Dear editor and reviewers,

We are grateful for all your comments. In this response document, we replied to the comments point by point. The detailed revisions are shown in our manuscript.

Replies to Referee #1

General comments:

Zhou et al. present a novel and interesting analysis of uncertainties inherent to global hydrodynamic models. Many studies over the past ~5 years have posited that global flood models produce widely divergent results, yet each of them fails to explain or evidence the cause of this divergence. The authors begin to address some of these previous gaps in the literature and have produced some (though limited) conclusions of genuine interest. However, there are fundamental problems with the methods, analysis, and wider interpretation/discussion of their conclusions that require addressing before publication is to be considered in NHESS. Do not be put off by the volume of comments. I do think the work has great value and is something the field desperately requires. To be the truly impactful contribution this work needs to be, I offer the following (I hope, constructive) comments. These broadly relate to:
- Model sensitivity vs. uncertainty analysis

Re: We agree that the contribution of this paper should be the model sensitivity as the reviewers suggested. So, we reframed the structures and revised the manuscript in the following aspects.

1. Added discussions in the Introduction how the uncertainty/sensitivity is discussed and addressed in the previous literature.
2. Emphasized in the introduction that we focused on uncertainty/sensitivity analysis rather than validation of the flood estimation.
3. The comparisons with two other flood hazard maps are removed from the results since they are not "validation" and they help little to uncertainty/sensitivity analysis.
4. We added uncertainties/sensitivity analysis on the population exposure and economic exposure for different floods.

- A single model with often extremely limited geographic scope

Re: Agree. The flood method is one of the uncertainties can lead to deviations in the estimation. However, we are not able to run different routing models with different runoff inputs. This may need more collaborations. Although we only use one single model, we can analyze model sensitivity to the inputs (i.e., runoff) and other conditions (e.g., fitting distribution, used variables for fitting). If new river models are added, we would further investigate the flood extent sensitivity to model choice. The limitation is discussed in the Discussions.

Trigg et al. (2016), Bernhofen et al. (2018) and Aerts et al. (2020) used 8 different models with different forcing, hydrological model, routing models. Although their results showed

deviations, they are not able to attribute the contribution a single process (e.g., routing) to the final uncertainty. (This has been added in the Introduction)

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G., and Fewtrell, T. J.: The credibility challenge for global fluvial flood risk analysis, Environmental Research Letters, 11, https://doi.org/10.1088/1748- 9326/11/9/094014, 2016.

Bernhofen, M. V., Whyman, C., Trigg, M. A., Sleigh, P. A., Smith, A. M., Sampson, C. C., Yamazaki, D., Ward, P. J., Rudari, R., Pappenberger, F., Dottori, F., Salamon, P., and Winsemius, H. C.: A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique, Environmental Research Letters, 13, https://doi.org/10.1088/1748-9326/aae014, 2018.

Aerts, J. P. M., Uhlemann-Elmer, S., Eilander, D., and Ward, P. J.: Comparison of estimates of global flood models for flood hazard and exposed gross domestic product: a China case study, Natural Hazards and Earth System Sciences, 20, 3245–3260, https://doi.org/10.5194/nhess-20-3245-2020, 2020.

- Considered uncertainties in the context of ’unconsidered’ uncertainties - The relevance of hydrologic variable choice

Re: Yes, we are using different variables in the fitting process (i.e., water level, water storage) because it is not linear between water level and water storage. They must have different distributions. Although in this paper, only the water level and water storage in the model unit catchment are used, it is still an attempt to investigate a new uncertainty source for the final flood extent mapping. Discharge is frequently used in engineering projects, while it is not used in our analysis because discharge and water level is not one-to-one consistent due to the loop rating curve. While with either of river water depth or the water storage we can estimated the flood extent and the floodplain water depth for any target region using CaMa-Flood. (This has been reflected in section 2.1)

