of a specific input variable*

influences the outcome of a model with respect to the other ones, whether the inclusion of more explicative variables may be considered as a possible source of variation, and to identify the most influencing parameters on the final output of the models. (…)"

R2-C19: Table 4: Could you split the column again like in the previous table.
Answer: This has been 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 show in the first column of the Table (in parentheses) the origin of the different models. We did the same also for Table 5.

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. In the revised version of the paper, we do 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 have then included the following text in the conclusions of the revised version of the manuscript (L632): "*Indeed, 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, as for instance for insurance data (André et al. 2013; Spekkers et al. 2013; Zhou et al. 2013; Wing et al. 2020). In this context, the blind test highlighted that "reality" depicted by observations is not univocal, so that data must be carefully investigated before their comparison with model outcomes, as they may be addressing different types of damage, damage to different components, or being incomplete. Based on this consideration, there is a need to be always cautious when drawing conclusions from validation analyses, given that if a model does not fit well some empirical data, this does not necessarily mean that it is not a "good" model and vice versa*".

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 have kept 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 explaining the aim of our analysis so as not to create misunderstandings. Therefore, when presenting Table 3 in the revised manuscript we have clarified that (L296): "*With the aim of understanding the impact of exposure estimation on damage assessment and identifying possible common features in the results, Table 3 shows the total*

exposure and loss figures obtained by applying the nine models to all buildings within the simulated inundation area (877 in total; see Figure 1)". The same was done when introducing Table 4 (L430): "*Table 4 summarises the results of the sensitivity analysis by comparing the total observed damage to the total damage estimates obtained with the implementation of the nine models to the subset of buildings".*

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 preferred to keep the sentence as it was, however we have written "Reality" in quotes in order to highlight its "non-uniqueness".

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 have modified 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 have amended 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. In the revised manuscript we have amended the text accordingly.

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 have deleted 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 original sentence in L205-208 has been 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 think that explanations on calculation findings, although obviously linked to modelling approaches, are fundamental to justify results, especially to non-expert readers. So we preferred to keep the text as it is.

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 have better specified 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 have amended 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 have fixed it in the revised manuscript.

Are flood damage models converging to "reality"? Lessons learnt from a blind test

Daniela Molinari[1], Anna Rita Scorzini[2], Chiara Arrighi[3], Francesca Carisi[4], Fabio Castelli[3], Alessio Domeneghetti[4], Alice Gallazzi[1], Marta Galliani[1], Frédéric Grelot[5], Patric Kellermann[6], Heidi Kreibich[6], Guilherme S. Mohor[7], Markus Mosimann[8], Stephanie Natho[7], Claire Richert[5], Kai Schroeter[6], Annegret H. Thieken[7], Andreas Paul Zischg[8] and Francesco Ballio[1]

[1] Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy
[2] Department of Civil, Environmental and Architectural Engineering, University of L'Aquila, Via Gronchi 18, 67100, L'Aquila, Italy
[3] Department of Civil and Environmental Engineering, University of Florence, Piazza San Marco 4, 50121, Firenze, Italy
[4] Department of Civil, Chemical, Environmental and Material Engineering, University of Bologna, Viale Risorgimento, 2 - 40136, Bologna, Italy
[5] G-EAU, Univ Montpellier, AgroParisTech, CIRAD, IRD, INRAE, Montpellier SupAgro, Montpellier, France
[6] GFZ German Research Centre for Geosciences, Section Hydrology, Telegrafenberg, 14473, Potsdam, Germany
[7] Institute of Environmental Science and Geography, University of Potsdam, Karl-Liebknecht-Strasse 24-25, 14476, Potsdam, Germany
[8] Institute of Geography, Mobiliar Lab for Natural Risks, Oeschger Centre for Climate Change Research, University of Bern, Hallerstrasse 12, 3012, Bern, Switzerland

Correspondence to: Daniela Molinari (daniela.molinari@polimi.it)

Abstract. Effective flood risk management requires a realistic estimation of flood losses. However, available flood damage estimates are still characterised by significant levels of uncertainty, questioning the capacity of flood damage models to depict real damages. With a joint effort of eight international research groups, the objective of this study was to compare, in a blind validation test, the performances of different ~~damage~~ models for the ~~estimation~~ assessment of the direct flood damage to the residential sector at the building level (i.e. micro scale) ~~in a blind validation test~~. The test consisted in a common flood case study characterised by high availability of hazard and building data, but with undisclosed information on observed losses in the implementation stage of the models. The selected nine models were chosen in order to guarantee a good mastery of the models by the research teams, variety of the modelling approaches and heterogeneity of the original calibration context, in relation to both hazard and vulnerability features. By avoiding possible biases in model implementation, this blind comparison provided more objective insights on the transferability of the models and on the reliability of their estimations, especially regarding the potentials of local and multi-variable models. From another perspective, the exercise allowed to increase awareness on strengths and limits of flood damage modelling, which are summarised in the paper in the form of take-home messages from a modeller's perspective.

1 Introduction

Efficient and effective flood risk management requires a realistic estimation of flood losses, implying the use of reliable models for flood hazard, damage and risk assessment (Meyer et al., 2013; Gerl et al., 2016; Zischg et al., 2018; Wagenaar et al., 2018; Molinari et al., 2019). Although several hydraulic models are available (Teng et al., 2017), their variety seems to be overtopped by the variety of flood damage models as, according to Gerl et al. (2016), only in Europe, 28 models (including 652 functions) exist to assess flood losses, whereas almost half of them focus on residential buildings.

Even within the residential sector and with respect to direct damage (i.e. damage due to the direct contact with the flooding water), the diversity of approaches is manifold. First, the models are classified according to the intended spatial scale of the analysis: while micro-scale models refer to the individual exposed building, meso-scale models work at more aggregated scales, like land use or administrative units, with large-scale spatial units (like regions or countries) being at the base of macro-scale models (Merz et al., 2010).

A second difference lies in the approach adopted for model development, with empirical models using damage data collected after flood events (~~see~~ e.g. Merz et al., 2004~~Huizinga et al., 2007~~) and synthetic approaches implementing information collected via what-if-questions (~~see~~ e.g. Penning-Rowsell et al., 2005). Still, both categories are characterised by a variety of methods; for example, empirical data can be interpreted by means of different statistical and mathematical tools, ranging from simple regression (e.g. Merz et al., 2004~~Huizinga et al., 2007~~) to more sophisticated machine learning algorithms and data mining approaches (e.g. Merz et al., 2013; Amadio et al., 2019). A distinction can also be made between absolute and relative damage models: the first directly return a value in a specific currency (Dottori et al., 2016; Rouchon et al., 2018), while relative damage models estimate the physical vulnerability or the degree of loss of an exposed asset (Fuchs et al., 2019a), to be multiplied by its monetary value to assess the damage. Linked to this point is the question of what is defined as exposure in the models: besides the distinction whether a model relies on the value of the whole building or just of the affected floors, it is also important to know if, for instance, the basement is considered as well. Moreover, exposure assessment may differ regarding the monetary value, whether it is based on e.g. market or replacement values (Röthlisberger et al., 2018), rather than full replacement costs or depreciated values (Merz et al., 2010).

A final important difference among the models lies in the number and type of considered input parameters, i.e. on model complexity. Simplest damage models (referred to, in the following, as "low-variable models") take into account a few number of variables, mostly the water depth at building location as well as building area and its monetary value (only in case of relative models). Even in their simplicity, these models can significantly differ from each other, due to the distinct shapes of the underlying damage functions, e.g. square root function (Dutta et al., 2003; Carisi et al., 2018), beta distribution function (Fuchs et al., 2019b) or graduated function (Jonkman et al., 2008; Arrighi et al., 2018a). On the contrary, multi-variable models consider numerous hazard and exposure/vulnerability input factors and, consequently, are supposed to be more accurate when detailed data is available (Thieken et al., 2008; Schröter et al., 2014; Wagenaar et al., 2017; Amadio et al., 2019). Nevertheless, simple models tend to be the most widely used, due to their ease for implementation and low requirements for input data.

Hence, flood damage modellers have always to envisage the trade-off in the model choice, ~~e.g~~i.e., ~~applying~~ using a complex, probably more accurate model with specific data requirements, ~~,~~ or a simple, probably less accurate one that can be applied without extensively available data. However, it ~~is~~ has been shown~~,~~ that even a small ensemble of models outperforms individual models, ~~and additionally has~~with the additional advantage of providing uncertainty information (Figueiredo et al., 2018).

What most models have in common is that they are calibrated in specific contexts, usually representative of a ~~specific~~ certain spatially limited region. In many cases, instead, validation of flood damage models is lacking (Merz et al., 2010; Gerl et al., 2016; Molinari et al., 2019). Where it is not lacking, the data used for model validation ~~is~~ are often either a subset of the dataset used for calibration or ~~is~~ are ~~obtained~~ collected in the same region or country of model development. This implies that, even if a model has been locally validated, it is not necessarily correct to apply it to any other region, unless ~~it~~ this latter reflects the context for which the model was derived. For instance, ~~to~~ the application of~~y~~ a damage model ~~which was~~that has been developed for alpine areas (i.e. house building tradition of the European Alps and flood processes involving significant sediment transport) to a coastal country like the Netherlands, and vice versa, is prone to lead to large discrepancies from reality (e.g. Cammerer et al., 2013). Hence, flood damage models need to be tested in regions other than those where they were calibrated in, ~~-~~to ~~assess~~ be confident with their transferability in space~~of flood damage models, they have to be tested in regions other than those where they were calibrated in~~.

Nevertheless, what all models and modellers deal with is the lack of data for model calibration and validation (Merz et al., 2010; Jongman et al., 2012; Meyer et al., 2013; Molinari et al., 2019). ~~Reality~~ The overall economic impact of a flood is hardly reproduced by ex-post data ~~– after a flood~~ and then biases have also to be taken into account when transferring models to different regions, e.g. due to different insurance conditions, uncompleted claims, etc.; moreover, even years after flood events, monetary losses can be revised due to long-term recovery: as an example, ~~(e.g.~~ monetary losses of the 2013 flood in Germany were estimated at 6.7 M€ in 2013 (Deutscher Bundestag, 2013) and changed over the following years to 8.2 M€ (Bundesministerium für Verkehr und digitale Infrastruktur, 2016)~~)~~. For this reason, comparative studies over a broad range of test cases (i.e. different validation datasets) are essential for acquiring a thorough understanding of the performances of the modelling tools that could help in enhancing the confidence in their reliability ~~more confidence in the reliability of modelling tools, based on a thorough understanding of their strengths and weaknesses~~.

The aim of this study is to contribute to the understanding of models' transferability and reliability by testing and comparing different damage models in a blind validation test. This joint effort of eight international research groups consists in a common flood case study characterised by high availability of hazard and building data, but with undisclosed information on observed losses in the implementation stage of the models. Tested models have been chosen among those mastered by the authors; indeed, the authors were either developers of the models or experienced users with significant knowledge of them, in order to prevent any possible bias in the results that could arise from an incorrect application of the models (for example, a non-expert user may misunderstand the meaning of some input variables, which would affect the final estimation). ~~With a joint effort of eight international research groups, the objective of this study was therefore to test and compare damage models used or developed by each group, by applying them in a blind validation test, consisting in a common flood case study characterised~~

~~by high availability of hazard and building data, but with undisclosed information on observed losses in the implementation stage of the models.~~

~~-~~Even though comparative analyses on the performance of damage models have ~~now~~ become more frequent in the literature (Jongman et al., 2012; Cammerer et al., 2013; Scorzini and Frank, 2017; Carisi et al., 2018; Figueiredo et al., 2018; Amadio et al., 2019), according to ~~the~~ authors' knowledge, this study would represent the first flood damage model comparison performed in a blind-mode. ~~By avoiding possible bias (participants cannot be influenced by validation data, being them unknown undisclosed in the implementation phase, e.g. by trying to adjust or tune their models in light of observed damages),~~ T~~t~~his type of comparison can provide more objective insights~~,~~ for a better understanding of models' capabilities and then for reducing modelling uncertainties, as already demonstrated in similar tests performed for other disciplines like seismology, hydrology and computational fluid dynamics (Smith et al., 2004; Soares-Frazao et al., 2012; Krogstad and Eriksen, 2013; ~~-~~Zelt et al., 2013; Andreani et al., 2019; Ransley et al., 2019; Skorek et al., 2019). Indeed, possible biases are avoided as participants cannot be influenced by validation data, being them undisclosed in the implementation phase of the models, e.g. by trying to adjust or tune their models, especially regarding the more qualitative input parameters, in light of observed damages.

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 and scale. ~~Given that most of the approaches for flood damage modelling (in Europe) were developed in relation to the direct damage to the residential sector and at the micro-scale (i.e. building level), the focus of this study lies in this specific set of models.~~

As the research groups use approaches representing many different types and characteristics of models (~~simple (~~low-variable ~~)~~ – multi-variable; absolute – relative; graduated – regression – machine learning – synthetic), being calibrated on the basis of observed data stemming from different countries (Austria, France, Germany, Italy, Japan, Netherlands), with different landscapes and level of complexity in exposure/vulnerability, the blind test as performed in this study can provide~~d~~ an in-depth understanding of the links between models features, their transferability and the reliability of the estimated damages~~an extensive comparison of models as well as an in-depth understanding of their transferability and reliability of the estimated damages~~.

~~The analysis of models' outcomes as a whole aimed at pointing out common patterns or divergent behaviours.~~ In particular, the blind test allowed to investigate these specific questions, raised from the evidence supplied by the literature (Thieken et al., 2008; Cammerer et al., 2013; Schröter et al., 2014; Dottori et al., 2016; Wagenaar et al., 2017; Amadio et al., 2019): do local models (i.e. models calibrated with data from a context similar to the investigated one) outperform other models? Do multi-variable models perform better than simplest ones and if so, why?

The paper is organised as follows. The methodology, models and case study implemented in the blind test are first presented in Sect. 2. Section 3 discusses results of the test, first by considering damage estimates obtained in a blind implementation of the models, and then by comparing damage estimates with ~~real damage data~~documented losses. Answers to the specific research questions are provided in Sect. 4. Finally, in Sect. 5, evidence from the blind test is synthesised in lessons learnt (on flood damage modelling) from a modeller's perspective, including the identification of research needs for further

improvements of flood damage models.