- The contextual relevance of distribution goodness-of-fit

Re: In the new manuscript, we conclude that the goodness-of-fit is likely to be poor in dry climate or mountainous areas. These regions are where the accumulative river discharge is small. The magnitude of water level (or river storage) is highly depended on single precipitation events, leading to an unstable relation between the high floods in different years. (This has been reflected in section 3.1)

- The need for some benchmark data

Re: As the reviewer mentioned, it is not easy to find appropriate benchmark data at the global scale due to lack of enough validation data especially in the developing countries. The two datasets (JRC and GAR) used in this study are also not appropriate because they are still replying on models. However, since we decided to focus on the sensitivity analysis at the

global scale, we first removed the current section 4.5 (comparison over the lower Mekong region) and we will not use any benchmark data (including any statistics or remote sensing images).

****

Specific comments:

The introduction is mostly good, but could contain a richer discussion on why unstitching the uncertainties of global flood models is needed and how past studies essentially have failed to do this. There are also many sweeping or incorrect generalisations, that may simply be a function of imprecise English (for which I am sympathetic and under- standing). This is the case throughout the manuscript (e.g. the sentence in line 26-27 p. 2 makes little sense as the same could be said for RFFA; line 3 p. 23 says the studies assess flood risk, when it does not).

Re: We have rewritten the Introduction. Additional literature and new discussions on the sensitivity analysis are provided.

To avoid English errors, the whole text has been checked with professional English editing service (https://www.proof-reading-service.com). A certificate is provided as a supporting material.

The overarching problem in the introduction is that it makes the reader think some quantification of the uncertainties – validation, against observations – is carried out by the authors, and this is not the case. Truly, this analysis is a sensitivity analysis of 1 model. While I appreciate this illustrates the 'uncertainty one should have about conclusions drawn using this model', really it just tells us what the model is sensitive to and by how much. The paper as a whole needs more of a framing as a sensitivity analysis rather than formal uncertainty quantification.

Re: Yes, we have restructured the whole paper and now this study is concentrated on the sensitivity analysis to various influential factors, rather than quantifying differences/biases compared to observations. The introduction has been rewritten as well.

The methodology, if framed as a sensitivity analysis of CaMa-Flood, appears thorough and fit-for-purpose. In general, justification of the use of a single model and subsequent analysis in specific regions (even specific grid cells) is needed. How universal are the conclusions in light of these methodological choices? Of course, these are only uncertainties related to the subjective choice of model tests. It is worth stressing that the reported uncertainties make the (of course, incorrect) assumption that terrain, channel bathymetry, human influence, and model parameterisation are certain.

Re: The compassions with other products have been removed.

There are indeed many uncertainty sources which are not studied in this paper. So, we assume others are certain and we had in our Discussions about how other uncertainties sources have a potential impact on our results and how important is the uncertainty in other studies..

[a.] It is not clear why river depth and water storage are chosen as variables of interest – this needs further explanation, as I can not yet see the significance of doing so. Common application of FFA is to discharge, yet this is not done here. [b.]Further discussion of the AIC is needed: what constitutes a 'good' result in this context is not specified. Equally, what is the relevance of this metric in terms of model uncertainty? What is the relevance of a good fitting distribution in the context of the uncertainty in the absolute values themselves? [c.]Are the authors saying that a variable with a poor AIC contains no relevant information for FFA? Really, it just shows a suitable distribution has not yet been found. I think section 3.1 fails to recognise the variable of choice is arbitrary and depends on the model used and the question asked. We all know that a 100-year rain-fall≠ 100-year streamflow≠ 100-year economic loss. So frame this strand of analysis in the context of why the variable you choose matters and why this is interesting.

Re: [a.] Water storage in the unit-catchment is the prognostic variable in CaMa-Flood. Water level is the diagnostic variable estimated from water storage. With either of these two variables we can estimated the flood extent and the floodplain water depth for any target region. Discharge is the variable frequently used in engineering design. However, with only discharge we cannot estimate the water level since the relation between discharge and water level is not one-to-one consistent due to the loop rating curve. (This has been reflected in 2.1 Experiment design).