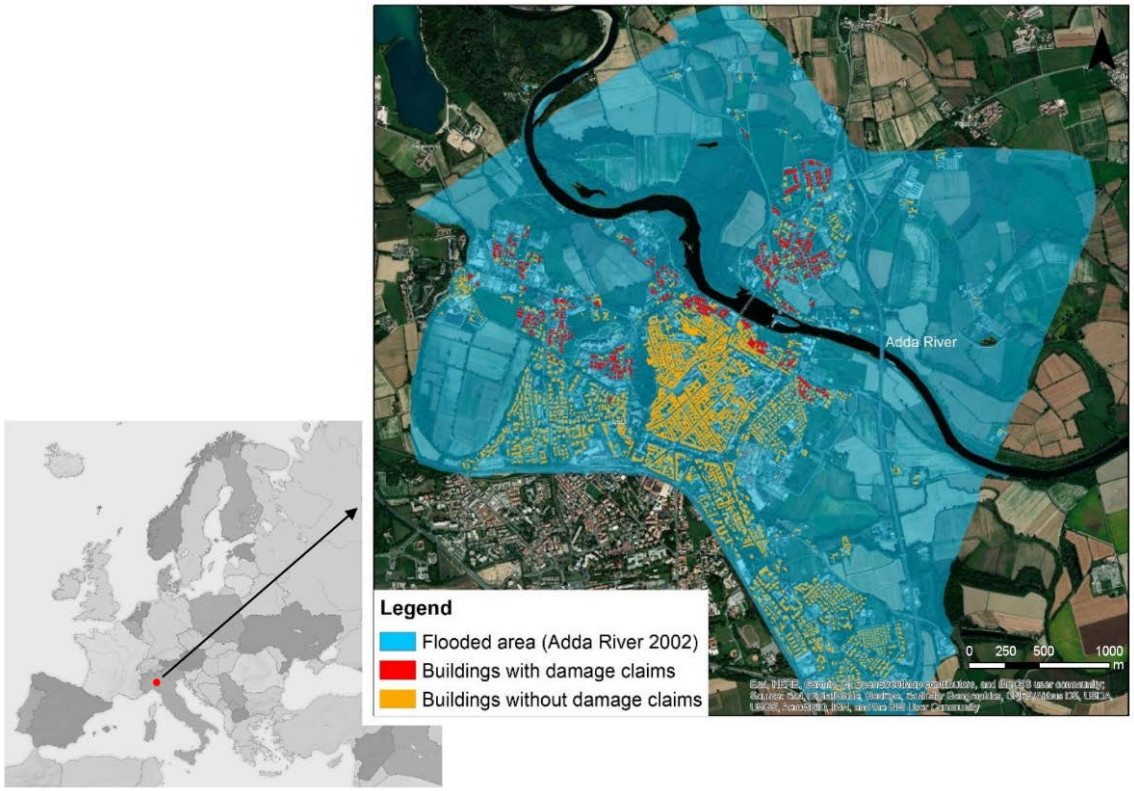
2 The blind test: case study, methodology, models

The main idea behind the blind test was to evaluate the performance of different flood damage models by their implementation to a common case study, to obtain enhanced information on their transferability, validity and reliability; the test is defined "blind" as, in order to avoid bias in the estimation process, the value of the observed damage was unknown to modellers in the implementation stage of the models. In particular, damage data were unblinded only to one group, which was the promoter of the initiative and responsible for data and results management. All required input data to reproduce the damage scenario for the examined event were made available to the participants, who were then asked to submit their results to the exercise manager in an established time frame. Once all contributions from the different groups had been gathered, observed data were disclosed, and models' performances were compared and analysed in a shared discussion between the participants.

2.1 Case study

The investigated context is the town of Lodi, North of Italy (Fig. 1), which ~~on 25-26 November 2002~~ was hit by a severe flood ~~(~~on 25-26 November 2002~~)~~, caused by the overflow of the Adda River as a result of two weeks of heavy rainfalls over North-w~~-W~~est~~ern ~~of~~ Italy. The flood caused severe damage to residential buildings, commercial activities and public services in the area, including the main hospital. Fortunately, no fatalities occurred. The event was chosen as reference for the exercise as it is well documented and characterised by a high availability of hazard, exposure and vulnerability data. ~~In particular, the~~ 2D ~~hydraulic modelling of the event was available~~obtained from a previous study (Scorzini et al., 2018). In detail, ~~, for the 2002 flood~~with respect to the hazard,~~,~~ information on observed water depths was available for more than 260 points within the inundated area, deriving from indications provided by municipal technicians and by citizens in damage compensation requests, as well as from interpretation of photographs taken during or immediately after the flood. These data were used for the validation of the 2D hydraulic simulation of the event: the resulting average absolute differences between observed and calculated water depths within the inundated area ranged from 0.2 to 0.4 m~~–~~, depending on the validation zone in which observed water depth data were aggregated (Scorzini et al., 2018). This is surely a possible source of uncertainty; however, reported differences could be considered to provide relatively small impacts on the damage estimation. Moreover, given that all tested damage model shared the same hazard data, this would be a common source of uncertainty that should not affect the overall results of the blind test.

~~as well as~~Available micro-scale ~~information~~ data on exposure and vulnerability of residential buildings ~~(~~are ~~instead~~ shown ~~see~~ in Table 1~~)~~. ~~Nonetheless~~Altogether, observed damage was known for 345 of the 877 buildings in the flooded area (after hydraulic simulation; Fig. 1), as derived from claims compiled by citizens after the ~~occurrence of the~~ flood~~,~~ to ask for public compensation.


Adda River
Legend
Flooded area (Adda River 2002)
Buildings with damage claims
Buildings without damage claims
0 250 500 1000
m

165

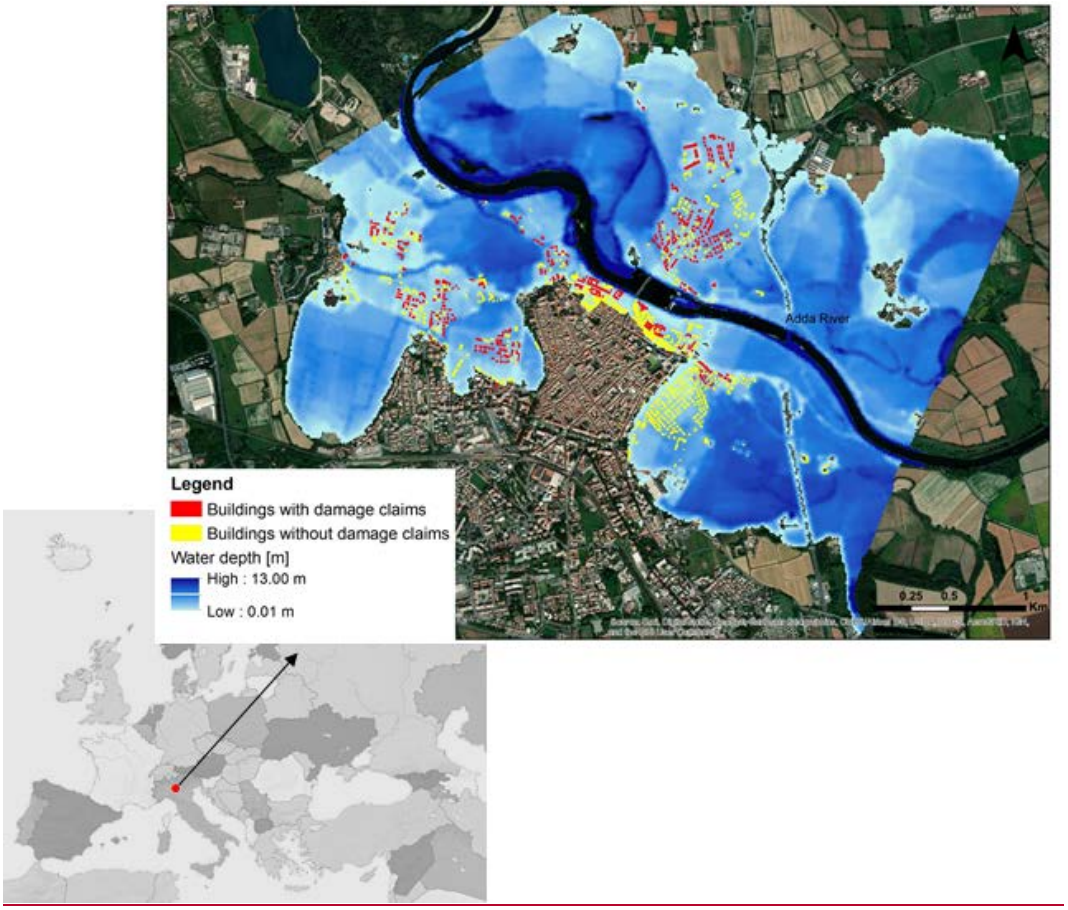


Figure 1: Map of the flooded area and affected buildings.

Claims were mostly collected by the Municipality of Lodi and, in a small part, by the Regional Authority of the Lombardy region after the event. Available claims data, in their original papery form, were then firstly ~~collected~~ acquired and successively stored in a georeferenced digital database, by a team of researchers of Politecnico di Milano in summer 2017. As regards data from the Municipality, original claims were organised in forms, including information on the owner, the address of the flooded building, its typology (e.g. apartment, single house), the number of affected floors, a description of the physical damage and its translation into monetary terms (distinguishing, for the different rooms of the building ~~is made of~~, among damage to walls, windows and doors, floor, systems and content). In few cases, ~~from the description,~~ information on water depth inside the building and on clean-up costs, non-usability of the building and intangible damage (e.g. loss of memorabilia) was also inferred from the qualitative damage description in the forms~~, as well as the value of water depth inside the building; the latter was used for the calibration of the hydraulic model (Scorzini et al., 2018)~~. The quality/reliability of data included in the claims was not uniform, ~~as~~ since only some of the owners justified the costs for fixing ~~the~~ damage by means of invoices. As regards data from the secondary source (i.e. the Regional Authority), they included limited information on the owner, the address of the

flooded building and the monetary value of damage, distinguished in damage to structure and contents.

Table 1: available micro-scale data for the blind exercise.

Data	Variable	Description	Source	Year
Area [m^2]	FA	Footprint area of the building	Regional topographical database	2010
Perimeter [m]	EP	External perimeter of the building	Regional topographical database	2010
Basement	BA	Presence of basement yes/no	Lodi cadastral data	2016
Building type	BT	Type of building ~~among~~ (apartment, detached ~~and~~ or semi-detached house) according to the cadastral data.	Lodi cadastral data	2016
Finishing level	FL	Quality of the building (low, medium or high) according to the cadastral data:	Lodi cadastral data	2016
Building structure	BS	Type of building structure ~~between~~ (masonry ~~and~~ or reinforced concrete) calculated as the most frequent value for the buildings in the census block it owns.	National Institute of Statistics (ISTAT)	2001
Floors	NF	Number of floors calculated as the most frequent value for the buildings in the census block it owns.	National Institute of Statistics (ISTAT)	2001
Level of maintenance	LM	State of conservation (low, medium or high) of the building calculated as the most frequent value for the buildings in the census block.	National Institute of Statistics (ISTAT)	2001
Water_depth [m]	h	Mean value of water depth in the building area.	2D hydraulic modelling	2018
Flow_velocity [$m\ s^{-1}$]	v	Mean value of flow velocity in the building area.	2D hydraulic modelling	2018
Presence of pollutants	q	Presence of fuel spillage or other pollutants	Claims forms / photos of the event	2002
Replacement value [€ m^{-2}]	RV	Reconstruction value of residential building given as a function of the building type and building structure of the building, based on existing literature and official studies	Cresme-Cineas-Ania	2014*
Market value [€ m^{-2}]	MV	Market value of residential buildings, as a function of building type, finishing level and building location	OMI (Osservatorio del Mercato Immobiliare) – Italian real estate and property price database	2014*

* for the objective of the exercise data were discounted to 2002 values

2.2 Methodology

The methodological approach followed in the test included the following steps:

Step 1: identification of damage models to be tested

The choice was based on several considerations: (i) good mastery of the models by the research team (i.e. damage models regularly used or initially developed by the groups), (ii) heterogeneity of the approaches, by considering simple and multi-variable models, empirical and synthetic approaches, absolute and relative models, and (ii) models being calibrated in a different context than the investigated one. The choice converged to the nine models described in Sect. 2.3.

Step 2: implementation of the models to the case study in a blind mode

The models were implemented independently by each research group to ~~estimate~~ calculate damage to all 877 buildings that were exposed to the 2002 Lodi flood, according to the inundation area simulated by the hydraulic model (Scorzini et al., 2018). All the groups used available and common data on hazard, exposure and vulnerability, as described in Table 1. While this step was ~~quite straightforward~~ simple for Italian models (which were originally developed to work with the same kind of data available for the case study), ~~significant~~ some efforts were required for the other models, particularly in the case of multi-variable ones. This is due to a ~~(possible)~~ lack of correspondence/consistency among exposure and vulnerability data available in the different countries, on which damage models are usually based. For instance, correspondence had to be defined among building types classified by the Italian cadastre and the ones adopted by the German and French models and the ones as classified by the Italian cadastre.

The damage ~~estimation~~ assessment was carried out only for building structures, ~~as~~ given that not all models are designed to simulate ~~include estimation of~~ damage to household contents. At this step, observed ~~damages~~ losses were still blinded to the research groups in order to avoid possible bias in the estimation.

Step 3: comparison of model outcomes

Exposure and damage estimates supplied by the different models were compared, at the aggregated and individual level, with the main objectives of (i) understanding the weight of exposure ~~estimation~~ assessment on damage ~~estimate~~ calculation, and (ii) pointing out common ~~patterns or~~ or divergent ~~behaviours in the~~ model outcomes.

Step 4: comparison of model features

Models were compared in terms of trends and variance of individual damage estimates, for homogeneous classes of input variables, by considering one variable at a time. The objective was to understand whether the inclusion of more explicative variables may be considered as a possible source of ~~difference~~variation, as well as to identify the most influencing ~~variables~~ parameters on the final output of the models.

Step 5: comparison between estimates and observations

This phase aimed at investigating the performances of the different models in the analysed context. Calculated ~~D~~damages ~~estimates supplied by the models~~ were compared to observed ~~damages~~ losses coming from claims. ~~Comparison~~ ~~This kind of~~ The comparison ~~analysis~~ was possible only for 345 of the buildings included in the flooded area, for which official claims were available. ~~The objective of this phase was to understand the performance of the models in the investigated context.~~

Step 6: analysis of claims

~~Official claims~~Claim data were analysed with the aim of identifying potential reasons for (in-) consistencies between estimates and observations.

Step 7: synthesis of results

Results obtained in the previous steps were critically analysed in order to gain knowledge on model transferability and reliability of damage estimates, with respect to their implementation in a same case study, and from a modeller's perspective. The analysis was conducted jointly by all groups, in the form of brainstorming, during several remote meetings and one face to face meeting.

2.3 Models

The main characteristics of the selected models are summarised in Table 2 and briefly described hereinafter.

- 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. The recovery cost is assumed equal to 15 % of the exposure, calculated as the market value of the flooded floor(s) based on the footprint area. The ratio between recovery cost and market value is based on the comparison between residential prices for new buildings and buildings requiring renovation (Italian real estate data). The model was created based on expert judgement for the city of Florence (Italy) and applied both at building and census block scale (Arrighi et al. 2018a, 2018b). It has been validated through comparison with other validated models (Arrighi et al., 2018b) and ex-post damage in another Italian context (Scorzini and Frank, 2017).
- **Carisi et al. - MV** (Carisi et al., 2018) is an empirical multi-variable model, which estimates relative building losses considering six explicative variables: maximum water depth, maximum flow velocity, flood duration, monetary building value per unit area (based on market value), structural typology and footprint area of each building (Carisi et al., 2018). Calibration data refer to the inundation event occurred in the province of Modena (Italy) in 2014, when a breach in the right embankment of the Secchia river caused about 52 km^2 of flooded area and € 500 million losses (see, e.g., Orlandini et al., 2015). Observed losses were derived from 1330 claim forms filled by citizens and collected by authorities for the purpose of compensation, while the maximum water depth was reconstructed by means of a fully 2D hydrodynamic model; economic building values per unit area were finally retrieved by the Italian Revenue Agency reports. The model does not consider damage to basements. The model uses the Random Forest approach (Breiman et al., 1984; Breiman, 2001), which is a tree-building algorithm for predicting variables, recursively repeating a subdivision of the given dataset into smaller parts in order to maximize the predictive accuracy. In order to avoid overfitting problems, several bootstrap replica of the learning data are used, for which regression trees are learned, then aggregating the responses from all trees to estimate the final result.

- **Carisi et al. - mono** (Carisi et al., 2018)~~; it~~ is an empirical simple model, calibrated on the previously cited 2014 Secchia flood event. The model supplies the relative damage to building (using the market value to relativize the observed monetary damage when developing the model), as a function of the maximum water depth. The model does not consider basements or garages, for coherence with the calibration context, where most of the buildings do not have these elements.
- The model developed by **CEPRI** (European Center for Flood Risk Prevention, (CEPRI, 2014a))~~; it~~ is a synthetic (expert-based) and multi-variable model that expresses absolute damage as the expected sum of the actions that must be performed after a flood to restore to the pre-flood state, including clean-up costs. ~~-~~The flood parameters taken into account are water depth and submersion duration. The considered building characteristics ~~taken into account~~ are the building type (single storey house, double storey house, or apartment), the floor area, the presence of a basement and its area. For each type of building, one damage curve indicates the damage to structural components, and one the damage to the furniture. Two separate damage curves are used to estimate the damage to the basements contained in houses or apartment blocks. Initially, the model was developed to estimate damage due to all types of floods. Its estimates have been compared to empirical damage due to fast rise floods (CEPRI, 2014a; Richert and Grelot, 2018) and coastal flooding (CEPRI, 2014b). The model was found acceptable in the first context, but needed calibration in the second case. The French State recommends using this model to conduct cost-benefit analyses of flood management projects (Rouchon et al., 2018).
- The model by **Dutta et al.** (2003)~~; it~~ was chosen because it is an early example of a model that describes the relationship between flood intensity and ~~degree of~~ damage. ~~(degree of loss, relative loss) with a mathematical function~~. It is a simple model supplying a relative damage (i.e. the degree of loss that describes the ratio of loss to the replacement value of the whole building) ~~on the basis~~based only ~~of~~ on flood depth; basement, number of exposed floors or other exposure variables are not separate inputs for the model, but are part of its variance. The stage-damage function was calibrated with data published by the Japanese Ministry of Construction, which are based on ~~the~~ site survey data accumulated since 1954. The validation with a flood event of 1996 showed reliable results for urban areas. The replacement value of the building has to be provided as input data.
- **FLEMO-ps** (Flood Loss Estimation MOdel for the private household sector)~~; it~~ is a multi-variable, rule-based model estimating relative monetary flood loss to residential buildings as a function of water depth, building type and building quality, without further differentiating between flooded floors and not explicitly considering the existence of a basement ~~-~~(Thieken et al., 2008). The model is empirically derived from data collected from 1697 households affected by the severe flooding of the rivers Elbe, Danube and some of their tributaries in August 2002 in Germany. It can be applied on both the micro- and the meso-scale. Model evaluations based on historical floods in Germany showed that FLEMO-ps is outperforming traditional stage-damage curves in estimating flood loss in the private household sector, except for damages caused by very high water depths (Thieken et al., 2008).
- The model by **Fuchs et al.** (2019b)~~; it~~ is a simple model, which supplies a relative damage (i.e. the degree of loss that

295 describes the ratio of loss to the replacement value of the whole building) considering water depth, building area (of all floors) and building (replacement) value as input variables. Differently ~~than~~ from other models, it is a function developed for mountain areas, i.e. referring to house building tradition of the Alps and flood processes with sediment transport. It was chosen to test the transferability of a model specialis~~z~~ed for mountain environments to a low-land situation. The model was fitted with empirical damage and hazard data. Model validation took place based on a 5-
300 fold cross validation.

- **INSYDE** (Dottori et al., 2016; Molinari et al., 2017b)~~; it~~ is a synthetic model based on the investigation and modelling of damage mechanisms triggered by floods, developed for the Italian context. The model is based on a what-if analysis, consisting of the simulated step-by-step inundation of the building and in the evaluation of the corresponding damage as a function of hazard and building characteristics. In total, INSYDE adopts 23 input variables, six describing
305 the flood event and 17 referring to building features; among them, there are all the variables available for the case study and included in Table 1. For the remaining ones, default values implemented in the model were adopted in the test. The model supplies damage in absolute terms by considering the replacement/reconstruction value of damaged components, and by referring only to flooded floors (including basement, if present); however, if required, the model can supply also an estimation of relative damage. INSYDE was validated for different Italian flood events and its
310 performance has been compared to those of other existing models (Dottori et al., 2016; Molinari et al., 2017b; Amadio et al., 2019).
- The model by **Jonkman et al.** (2008)~~;~~ is a simple relative damage model considering water depth and building (replacement) value of all floors as explicative variables, developed on the basis of empirical flood damage data collected in the Netherlands in combination with existing literature and expert judgment.~~it is a simple relative damage~~
315 ~~model considering water depth, building area (of all floors) and building (replacement) value as explicative variables, calibrated on loss data in the Netherlands, combined with existing literature and expert judgment.~~ There is no information concerning validation or the robustness of this model. The model is a combined function of content and structure loss. Therefore, to only consider damage on building structure, the original function was rescaled to possibly reach "total destruction" (degree of loss = 1).

320

Table 2: main features of the models implemented in the blind test.

Model	Country and year of development	Hazard context of development	Considered explicative variables	Type of model	Type of results	Economic evaluation	Exposure estimation	Other features
Arrighi et al.	Italy, 2018	Riverine floods	h, FA, BA, economic value of the building	synthetic	relative damage	Recovery (based on market value)	flooded floors, (considering also FL and LM)	– zero-damage threshold at water depth 0.25 m – the model estimates also absolute damage
Carisi et	Italy, 2018	Riverine	h, v, FA,	empirical	relative	market	flooded	

al. - MV		floods	BS, economic value of the building		damage	value	floors (considering also FL and LM)	
Carisi et al. - mono	Italy, 2018	Riverine floods	h, FA, economic value of the building	empirical	relative damage	market value	flooded floors (considering also FL and LM)	
CEPRI	France, 2014	Riverine, coastal floods	h, BT, FA, BA, NF	synthetic	absolute damage	replacement value	flooded floors	– the model estimates also damage to contents (not considered here)
Dutta et al.	Japan, 2003	Riverine floods	h, FA, economic value of the building	empirical	relative damage	replacement value	whole building	
FLEMO-ps	Germany, 2008	Riverine floods	h, q, BT, FL, economic value of the building	empirical	relative damage	replacement value	whole building	– the model is also capable of estimating damage to household contents (not considered here)
Fuchs et al.	Austria/ Switzerland, 2019	Mountain (high velocity) floods, debris flows	h, FA, economic value of the building	empirical	relative damage	replacement value	whole building	
INSYDE	Italy, 2016	Riverine floods	h, v, q, FA, EP, BA, BT, FL, BS, NF, LM	synthetic	absolute damage	replacement value	flooded floors (considering FL and LM)	– the model estimates also relative damage
Jonkman	The Netherlands, 2008	Riverine floods	h, FA, economic value of the building	empirical	relative damage	replacement value	whole building	

3 Results ~~Critical presentation of results~~

3.1 Implementation of the models to the case study in a blind mode

With the aim of understanding the impact of exposure estimation on damage assessment and identifying possible common features in the results, Table 3 shows the total exposure and loss figures obtained by ~~the implementation of~~applying the nine models to all ~~the 877~~ buildings ~~included in the~~within the simulated inundation area (877 in total; see Figure 1); n~~, with respect to both the monetary value of exposed assets and the monetary value of damage.~~ ote that at this stage of the analysis damage observations were not considered yet for ~~no~~ comparison ~~with damage observations was made~~purposes ~~at this stage of the analysis~~ (see Section 2).

Total exposure estimates differ among the models by a maximum factor of 2.75. With respect to the mean exposure value ~~of~~

~~total exposure estimates~~, single estimations diverge instead by a maximum factor of 1.77. ~~Total exposure estimates diverge by a maximum factor of 2.75, and by a maximum factor of 1.77 with respect to the average estimation.~~ These ~~se~~ significant differences mainly result from the fact that some models ~~evaluate as~~calculate exposure as the monetary value of flooded floors, while others refer to the whole building (see Table 2). Indeed, focusing ~~the comparison~~ on the four models that consider only flooded floors (i.e. Arrighi et al., Carisi et al.-MV, Carisi et al.-mono, and INSYDE, see Table 2), total exposure estimates differ by a maximum factor of 1.22. ~~When comparing models that focus only on flooded floors, estimates differ by a maximum factor of 1.22.~~ Minor differences are due to the (non-)consideration of the presence of a basement as well as to the adoption of replacement/recovery values rather than market values as parametric cost for the estimation. These results point out that a first source of variability among model outcomes lies in the approach for exposure assessment.

Total damage estimations differ among the modelling approaches by a maximum factor of 12.6, which is limited to 3.1 with respect to the mean value of total damage estimations~~,Total damage estimations differ by a maximum factor of 12.6, and by a maximum factor of 3.1 with respect to the average estimation~~, suggesting that the shape of the damage functions exacerbates the variability of models' outcomes due to exposure estimation.

Similar conclusions can be drawn when looking at individual (i.e. building by building) ~~building~~ estimations reported in Fig. 2 (exposure values) and Fig. 3 (damage values). Individual estimations of exposure differ by a mean factor of 3.5.~~The mean difference among individual estimations of exposure amounts to 3.5, whereby most of the models rather differ by a factor of approximately 2.~~ The models of Fuchs et al., Jonkman et al., Dutta et al. and FLEMo-ps use the replacement value of the whole building as a reference for calculating the degree of loss and are thus relying on sensibly higher exposure values than others. Individual damage estimates differ on average by a factor of 28, with the~~8, with the more frequent factor around 10.~~ ~~Highest~~ highest differences ~~are~~ due ~~by~~ to the models of Fuchs et al. and Dutta et al.~~, which estimate the highest~~(maximum expected damage)~~,~~ and ~~by~~ to the model of Arrighi et al.~~., which estimates the lowest damage~~ (minimum expected damage). Such results can be partly explained by the adoption of the whole building value for exposure estimation (see also Sect. 3.2)~~,~~ as regards high estimations, and by the zero damage threshold for water depths lower than 0.25 m~~,~~ for low estimations. In detail, the weight of the zero damage threshold on the final damage figure has been calculated as a percentage ranging from 7 to 32 %, depending on the considered model.

~~**Table 3: Estimates of the monetary value of exposed assets and damage, for all the buildings in the flooded area. The first column reports the total value of exposed assets (n.a.= not applicable). The second column reports the total damage and the unit damage per m^2 (in brackets). The third and the fourth columns report the ratio between estimates and mean value of estimates (reported in the last row), for exposed assets and damage respectively.**~~

~~Model~~	~~Monetary value of exposed assets [M€]~~	~~Monetary damage [M€] (Unitary monetary damage [€ m^{-2}])~~	~~Monetary value of exposed assets/mean value [-]~~	~~Monetary value of damage/mean value [-]~~
~~Arrighi et al~~	~~392~~	~~12 (35)~~	~~0.78~~	~~0.25~~
~~Carisi et al. - MV~~	~~368~~	~~20 (80)~~	~~0.73~~	~~0.40~~
~~Carisi et al. - mono~~	~~368~~	~~30 (118)~~	~~0.73~~	~~0.59~~

CEPRI	n.a.	25 (71)	n.a.	0.50
Dutta et al.	889	155 (225)	1.77	3.10
FLEMO-ps	468	58 (230)	0.93	1.15
Fuchs et al.	889	102 (147)	1.77	2.03
INSYDE	395	21 (69)	0.79	0.41
Jonkman et al.	889	29 (42)	1.77	0.58
Mean	502	50	-	-

Table 3: Estimates of the monetary value of exposed assets and damage, for all the buildings in the flooded area. The first column reports the total value of exposed assets (n.a.= not applicable). The second and the third column report, respectively, the total damage and the unit damage per m^2. The fourth and the fifth columns report the ratio between estimates and mean value of estimates (reported in the last row), for exposed assets and damage, respectively.

Model	Monetary value of exposed assets [M€]	Monetary damage [M€]	Unitary monetary damage [€ m^{-2}]	Monetary value of exposed assets/mean value [-]	Monetary value of damage/mean value [-]
Arrighi et al	392	12	35	0.78	0.25
Carisi et al. - MV	368	20	80	0.73	0.40
Carisi et al. - mono	368	30	118	0.73	0.59
CEPRI	n.a.	25	71	n.a.	0.50
Dutta et al.	889	155	225	1.77	3.10
FLEMO-ps	468	58	230	0.93	1.15
Fuchs et al.	889	102	147	1.77	2.03
INSYDE	395	21	69	0.79	0.41
Jonkman et al.	889	29	42	1.77	0.58
Mean	**502**	**50**	-	-	-

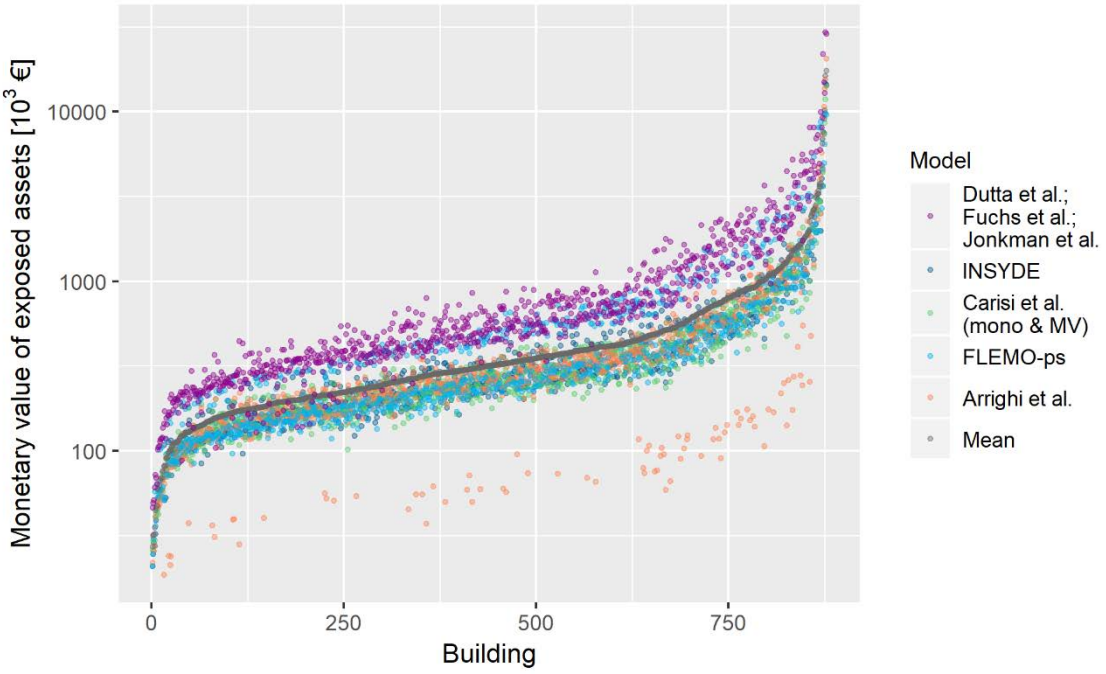


Figure 2: Individual estimates of the monetary value of the exposed assets for all the buildings in the flooded area. Data are ordered according to increasing value of mean estimate (in grey).

Figures 2 and 3 further highlight a common trend in exposure and damage ~~estimates~~ values supplied by the different models, also confirmed in Fig. 4 and 5, showing the Pearson's correlation coefficients for individual (i.e. building by building) exposure and damage estimates. The figures ~~, which~~ show a very high correlation of exposure estimations and a weaker, but still notable, correlation of damage estimations. This finding supports previous results on the importance of damage functions in determining the main differences in model outcomes. In particular, Fig. 5 shows that a higher correlation exists between absolute damage estimates supplied by the two synthetic models INSYDE and CEPRI, among multi-variable models (INSYDE, CEPRI, Carisi et al. - MV and FLEMO-ps), and among simple models (Carisi et al. - mono, Dutta et al., Fuchs et al. and Jonkman et al.), which reflects the consistency between models based on comparable conceptual frameworks.

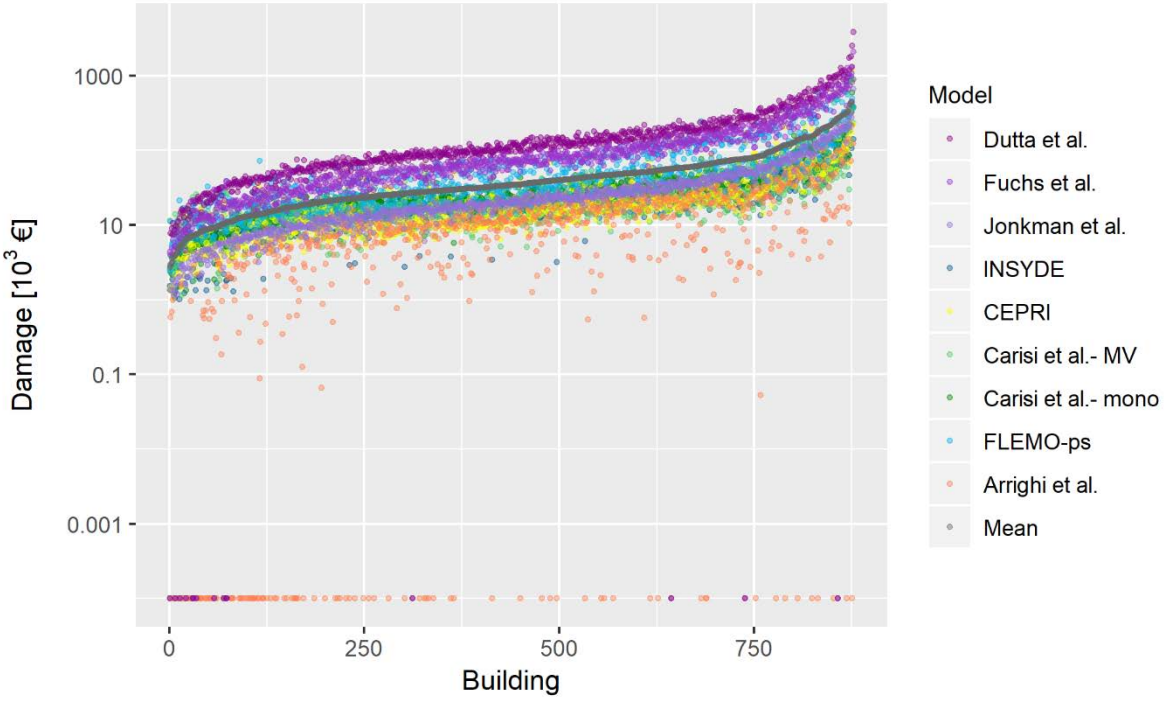


Figure 3: Individual estimates of the monetary damage for all the buildings in the flooded area. Data are ordered according to increasing value of mean estimate (in grey). Zero damages are due to the modelling assumptions behind the specific damage models (i.e. 0.25 m water depth threshold for damage occurrence in Arrighi et al. and 0.01 m water depth threshold in Dutta et al. and Jonkman et al. to distinguish between flooding and surface water runoff)

A c~~C~~omparison between correlation coefficients for absolute and relative damage estimations in Fig. 5 conversely highlights the importance of exposure assessment on the final damage figures. For instance, the low correlation among absolute damage estimates supplied by the model of Arrighi et al. with those from similar models (i.e. simple, low-variable models like Carisi et al. - mono, Dutta et al., Fuchs et al. and Jonkman et al.) can be explained by the fact that the approach adopted by Arrighi et al. for the evaluation of exposure is considerably~~significantly~~ different ~~than~~ from those adopted by the other comparable models; specifically, the model calculates the monetary value of damage as a function of the recovery cost, which is assumed equal to 15 % of the market value of exposed floors (see Sect. 2). Accordingly, when relative damage estimations are considered, the values of Pearson's correlation coefficient increase. The weight of exposure assessment is also evident when correlation among absolute damage estimates supplied by the four simple, empirical models (i.e. Carisi et al. – mono, Dutta et al., Fuchs et al. and Jonkman et al.) are considered, with models of Dutta et al., Fuchs et al. and Jonkman et al. using the same exposure assessment approach (see Sect. 2) and thus being more correlated among them than with the model Carisi et al. – mono; on the opposite, when relative damage estimations are considered, the correlation coefficients for the four models are comparable. At last, the weight of exposure arises when correlation between absolute damage estimates supplied by Carisi et

al. – mono versus INSYDE are considered. The couple ~~compares~~ consists of two conceptually different models (in particular, a simple, empirical model versus a multi-variable model~~s~~), but it show~~sing~~ high correlation. This can be explained by the adoption of very similar approaches for exposure estimation by the considered models (see Sect. 2 and Table 3); in fact, when relative damage estimates are considered correlation decreases.

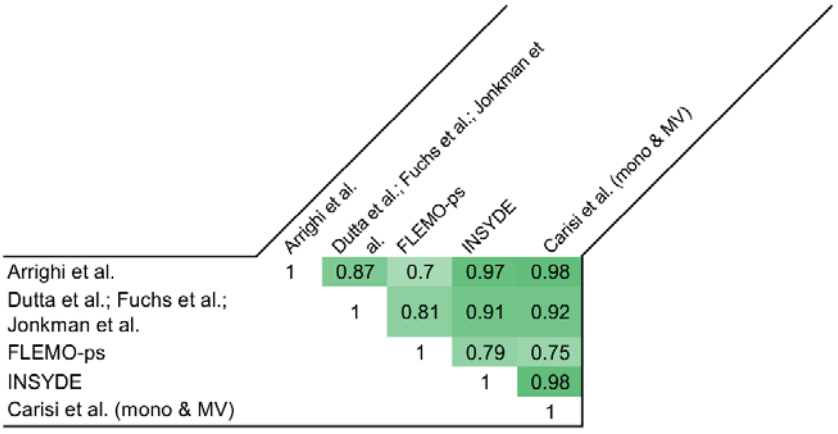


Figure 4: Pearson's correlation coefficient for individual exposure estimates supplied by the models with reference to all the buildings in the flooded area (the darker the colour, the stronger the correlation).

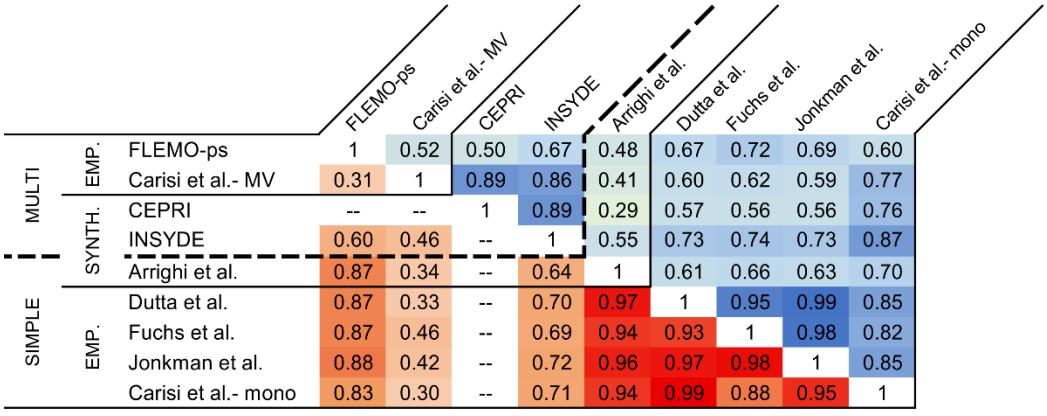


Figure 5: Pearson's correlation coefficients for absolute damage estimations (top-right of the matrix, in blue) and relative damage estimations (bottom-left of the matrix - in red) supplied by the models with reference to all the buildings in the flooded area (the darker the colour, the stronger the correlation).

3.2 Role of input variables in the determination of divergent models' outcomes ~~Critical analysis of models (analysis of~~

~~models' behaviour with respect to explicative variables)~~

In order to explain the differences observed in the blind implementation, models were compared in terms of trends and variance of individual damage estimates, for classes of values of input variables, and by considering one variable at a time. The objectives of the analys~~e~~is were to investigate whether the consideration of a specific input variable influences the outcome of a model with respect to the other ones, whether the inclusion of more explicative variables may be considered as a possible source of variation, and to identify the most influencing parameters on the final output of the models.

The input variables considered were: ~~In order to investigate divergent behaviours in model outcomes, individual damage estimates were analysed for different classes of the influencing input variables (see Table 1), namely:~~ the mean value of the water depth in the building area (h), the footprint area of the building (FA), its external perimeter (EP), the presence of basement (BA), the building type (BT), the building structure (BS), the finishing level of the building (FL), the number of floors (NF), and the level of maintenance (LM). The results are shown in the boxplots reported in Fig. 6 and 7.

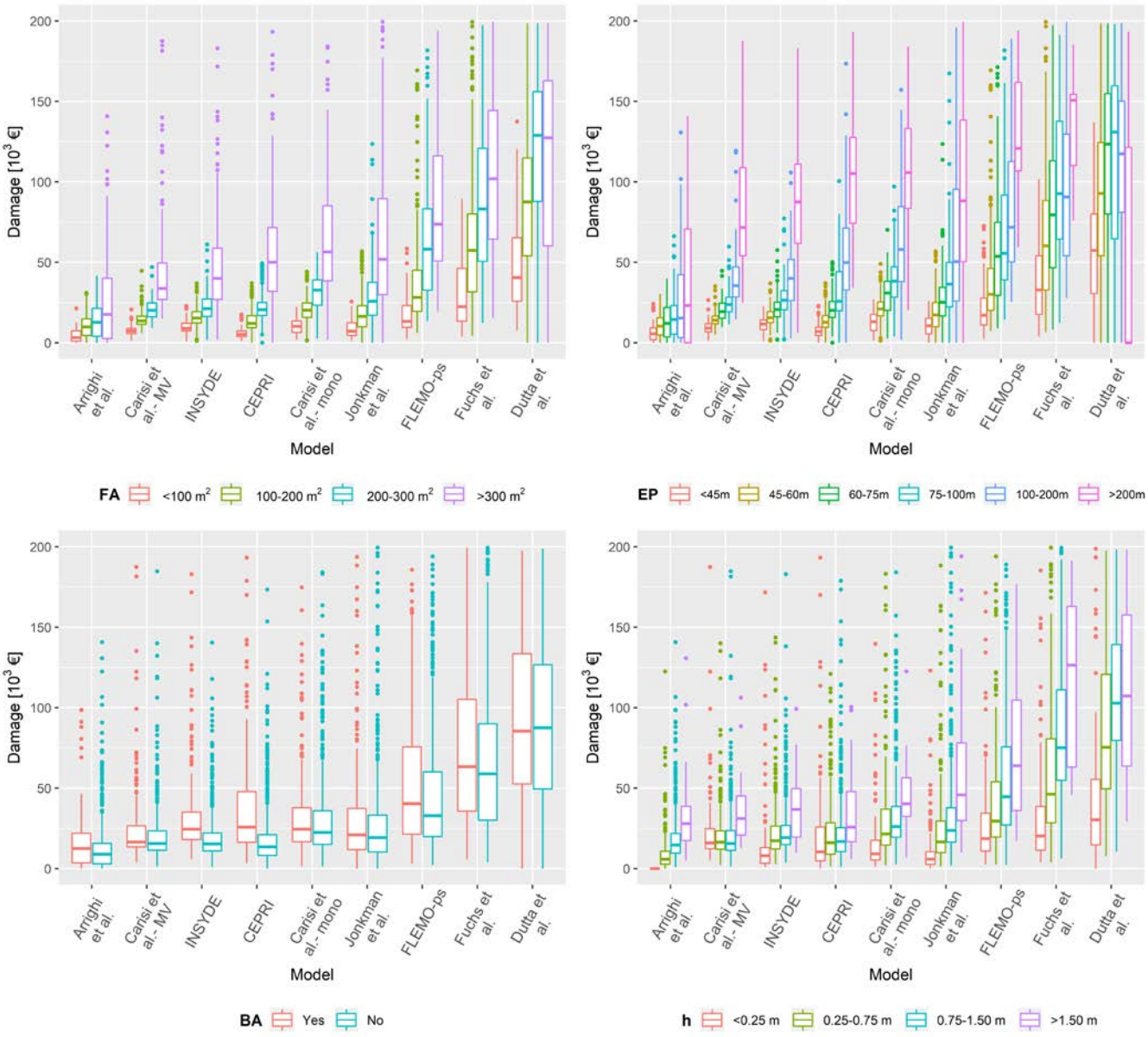
An expected increasing trend in damage as a function of the variables related to the extensive properties of the buildings (FA and EP) can be seen, with limited data variance in the case of those models considering other explicative variables than FA (e.g. EP), as INSYDE. As highlighted in the previous section, the models of Dutta et al. and Fuchs et al. show markedly different results, i.e. higher estimates than other models in all classes. This cannot be totally attributed to the fact that such models ~~use~~ consider the whole building ~~value as exposure value~~for calculating exposure, as this is true also for the model of Jonkman et al., which supplies results that are comparable ~~results with respect~~with the ones of ~~the to~~ other models. Instead, one possible reason ~~relates~~ may be found in~~to~~ the different origins of the models. In fact, contrarily to all other models, the model of Fuchs et al. was developed for mountainous regions where floods are usually characterised by high sediment transport and deposition, which increases the damage, other variables being equal. In the case of Dutta et al. the detection of the reason for the remarkably higher damages ~~estimations~~ is more elusive, ~~as~~ given the lack of ~~there are no further~~ detailed information on~~s~~ ~~in the~~ model derivation, ~~and therefore,~~which makes the original model environment ~~is~~ not known ~~n~~either for hazard ~~n~~or exposure variables. In addition, this model is based on survey data collected since 1954 in Japan, meaning that the data used might not be consistently representative for the current flood vulnerability ~~of today~~ (and in a European environment). The general increasing variance of the estimates with FA and EP classes can be explained by the intrinsic variability of the features characterising larger buildings: they can be apartment buildings rather than semi-detached houses or big villas, with one or more floors; moreover, in the case of apartment buildings, the level of maintenance can change from flat to flat.

Figure 6 indicates the importance of BA as an influencing variable in modelling flood damage for the given event. This is particularly evident in the results provided by CEPRI and INSYDE, which estimate median damages ranging respectively from 13~~.~~600 € and 15~~.~~400 € for buildings without basement to 26~~.~~300 € and 24~~.~~500 € for buildings with basement, as opposed to the performances of other models, which did not differ significantly for the two building categories.

Regarding damage estimates for different water depth classes, Fig. 6 indicates an acceptable convergence among model results, especially for the shallower water depth classes, if excluding the results of the models of Dutta et al. and Fuchs et al. (as

discussed earlier). However, larger differences are apparent for the highest water depth class (h > 1.5 m). Overall, this result seems reasonable as most of the tested models were calibrated and/or validated for flood events characterised by shallow or medium inundation depths.

Finally, as also emerged in previous studies (Wagenaar et al., 2017; Amadio et al., 2019), Fig. 7 denotes that other variables related to building features do~~es~~ not significantly influence model behaviour. Larger scatter is observed only for the "Apartment" category, which is intrinsically characterised by larger variability, especially in terms of extensive parameters.



455 **Figure 6: Boxplots of damage estimates obtained with the tested models, for different classes of: footprint area – FA (Top-left), external perimeter – EP (Top-right), presence of basement – BA (Bottom-left) and water depth – h (Bottom-right). Models are organised according to increasing value of total damage estimates.**

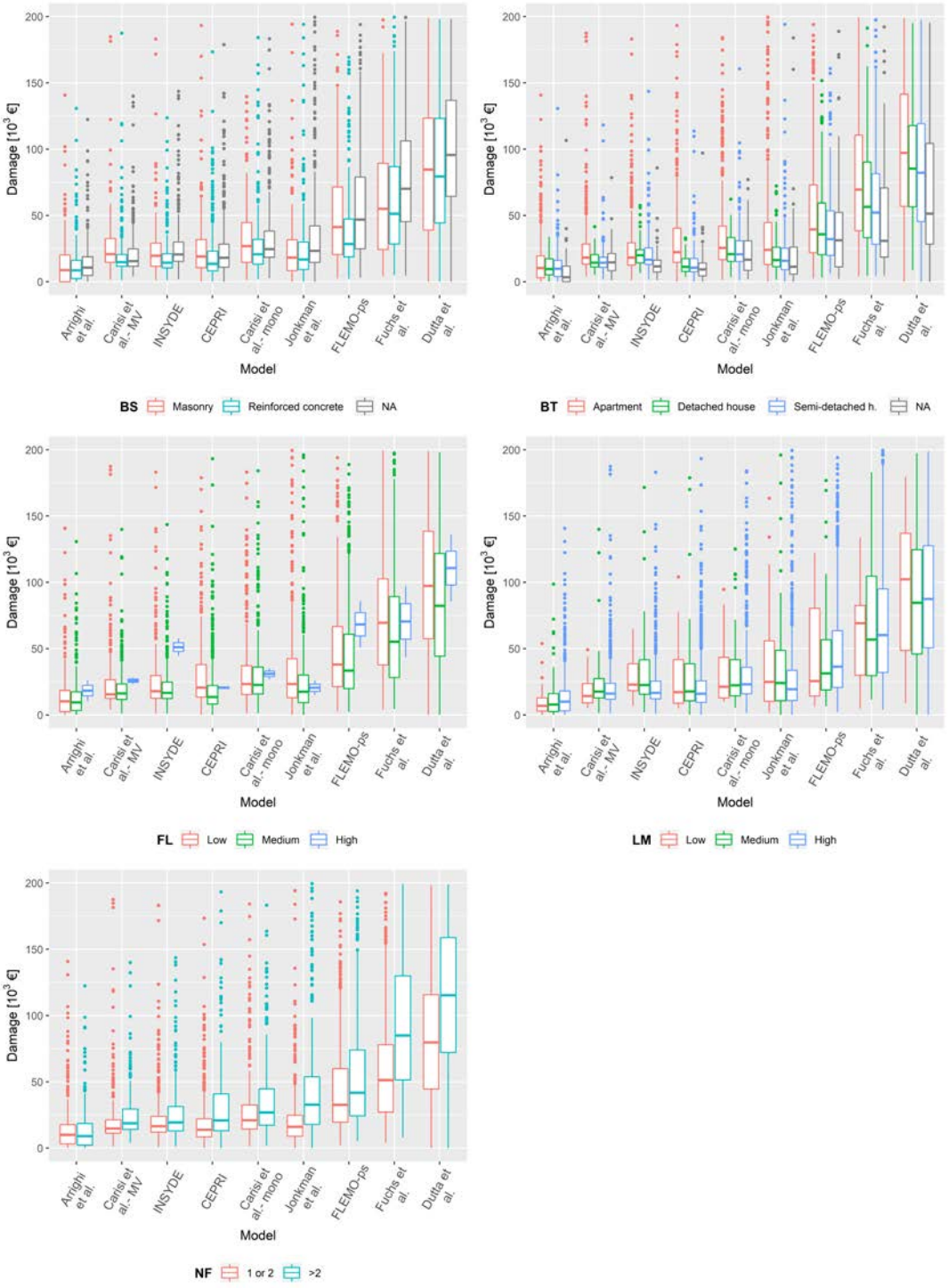


Figure 7: Boxplots of damage estimates obtained with the tested models, for different classes of building structure – BS (Top-left), building type – BT (Top-right), finishing level – FL (Middle-left), level of maintenance – LM (Middle-right) and number of floors – NF (Bottom-left). Models are organised according to increasing value of total damage estimates.

3.3 Comparison between estimates and observations

In order to gain knowledge on ~~the~~ models' reliability in the investigated context, estimated losses were compared to observed damages ~~observations provided in the form of official~~derived from claims. For this purpose, a subset of the buildings within the simulated inundation area was considered~~, (~~ given that claims presented by private owners were available for only 345 buildings~~)~~. Table 4 summarises the results of the sensitivity analysis by comparing~~s~~ the total observed damage to the total damage estimates obtained ~~by~~ with the implementation of the nine models to the subset of buildings. The table confirms the results ~~from~~ presented in Sect. 3.1 (i.e. models estimations differ by a factor of around 13) and highlights the systematic overestimation provided by the ~~of~~ models with respect to observed damage, up to a maximum difference ratio of 13.97. The table also ~~shows~~ indicates the better performances of the Italian/local models (marked with the "IT" suffix in the table), with Arrighi et al. showing the lowest difference. However, by looking at its features, it is possible to state that even this last model tends to overestimate damage. First, because it does not consider clean-up costs (like INSYDE and CEPRI), which are instead included in the observations. Second, because the lower value of the total damage with respect to other models is partly due to the effect of the zero damage threshold for water depths lower than 0.25 m (see Sect. 3.1); indeed, as highlighted in Fig. 8 (showing the comparison between individual observed and estimated damages), a zero damage was expected by this model also for those buildings which experienced a significant loss. Interestingly, Table 5 finally shows that some of the ~~foreign~~ imported models perform similarly or better than Italian models, with specifically high performance of CEPRI.

~~**Table 4: Observed damage data versus estimates of the total monetary damage for the subset of buildings with claims (n.a.= not applicable). The second column reports the total damage and the unit value of damage per m^2 (in brackets). Mean value of estimates is reported in the last row. The third column reports the ratio between estimates and observed damage.** "IT" suffix is used to mark Italian models.~~

~~**Model**~~	~~**Monetary damage (M€) (Unitary monetary damage [€ m^{-2}])**~~	~~**Calculated damage/observed damage [-]**~~
~~*observed*~~	~~*6 (n.a.)*~~	~~-~~
~~Arrighi et al. (IT)~~	~~6 (43)~~	~~1.00~~
~~Carisi et al. – MV (IT)~~	~~8 (85)~~	~~1.4~~
~~Carisi et al. – mono (IT)~~	~~12 (132)~~	~~2.19~~
~~CEPRI~~	~~10 (74)~~	~~1.72~~
~~Dutta et al.~~	~~77 (265)~~	~~13.97~~
~~FLEMO-ps~~	~~30 (320~~	~~5.30~~
~~Fuchs et al.~~	~~50 (171)~~	~~9.03~~
~~INSYDE (IT)~~	~~9 (85)~~	~~1.69~~
~~Jonkman et al.~~	~~14 (49)~~	~~2.61~~
~~Mean~~	~~24 (n.a.)~~	~~4.06~~

Table 4: Observed damage data versus estimates of the total monetary damage for the subset of buildings with claims (n.a.= not applicable). The second and the third columns report, respectively, the total damage and the unit value of damage per m^2. Mean

value of estimates is reported in the last row. The fourth column reports the ratio between estimates and observed damage. Suffixes are used to track the original country of the models (IT=Italy, FR=France, JP=Japan, DE=Germany, AT= Austria, NL= The Netherlands).

Model	Monetary damage (M€)	(Unitary monetary damage [€ m^{-2}])	Calculated damage/observed damage [-]
observed	*6*	*60*	-
Arrighi et al. (IT)	6	43	1.00
Carisi et al. - MV (IT)	8	85	1.4
Carisi et al. – mono (IT)	12	132	2.19
CEPRI (FR)	10	74	1.72
Dutta et al. (JP)	77	265	13.97
FLEMO-ps (DE)	30	320	5.30
Fuchs et al. (AT)	50	171	9.03
INSYDE (IT)	9	85	1.69
Jonkman et al. (NL)	14	49	2.61
Mean	24	n.a	4.06

Figure 8 generally corroborates findings of Sect. 3.1, ~~i.e.~~depicting ~~–~~a common trend in the models with largely different individual damage estimates. Moreover, it also emphasises the overestimation made by the models with respect to observations, with ~~observations~~ the latter not showing the common trend followed by the models. This evidence is supported by the results of the correlation analysis (Table 5), which reveals only marginal correlation between ~~model estimates~~calculated losses and reported claims. On the contrary, the high correlation among models (see Fig. 5) raises the question of whether reported claims and damage estimation are comparable.

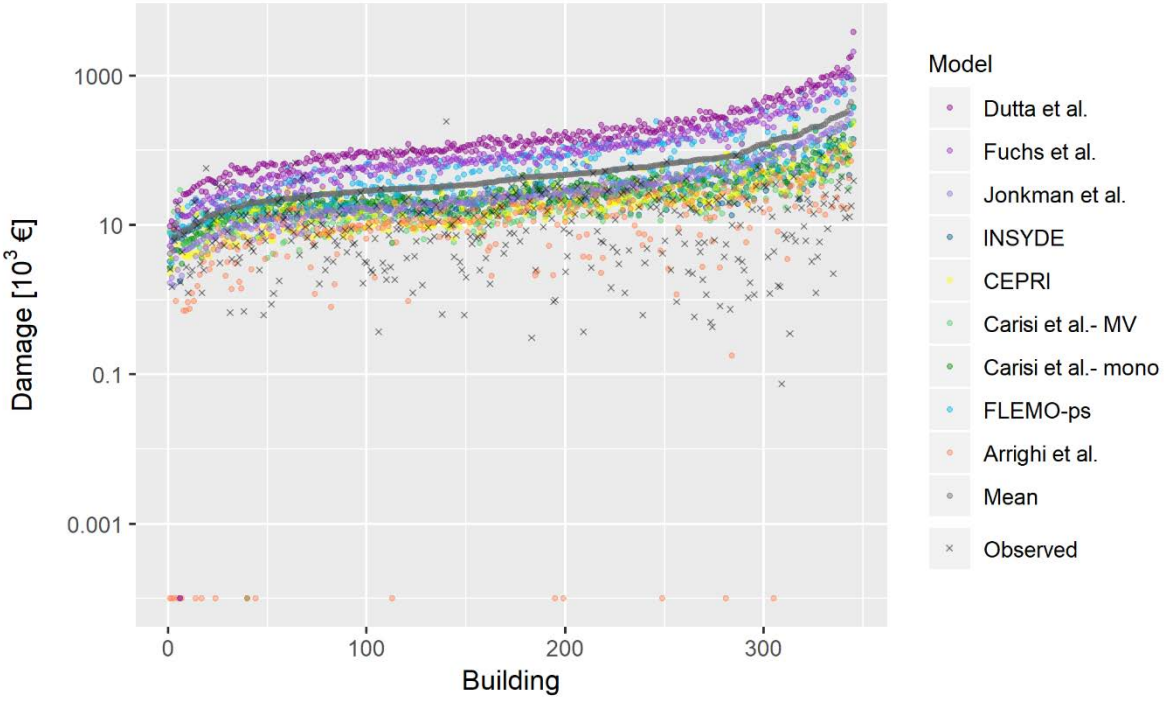


Figure 8: Observed damage versus individual estimates of the monetary damage for the subset of buildings with claims. Data are ordered according to increasing value of mean estimate (in black).

Table 5: Pearson's correlation coefficient of observed damage and estimates supplied by the models with reference to the subset of buildings with claims. ~~Suffixes~~The acronyms in parentheses indicate ~~are used to track~~ the original countries~~y~~ of the models (IT=Italy, FR=France, JP=Japan, DE=Germany, AT= Austria, NL= The Netherlands).

	Observed
Arrighi et al. (IT)	0.26
Carisi et al. - MV (IT)	0.10
Carisi et al. – mono (IT)	0.12
CEPRI (FR)	0.15
Dutta et al. (JP)	0.13
FLEMO-ps (DE)	0.13
Fuchs et al. (AT)	0.15
INSYDE (IT)	0.18
Jonkman et al. (NL)	0.13

3.4 Analysis of damage claims

In order to explain the differences between model results and observations, a thorough analysis of claims data was carried out. Given the general overestimation ~~made~~ provided by the models, first we focused our attention on 44 buildings that are characterised by very low values of observed damage (less than 1500 € in 2002 currency), referred to as "outliers" hereinafter.

Table 6 reports the mean value of water depth, footprint area and external perimeter (i.e. the variables which most influence damage according to the analysis performed in Sect. 3.2) calculated for this subset of buildings and for all the buildings with claims. Table 6 indicates that low damages cannot be explained by significant differences in these influencing variables, given that both datasets show comparable values~~. in, as~~. Moreover, based on informal conversation with representatives of the Committee of Flooded Citizens in Lodi, it is possible to postulate that existing outliers cannot even be explained by the adoption of individual mitigation actions (like temporary flood barriers or pumps), because no official flood warning was issued and, consequently, no lead time was available to undertake precautionary measures. Finally, from the analysis of building pictures available in Google Street View, we can state that outliers are not due to the presence of steps or other elements which increase the height of the building with respect to the ground level, reducing its ~~exposure to hazard~~vulnerability.

Table 6: Mean value of water depth (h), footprint area (FA) and external perimeters (EP) for all buildings with claims and for the outliers' subset.

Dataset	Mean value of influence variables		
	H [m]	FA [m^2]	EP [m]
outliers	0.79	264.80	78.07
all claims	0.86	265.56	77.32

On the contrary, examining in detail the outlier claims, the following evidence~~s~~ arose:

- 27 % of outliers refer to claims with no detailed information about the type of damage, hindering the thorough understanding of low loss values in these cases;
- 32 % of outliers can be explained by the fact that declared damage regards only garages or boilers, while damage models typically assume a residential use of the building, with the presence/damage of all technical systems (i.e. heating, electrical, and water);
- 41 % of outliers refer to paltry claims, even in case of significant water depths (around 1 m), which are mostly related to painting of walls and replacement of doors and windows.

In view of the large proportion of paltry claims, it was attempted to understand the causes of declared damages. For this, we calculated the frequency of damage occurrence to different building components (i.e. damage to walls, damage to floor, damage to doors and windows and damage to systems) in the different claims and for three water depth classes (Fig. 9). Findings reveal an unexpected behaviour with respect to existing knowledge on damage mechanisms~~; and~~ in particular:

- damage to floors is found to be declared mostly for water depths higher than 1.5 m, although in principle this type of damage should be poorly related to water depth;
- frequency of damage to doors and windows decreases moving from the middle to the highest water depth class, as opposed to expectations (because of the occurrence of damage to windows with higher water depths);

- no damage to water, sanitary and heating systems is found to be declared for water depths higher than 1.5 m, contrarily to what can be expected by considering the typical height of the technical installations in Italian houses (Dottori et al., 2016).

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 budget~~money~~ required to replace or reconstruct the whole physical damage suffered by the building, but rather to the amount of money needed to bring the building back to a desired level of functionality, according to the financial resources of the owner: for this reason, for example, not all flooded doors are replaced ~~or~~and flooded floors ~~are not~~ always rebuilt. This would explain why synthetic models overestimate observed damage, as they are usually based on full replacement/reconstruction costs. Likewise, it would explain why the model by Arrighi et al. performs better than others: indeed, the recovery value adopted by this model is defined as the average difference between the market value of new buildings and that of equivalent, older buildings requiring renovation. It is then sensible that this value reflects a balance between the two opposite extreme behaviours of buyers (which, ~~on~~ in turn, depend on their financial resources): i.e. to completely renovate the building or to bring~~ing~~ ~~the building~~it back to a minimum level of functioning. In our view, such behaviours can be compared with those of flooded owners.

Moreover, declared monetary damage is strongly correlated to the expectations that citizens have to be reimbursed. This expectation is low in Italy, when in most cases limited funding is available for the compensation of private damage, which implies strict criteria and thresholds for compensation (often much lower than the effective damage). In addition, all costs must be proved by the citizens by means of official invoices. For all these reasons, citizens often prefer taking advantage of the "black market" rather than declaring damage (Cellerino, 2004). This would also explain why empirical models (derived from claims) developed in regions with high expectations and then high values of declared damage (like Germany~~ or the Netherlands~~), overestimate the observed damage in this case study.

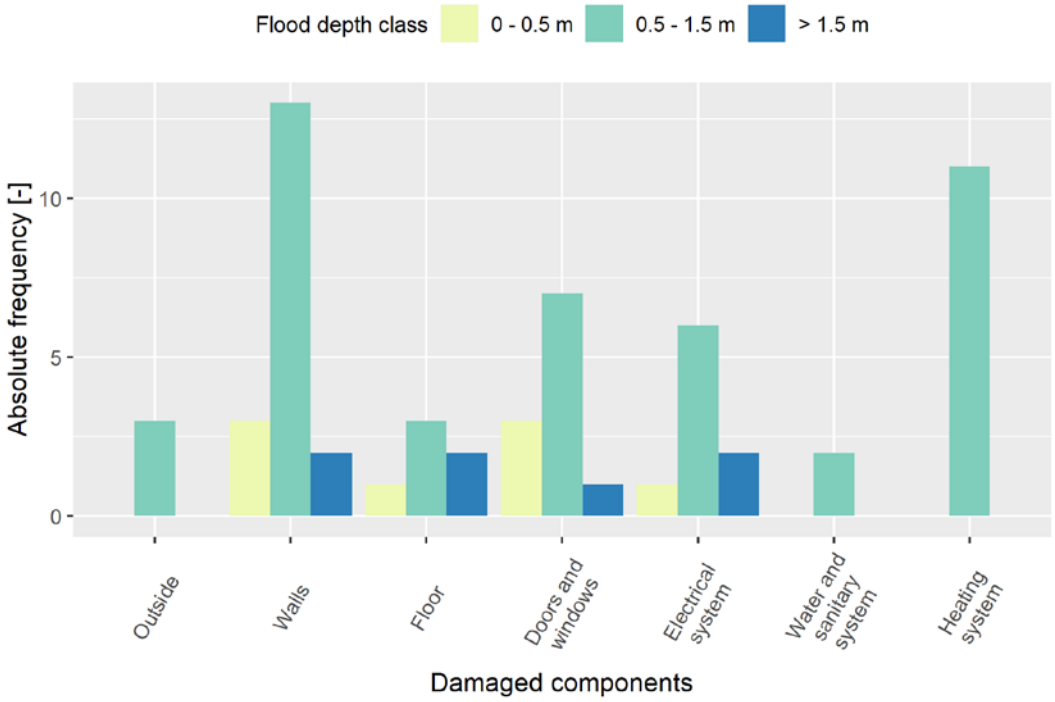


Figure 9: Absolute frequency of declared damage to the different building components in the outlier dataset for different water depth (h) classes.

From another perspective, in order to explain the scatter that is generally observed in real damage data with respect to water depth (note that the value of ~~the~~ Pearson's correlation coefficient between observed damage and water depth is 0.11), we focused the attention on 13 paired buildings, whereby the term "paired" refers to buildings with the same vulnerability characteristics (i.e. building type, building structure, level of maintenance and finishing level) as well as similar values of hazard parameters (i.e. water depth and flow velocity), but significant difference in the declared unit~~ary~~ damage (€ m^{-2}).

The analysis revealed that:

- considerable differences are attributable to declared or undeclared replacement costs of systems, rather than of doors and windows; this can be explained again by what is considered as monetary damage by citizens.
- in other cases, costs related to similar damage (e.g. cost of painting, cost of replacement of doors) differ a lot, even by a factor of 10. This discrepancy might be explained by wrong assumptions concerning the finishing level and/or the building type. More specifically, the actual conditions of buildings with high damage values could have been better than what was assumed for the blind test, using cadastral data as reference (see Table 1).
- sometimes the above two factors add up, further increasing the differences among paired buildings in terms of declared damage.

Scatter in claims data can then be partially explained by the influence of local parameters (like the finishing level or the building

type) which are difficult to assess at the micro-scale without a detailed field survey; nonetheless, it seems that the influence of such parameters on damage estimation for the analysed models is very low (see Sect. 3.2) so that the latter are reliable only when applied at the meso-scale.

Overall, the analysis of claims highlighted that observed damage data need to be carefully analysed before being used for model validation, since their comparability with damage estimates is not always guaranteed.

4 Discussion

Results from the previous analyses were critically analysed in order to gain general knowledge on the transferability of damage models and reliability of damage estimates, and, in particular, to answer to the two specific research questions set in the Introduction.

Concerning the performance of local versus imported models, the blind test corroborated literature results (Cammerer et al., 2013), suggesting that models transferability depends on the consistency between the context of implementation and the original calibration context, as far as both hazard and exposure/vulnerability features of exposed buildings are concerned. In fact, in the blind test, models developed for the Italian territory and for riverine floods performed generally better than models derived in other countries or for different flooding features, e.g. mountain areas. Such a result was not surprising as models providing good results have proven to perform well also in other Italian validation case studies, (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 the two models by Carisi et al., which were tested in other Italian flood events (Amadio et al. 2019)). On the contrary, the imported model of Dutta et al. was already found to not properly work in Italian cases (Scorzini and Frank 2017)). Still, the analysis of damage claims revealed that, as far as empirical models are considered, transferability could depend also on comparability of the compensation contexts, given that observed losses on which empirical models are calibrated may depend on citizens' expectations of reimbursement.

Regarding instead the second question, literature suggests that the inclusion of several influencing variables should increase the accuracy of a model (Merz et al., 2013; Schröter et al., 2014; Van Ootegem et al., 2018). Still, the blind test highlighted that such an evidence can be invalidated by the lack of availability/consistency of input data between the calibration and the implementation context. Indeed, the models implemented considered in the blind test were designed to be used with the type of data usually available in the original context, which generally differ from the data available in the Lodi case study, (i.e., models use different proxy variables for the same explicative parameters). For this reason, a variety of assumptions had to be undertaken to allow the application of a model in the case studygiven area (see Sect. 2). For example, the building categories (BT) assumed by CEPRI ("apartment"/ "single storey building"/ "multi-storeys buildings") are different than the Italian ones ("apartment"/ "detached"/ "semi-detached") so that a correspondence has to be defined, also on the basis of the number of floors (NF); specifically apartment is defined by BT ="apartment", "single-storey" is defined by BT = "detached" or "semi-detached" and NF = 1, "multiple-storeys" is defined by BT ="detached" or "semi-detached" and NF > 1. Correspondence

among building categories was defined also for the implementation of FLEMO-ps, although in this case the task was quite straightforward, since the German building categories are almost coincident with the Italian ones (FLEMO-ps distinguishes between "Multi-family house" / "Semi-detached house" / "One-family home"). Assumptions on input variables may reduce the reliability of the original model because of an improper/inaccurate "adaptation" of the available data, thus reducing the advantage of using many variables. This also explains why the simple models by Jonkman et al. and Carisi et al. - mono provided comparable or better results ~~to~~ than those obtained from multi-variable models like FLEMO-ps or CEPRI. ~~couldifwereto~~Also, the use of ~~such~~ additional variables may have different impact depending if, in the application area and differently for the original model development strategy, this information is retrieved at the building scale or known as aggregated variable. Consultations of experts with local knowledge were needed to ~~ensure~~ help in the correct interpretation and use of the available input data for the Lodi case study.

Importantly, the blind test highlighted that none of the tested models (being them local or imported, simple or multi-variable) seemed appropriate to estimate flood damage at the building scale in the given context; still, models' performance improved when aggregated damage data were taken into account. In fact, considering the 345 buildings for which a claim was known, all models' estimates differed significantly individually (Fig. 8), but some of them indicated a total damage figure close to the ~~total of claims~~observations (Table 4). Besides the already discussed potential biases of claim data, this duality suggests that model~~s~~ uncertainty may be balanced in ~~the~~ aggregated results, i.e. the lump-sum might be more reliable than the individual results. This raises the question of which is the right spatial scale (that is the level of complexity) of analysis to get reliable results, and for which objective. For example, by implementing the simpler, lump-sum model DELENAH_M (Natho and Thieken, 2018), an adaptation of the UNISDR method for national damage estimates (UNISDR 2015) in developed countries taking Germany as a study case, the estimate of the aggregated damage for the 345 buildings with claim data is 4.3 M€. This estimation is affected by an error which is comparable or lower than errors supplied by the micro-scale models (see Table 7), although being obtained with a simple calculation and in a blind mode, i.e. using the average damage ratio for severe floods and the average housing size derived from German survey data (Thieken et al., 2017) on flood losses in the housing sector (note that in this case underestimation of total damage is due ~~by~~ to the adoption of a conservative housing size, so that the estimation must be intended as a minimum estimate or a lower bound). Is this assessment useful for flood risk mitigation? Which is then the advantage of using micro-scale models? Is there a level of spatial aggregation which supply reliable, more informative estimation than a simple lump-sum at the municipality level? Answers to these questions will be objective of further investigations by the research groups involved in the test.

5 Conclusions: lessons learnt from a modeller's perspective

The blind test conducted in this study represented an opportunity not only to deeply investigate the transferability of tested models and the reliability of their estimations, especially regarding the potentialities of local and multi-variable models, but also to increase authors' awareness on strengths and limits of flood damage modelling tools. As concluding remarks, we report

in the following section take-home messages synthesising lessons learnt from the blind test, from a modeller's perspective.

First, ~~results from the blind test pointed out that~~ a former source of variability among models' outcomes lies in the approach for exposure assessment, which then represents a critical, often overlooked, step in flood damage modelling. In particular, assessing exposure coherently with the approach originally adopted in model development is key to preserve the original reliability ~~of damage estimates~~; in this regard, the blind test showed that the different approaches applied within the models demand for a clear definition and differentiation of the terms "exposure value" and "building value". Nonetheless, the blind test indicated a common overestimation, confirmed also in other case studies (Zischg et al., 2018; Cammerer et al., 2013; Thieken et al., 2008; Fuchs et al., 2019b; Arrighi et al., 2018a, 2018b), in terms of number of buildings damaged by a flood event (i.e. the number of buildings with claims is significantly lower than those exposed to the flood). This might be attributed to the fact that not all affected building owners asked for compensation, or that some buildings are not affected by the flood due to local micro-topographical conditions or due to the installation of ~~object~~ protection measures. ~~But~~However, it might also highlight problems in the current strategy adopted to identify ~~the~~ exposure (e.g. by not considering building elevation).

A second critical issue in flood damage modelling is the transfer of models in space and time~~, as also well known and documented in the literature~~, with difficulties on predicting the expected performance of a given model when applied in a different context ~~-~~(e.g. Jongman et al. 2012, Cammerer et al., 2013; Wagenaar et al., 2018). Accordingly, flood damage modellers should always be cautious when applying a flood damage model to a new context. Their general trust towards the model performance in the new study area must be in the first instance limited; however, model validation (ideally, ~~which more than a single~~over multiple datasets) can significantly increase the trust level.

But validation of damage models invariably relies on observed damage data, either from insurance claims, governmental reimbursement claims or~~,~~ direct surveys~~, etc.~~, all of which are generally intended as "reality". Indeed, 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, as for instance for insurance data (André et al. 2013; Spekkers et al. 2013; Zhou et al. 2013; Wing et al. 2020). ~~However~~In this context, the blind test highlighted that "reality" depicted by observations is not univocal, so that observed data must be carefully investigated before their comparison with model outcomes, as they may be addressing different types of damage, damage to different components, or being incomplete. Based on this consideration, ~~there is a need to~~flood damage modellers must be always cautious when drawing conclusions from validation analyses: if a model does not fit well to some empirical data, this does not necessarily ~~mean~~ reflect the inability of~~that it is not a "good" model~~the model in genereal terms but attention has to be drawn to input data quality and vice versa. This also points out the importance of collecting not only flood damage data, but also ancillary information on flood hazard and vulnerability of affected assets~~,~~ in the ex-post flood phase ~~arises~~ (Merz et al., 2004; Thieken et al., 2005; Ballio et al., 2015; Thieken et al., 2016; Molinari et al., 2017a; Molinari et al., 2019). Moreover, c~~Co~~onsultations of experts with local knowledge can help in ~~ensure~~ the correct interpretation and use of observed damage data~~. From another perspective, the importance of collecting not only flood damage data, but also ancillary information on flood hazard and vulnerability of affected assets in order to validate flood damage models arises (Merz et al., 2004; Thieken et al., 2005; Ballio et al., 2015; Thieken et al., 2016; Molinari et al.,~~

~~2017a; Molinari et al., 2019)~~.

In absence of data (or appropriate data) for validation, the application of several models might ~~help~~ be useful to quantify mean and variance and provide a range of uncertainty of the estimations (Figueiredo et al., 2018); a good agreement of model results, in particular with the models developed for context similar to the one under investigation, can significantly increase the trust level in model performance. In this regard, the blind test stressed that damage models have to be compared in their original form, meaning that, for instance, relative damage models relying on the total building value cannot be directly compared to the ones relying on only the first floor.

-As a general recommendation, to select a damage model for an application in 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. Moreover,

~~W~~when transferring a model (in space or time), proxies of input variables are frequently needed, and the modeller must be prudent in this step. A good understanding of both the data used during the model development and the data gathered for the new application is crucial, as the attribution of uncertainty becomes elusive afterwards, if this step is neglected. The blind test highlighted that the real effort of transferring the models to the given implementation context was related to finding the "right" required data, while the costs of implementing assumptions about exposure and calculating the damage value were negligible. To support transferability, there is then a need to precisely describe how the models were developed, which variables were included and for which specific context. In this regard, a protocol or standardised information for all models would help in finding the most appropriate ~~model~~ tool in a given context; in fact, at present, details about origin, calibration, assumptions, field of application, etc. of existing models in the literature are few and sparse. A new promising attempt in this direction is represented by the Flood Damage Model Repository, recently launched by Politecnico di Milano (www.fdm.polimi.it) as a research community effort.

Given these considerations, and in contrast with the general approach in which each research group develops ~~their~~ its own model~~s~~ for a limited context, authors support a call for a community effort in setting up a common model, with different sub-modules useable for many purposes and regions, and with a flexibility in the required input data.

Author contributions. *Conceptualisation of the blind test:* Francesco Ballio; *Management of the blind test:* Daniela Molinari; *Data and results management:* Daniela Molinari, Alice Gallazzi, Marta Galliani; *Models implementation*: Chiara Arrighi, Francesca Carisi, Marta Galliani, Patric Kellermann, Markus Mosimann, Stephanie Natho, Claire Richert; *Elaboration of results:* Daniela Molinari, Anna Rita Scorzini, Alice Gallazzi; *Interpretation of Results – Original Draft:* Daniela Molinari, Anna Rita Scorzini, Francesco Ballio; *Interpretation of Results – Review:* all; *Writing-Original Draft:* Daniela Molinari, Markus Mosimann, Francesca Carisi, Alessio Domeneghetti, Guilherme S. Mohor; *Writing-Review:* Daniela Molinari, all; *Figures and Tables:* Anna Rita Scorzini, Daniela Molinari.

Acknowledgements. Authors acknowledge with gratitude Andrea Nardini (from the Italian Centre for River Restoration – CIRF) and Marianne ~~Skow~~ Skov (from Rambøll, Denmark) for their fruitful suggestions and hints during the developing of the test. Authors are also grateful to three anonymous Reviewers for their meaningful comments and suggestions.

References

Amadio, M., Scorzini, A.R., Carisi, F., Essenfelder, A.H., Domeneghetti, A., Mysiak, J., and Castellarin, A.: Testing empirical and synthetic flood damage models: the case of Italy, Nat. Hazards Earth Syst. Sci., 19 (3), 661–678, doi:10.5194/nhess-19-661-2019, 2019.

André, C., Monfort, D., Bouzit, M., and Vinchon, C.: 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, doi: 10.5194/nhess-13-2003-2013, 2013.

Andreani, M., Gaikwad, A.J., Ganju, S., Gera, B., Grigoryev, S., Herranz, L.E., Huhtanen, R., Kale, V., Kanaev, A., Kapulla, R., Kelm, S., Kim, J., Nishimurai, T., Paladino, D., Paranjape, S., Schramm, B., Sharabi, M., Shen, F., Wei, B., Yan, D., and Zhang, R.: Synthesis of a CFD benchmark exercise based on a test in the PANDA facility addressing the stratification erosion by a vertical jet in presence of a flow obstruction, Nuclear Engineering and Design, 354, 110177, doi:10.1016/j.nucengdes.2019.110177, 2019.

Arrighi, C., Brugioni, M., Castelli, F., Franceschini, S., and Mazzanti, B.: Flood risk assessment in art cities: the exemplary case of Florence (Italy), Journal of Flood Risk Management, 11, 616-631, doi: 10.1111/jfr3.12226, 2018a.

Arrighi, C., Rossi, L., Trasforini, E., Rudari, R., Ferraris, L., Brugioni, M., Franceschini, S., and Castelli, F.: Quantification of flood risk mitigation benefits: A building-scale damage assessment through the RASOR platform, Journal of Environmental Management, 207, 92-104, doi:10.1016/j.jenvman.2017.11.017, 2018b.

Ballio, F., Molinari, D., Minucci, G., Mazuran, M., Arias Munoz, C., Menoni, S., Atun, F., Ardagna, D., Berni, N., and Pandolfo, C.: The RISPOSTA procedure for the collection, storage and analysis of high quality, consistent and reliable damage data in the aftermath of floods, Journal of Flood Risk Management, 11, S604–S615 , 2018. https://doi.org/10.1111/jfr3.12216

Breiman, L., Friedman, J., Olshen, R.A., and Stone, C.J.: CART: Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Bundesministerium für Verkehr und digitale Infrastruktur (2016). "Hochwasserkatastrophe 2013 - Bericht über die Verwendung der Finanzhilfe aus dem EU-Solidaritätsfonds zur Bewältigung der durch das Hochwasser 2013 in der Bundesrepublik Deutschland entstandenen Schäden der öffentlichen Hand". (Berlin: Bundesministerium für Verkehr und

digitale Infrastruktur, Projektgruppe Hochwasser). URL: https://www.bmvi.de/SharedDocs/DE/Anlage/WS/hochwasserkatastrophe-2013-bericht.pdf?__blob=publicationFile; last download January 13th 2020.

Cammerer, H., Thieken, A.H., and Lammel, J.: Adaptability and transferability of flood loss functions in residential areas, Nat. Hazards Earth Syst. Sci., 13(11), 3063–3081, doi:10.5194/nhess-13-3063-2013, 2013.

Carisi, F., Schröter, K., Domeneghetti, A., Kreibich, H., and Castellarin, A.: Development and assessment of uni- and multi-variable flood loss models for Emilia-Romagna (Italy), Nat. Hazards Earth Syst. Sci., 18, 2057–2079, doi:10.5194/nhess-18-2057-2018, 2018.

Cellerino, R.: L'Italia delle alluvioni. Un'analisi economica, Franco Angeli Editore, 2004

CEPRI. (2014a). Evaluation des dommages liés aux inondations sur les logements.

CEPRI. (2014b). Evaluation des dommages aux logements liés aux submersions marines.

Deutscher Bundestag (2013). "Bericht zur Flutkatastrophe 2013: Katastrophenhilfe, Entschädigung, Wiederaufbau". (Berlin). URL: dip21.bundestag.de/dip21/btd/17/147/1714743.pdf; last download January 13th 2020.

Dottori, F., Figueiredo, R., Martina, M.L.V., Molinari, D., and Scorzini, A.R.: INSYDE: a synthetic, probabilistic flood damage model based on explicit cost analysis, Nat. Hazards Earth Syst. Sci., 16, 2577-2591, doi:10.5194/nhess-16-2577-2016, 2016.

Dutta, D., Herath, S., and Musiake, K.: A mathematical model for flood loss estimation, Journal of Hydrology, 277 (1-2), 24–49, doi:10.1016/S0022-1694(03)00084-2, 2003.

Figueiredo, R., Schröter, K., Weiss-Motz, A., Martina, M.L.V., and Kreibich, H.: Multi-model ensembles for assessment of flood losses and associated uncertainty, Nat. Hazards Earth Syst. Sci., 18(5), 1297–1314, doi:10.5194/nhess-18-1297-2018, 2018.

Fuchs, S., Keiler, M., Ortlepp, R., Schinke, R., and Papathoma-Köhle, M.: Recent advances in vulnerability assessment for the built environment exposed to torrential hazards: Challenges and the way forward, Journal of Hydrology, 575, 587–595, doi:10.1016/j.jhydrol.2019.05.067, 2019a.

Fuchs, S., Heiser, M., Schlögl, M., Zischg, A., Papathoma-Köhle, M., and Keiler, M.: Short communication: A model to predict flood loss in mountain areas, Environmental Modelling & Software, 117, 176–180, doi:10.1016/j.envsoft.2019.03.026, 2019b.

Gerl, T., Kreibich, H., Franco, G., Marechal, D., and Schroter, K.: A Review of Flood Loss Models as Basis for Harmonization and Benchmarking, PLoS ONE, 11 (7), e0159791, doi:10.1371/journal.pone.0159791, 2016.

Huizinga, J., de Moel, H., and Szewczyk, W.: Flood damage functions for EU member states. Technical report, HVK Consultants. Implemented in the framework of the contract #382441-F1SC awarded by the European Commission – Joint Research Centre, 2007.

Jongman, B., Kreibich, H., Apel, H., Barredo, J.I., Bates, P.D., Feyen, L., Gericke, A., Neal, J., Aerts, J.C.J.H., and Ward, P.J.: Comparative flood damage model assessment: towards a European approach, Nat. Hazards Earth Syst. Sci., 12, 3733–3752, doi:10.5194/nhess-12-3733-2012, 2012.

Jonkman, S.N., Bočkarjova, M., Kok, M., and Bernardini, P.: Integrated hydrodynamic and economic modelling of flood damage in the Netherlands, Ecological Economics, 66 (1), 77–90, doi:10.1016/j.ecolecon.2007.12.022, 2008.

Krogstad, P.Å., & Eriksen, P.E.: "Blind test" calculations of the performance and wake development for a model wind turbine, Renewable Energy, 50, 325-333, doi:10.1016/j.renene.2012.06.044, 2013.

Merz, B., Kreibich, H., Thieken, A., and Schmidtke, R.: Estimation uncertainty of direct monetary flood damage to buildings, Nat. Hazards Earth Syst. Sci., 4(1), 153-163, SRef-ID: 1684-9981/nhess/2004-4-153, 2004.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Review article "Assessment of economic flood damage", Nat. Hazards Earth Syst. Sci., 10, 1697–1724, doi:10.5194/nhess-10-1697-2010, 2010.

Merz, B., Kreibich, H., and Lall, U.: Multi-variate flood damage assessment: a tree-based data-mining approach, Nat. Hazards Earth Syst. Sci., 13, 53–64, doi:10.5194/nhess-13-53-2013, 2013.

Meyer, V., Becker, N., Markantonis, V., Schwarze, R., van den Bergh, J.C.J.M., Bouwer, L.M., Bubeck, P., Ciavola, P., Genovese, E., Green, C., Hallegatte, S., Kreibich, H., Lequeux, Q., Logar, I., Papyrakis, E., Pfurtscheller, C., Poussin, J., Przyluski, V., Thieken, A.H., and Viavattene, C.: Review article: Assessing the costs of natural hazards – state of the art and knowledge gaps, Nat. Hazards Earth Syst. Sci., 13, 1351–1373, https://doi.org/10.5194/nhess-13-1351-2013, 2013.

Molinari, D., Menoni, S., and Ballio, F. (Eds): Flood Damage Survey and Assessment: New Insights from Research and Practice, AGU-Wiley, 2017a.

Molinari, D. and Scorzini A.R.: On the Influence of Input Data Quality to Flood Damage Estimation: The Performance of the INSYDE Model, Water, 9(9), 688, doi:10.3390/w9090688, 2017b.

Molinari, D., de Bruijn, K.M., Castillo-Rodríguez, J.T., Aronica, G.T., and Bouwer, L.M.: Validation of flood risk models: Current practice and possible improvements, International Journal of Disaster Risk Reduction, 33, 441–448, doi:10.1016/j.ijdrr.2018.10.022, 2019.

Natho, S. and Thieken, A.H.: Implementation and adaptation of a macro-scale method to assess and monitor direct economic losses caused by natural hazards, International Journal of Disaster Risk Reduction, 28, 191-205, https://doi.org/10.1016/j.ijdrr.2018.03.008, 2018.

Orlandini, S., Moretti, G., and Albertson, J.D.: Evidence of an emerging levee failure mechanism causing disastrous floods in Italy, Water Resour. Res., 51, 7995–8011, https://doi.org/10.1002/2015WR017426, 2015.

Penning-Rowsell, E., Johnson, C., Tunstall, S., Tapsell, S., Morris, J., Chatterton, J., and Green, C.: The benefits of flood and coastal risk management: a handbook of assessment techniques, Middlesex University Press, UK, 2005.

Ransley, E., Yan, S., Brown, S.A., Mai, T., Graham, D., Ma, Q., Musiedlak, P-H., Engsig-Karup, A.P., Eskilsson, C., Li, Q., Wang, J., Xie, Z., Venkatachalam, S., Stoesser, T., Zhuang, Y., Li, Q., Wan, D., Chen, G., Chen, H., Qian, L., Ma, Z., Mingham, C., Causon, D., Gatin, I., Jasak, H., Vukcevic, V., Downie, S., Higuera, P., Buldakov, E., Stagonas, D., Chen, Q., Zang, J., and Greaves, D.: A Blind Comparative Study of Focused Wave Interactions with a Fixed FPSO-like Structure (CCPWSI Blind Test Series 1), International Journal of Offshore and Polar Engineering, 29(2), 113-127, doi:10.17736/ijope.2019.jc748, 2019.

Richert, C. and Grelot, F.: Comparaison des modèles de dommages nationaux avec les données de sinistralité, Tech. rep., IRSTEA, Montpellier, France, 2018.

Röthlisberger, V., Zischg, A.P., and Keiler, M.: A comparison of building value models for flood risk analysis, Nat. Hazards Earth Syst. Sci., 18, 2431–2453, doi:10.5194/nhess-18-2431-2018, 2018.

Rouchon, D., Christin, N., Peinturier, C., and Nicklaus, D. (2018): Analyse multicritère des projets de prévention des inondations. Guide méthodologique 2018. Théma - Balises. Ministère de la Transition Écologique et Solidaire, Commissariat général au développement durable. Available online at https://www.ecologique-solidaire.gouv.fr/sites/default/files/Th%C3%A9ma%20-%20Analyse%20multicrit%C3%A8re%20des%20projets%20de%20pr%C3%A9vention%20des%20inondations%20-%20Guide.pdf.

Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F., and Merz, B.: How useful are complex flood damage models?, Water Resources Research, 50(4), 3378-3395, https://doi.org/10.1002/2013WR014396, 2014.

Scorzini, A.R. and Frank, E.: Flood damage curves: new insights from the 2010 flood in Veneto, Italy, Journal of Flood Risk Management, 10 (3), 381-392, doi:10.1111/jfr3.12163, 2017.

Scorzini, A.R., Radice, A., and Molinari, D.: A New Tool to Estimate Inundation Depths by Spatial Interpolation (RAPIDE): Design, Application and Impact on Quantitative Assessment of Flood Damage, Water 2018, 10, 1805, doi:10.3390/w10121805, 2018.

Skorek, T., de Crécy, A., Kovtonyuk, A., Petruzzi, A., Mendizábal, R., de Alfonso, E., Reventós, F., Freixa, J., Sarrette, C., Kyncl, M., Pernica, R., Baccou, J., Fouet, F., Probst, P., Chung, B., Tram, T.T., Oh, D., Gusev, A., Falkov, A., Shvestov, Y., Li, D., Liu, X., Zhang, J., Alku, T., Kurki, J., Jäger, W., Sánchez, V., Wicaksono, D., Zerkak, O., and Pautz, A.: Quantification of the uncertainty of the physical models in the system thermal-hydraulic codes–PREMIUM benchmark, Nuclear Engineering

and Design, 354, 110199, doi:10.1016/j.nucengdes.2019.110199, 2019.

Smith, M.B., Seo, D.J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, Journal of Hydrology, 298(1-4), 4-26, doi:10.1016/j.jhydrol.2004.03.040, 2004.

Soares-Frazao, S., Canelas, R., Cao, Z., Cea, L., Chaudhry, H.M., Die Moran, A., El Kadi, K., Ferreira, R., Cadórniga, I.F., Gonzalez-Ramirez, N., Greco, M., Huang, W., Imran, J., Le Coz, J., Marsooli, R., Paquier, A., Pender, G., Pontillo, M., Puertas, J., Spinewine, B., Swartenbroekx, C., Tsubaki, R., Villaret, C., Wu, W., Yue, Z., and Zech, Y.: Dam-break flows over mobile beds: experiments and benchmark tests for numerical models, Journal of Hydraulic Research, 50(4), 364–375, doi:10.1080/00221686.2012.689682, 2012.

Spekkers, M.H., Kok, M., Clemens, F.H.L.R., and Ten Veldhuis, J.A.E.: A statistical analysis of insurance damage claims related to rainfall extremes. Hydrol. Earth Syst. Sci., 17(3), 913-922, doi: 10.5194/hess-17-913-2013, 2013.

Teng, J., Jakeman, A.J., Vaze, J., Croke, B.F.W., Dutta, D., and Kim, S.: Flood inundation modelling: A review of methods, recent advances and uncertainty analysis, Environmental Modelling & Software, 90, 201–216, https://doi.org/10.1016/j.envsoft.2017.01.006, 2017.

Thieken, A.H., Müller, M., Kreibich, H., and Merz, B.: Flood damage and influencing factors: New insights from the August 2002 flood in Germany, Water Resources Research, 41(12), W12430, doi:101029/2005WR004177, 2005.

Thieken, A.H., Olschewski, A., Kreibich, H., Kobsch, S., and Merz, B.: Development and evaluation of FLEMOps - a new Flood Loss Estimation MOdel for the private sector, in: Flood Recovery, Innovation and Response I, edited by: Proverbs, D., Brebbia, C.A., and Penning-Rowsell, E. (Eds.), WIT Press, pp. 315-324, doi:10.2495/FRIAR080301, 2008.

Thieken, A.H., Bessel, T., Kienzler, S., Kreibich, H., Müller, M., Pisi, S., and Schröter, K.: The flood of June 2013 in Germany: how much do we know about its impacts?, Nat. Hazards Earth Syst. Sci., 16, 1519-1540, doi:10.5194/nhess-16-1519-2016, 2016.

Thieken, A.H., Kreibich, H., Müller, M., and Lamond, J.: Data collection for a better understanding of what causes flood damage – experiences with telephone surveys, in: Flood Damage Survey and Assessment: New Insights from Research and Practice, edited by: Molinari, D., Menoni, S., and Ballio, F., AGU Wiley, pp. 95-106, doi:10.1002/9781119217930.ch7, 2017.

UNISDR 2015, Concept note on Methodology to Estimate Direct Economic Losses from Hazardous Events to Measure the Achievement of Target C of the Sendai Framework for Disaster Risk Reduction: A Technical Review, Report, 51 Pages. Available online: https://www.preventionweb.net/documents/framework/Concept%20Paper%20-%20Direct%20Economic%20Loss%20Indicator%20methodology%2011%20November%202015.pdf

Van Ootegem, L., van Herck, K., Creten, T., Verhofstadt, E., Foresti, L., Goudenhoofdt, E., Reyniers, M., Delobbe, L., Murla

Tuyls, D., and Willems, P.: Exploring the potential of multivariate depth-damage and rainfall-damage models, J Flood Risk Management, 11, S916-S929, doi:10.1111/jfr3.12284, 2018.

Wagenaar, D., de Jong, J., and Bouwer, L.M.: Multi-variable flood damage modelling with limited data using supervised learning approaches, Nat. Hazards Earth Syst. Sci., 17(9), 1683-1696, https://doi.org/10.5194/nhess-17-1683-2017, 2017.

Wagenaar, D., Lüdtke, S., Schröter, K., Bouwer, L.M., and Kreibich, H.: Regional and Temporal Transferability of Multivariable Flood Damage Models, Water Resources Research, 54 (5), 3688–3703, doi:10.1029/2017WR022233, 2018.

Wing, O.E., Pinter, N., Bates, P.D., and Kousky, C.: New insights into US flood vulnerability revealed from flood insurance big data. Nat. Commun., 11(1), 1-10, doi: 10.1038/s41467-020-15264-2, 2020.

Zelt, C.A., Haines, S., Powers, M.H., Sheehan, J., Rohdewald, S., Link, C., Hayashi, K., Zhao, D., Zhou, H., Burton, B.L., Petersen, U.K., Bonal, N.D., and Doll, W.E.: Blind test of methods for obtaining 2-D near-surface seismic velocity models from first-arrival traveltimes, Journal of Environmental and Engineering Geophysics, 18(3), 183-194, http://dx.doi.org/10.2113/JEEG18.3.183, 2013.

Zhou, Q., Panduro, T.E., Thorsen, B.J., and Arnbjerg-Nielsen, K.: Verification of flood damage modelling using insurance data. Water Sci. Technol., 68(2), 425-432, doi: 10.2166/wst.2013.268, 2013.

Zischg, A.P., Mosimann, M., Bernet, D.B., and Röthlisberger, V.: Validation of 2D flood models with insurance claims, Journal of Hydrology, 557, 350–361, doi:10.1016/j.jhydrol.2017.12.042, 2018.