[b.] Instead of defining the best/good fitting result with aic, we decided to compare the sensitivity of the aic to different experiment. The aic itself is not able to show the model uncertainty, but when we compare the aic for all experiments, we can conclude that aic is sensitive to the selection of the fitting functions, and then the variables. It is less sensitive to the runoff inputs. (additional explanation of the aic is added in section 2.3, the results are reflected in section 3.1)

[c.] No, as the reviewer mentioned, we haven't found a best fitting distribution for any variable that is applicable for all the places. But we can still use the uncertainty information from experiments with different variables. Regarding other variables, because we are working on floods, 100-year rainfall is not applicable. 100-year streamflow is discussed in the reply to this question a. 100-year economic loss could be more complex. But we added one subsection in the result to discuss what are the population and economic exposure to the floods (see section 3.3.3).

I can not see evidence that WAK is the best distribution because of it having 5 parameters. As the authors mentioned, it may just be overfit to the simulation record. The reality is we have no idea which distribution we should extrapolate with – and this is not something the AIC can test.

Re: Yes, we agree. Therefore, in the new manuscript, the aic is used for testing the sensitivity of fitting performance by comparing aic for various experiments.

The section 3.3 analysis of runoff is interesting, but the results are stated in such a way that the authors expected the analysis to produce a 'preferable' runoff product. No feature of the analysis performed could identify such a thing. It is not clear why being a runoff product in 'the middle' is the best place to be: it could be that the lowest estimate types are actually best! It is a problem throughout the paper, where a suitable perfor- mance benchmark has not been found. Ensure the results are framed and reported as sensitivities, not as good/bad.

Re: This part has been removed in the new manuscript. Regarding the "middle" one, we have one pre-assumption that the users don't know which runoff should be the best. In this case, the users tend to use the ensemble mean rather than a single runoff input. If the system is too heavy to run for all runoffs, it is better to choose the one in the middle. (But anyway, this has been removed from the previous manuscript.)

I like the analysis in section 4.1, but I'm not sure why this could not be done for every global grid cell – with normalised results – and presented in the same way. How representative is this grid cell? It may also be interesting here to compare the AIC results to Figure 7c: exploring some of my above comments on why AIC matters more quantitatively (i.e., does high/low AIC [thus, how good the distribution fit is] matter in the context of inundation?)

Re: Probably, it is not necessary to work on the normalized results since they will neglect the variations caused by the mean values. Conducting the point analysis for all the global grids might not be realistic with regard to the necessity or the difficulty for visualization. Instead, we selected five more river basins (e.g., Amazon, Yangtze, Mississippi, Lena and Nile), for each river basin we investigated the results at one specific GRDC gauge near the river outlet.

The differences of the fitting performance are mainly due to the degree of freedom of each fitting distribution as the WAK has five parameters, GAM and GUM have two parameters while the others have three. With higher degree of freedom, the fitting performance will be better. Regarding the spatial pattern, we found that the aic is higher (indicating poorer fitting performance) in dry zones and mountainous regions. The accumulative river discharge over those regions is small. The magnitude is thus highly depended on single precipitation events, leading to an unstable relation between the high floods in different years. (This has been reflected in section 3.1).

As for the rest of section 4, the analysis is good. While I appreciate visualising the globe at this scale is difficult, a lot of the calculations could still easily be done globally. It leaves the reader wondering whether different climatologies and geomorphic settings might have different conclusions. Deltas are difficult to model – particularly for models with poor/no representation of coastal boundaries – and so may have distinct features of uncertainty to other areas. I see no reason for the authors not to report findings elsewhere.

Re: Thank you very much for this suggestion. In the new manuscript, the analysis over the global maps has been added as current section 3.2. For short, the regions with high coefficient of variation (Cv, ratio of standard deviation to the mean) are likely to be the dry zones (e.g., Sahara, Center Australia, Center Asia) and the originating river basins in mountainous regions (e.g., the Rocky Mountain, the Andes, Tibet Plateau). We didn't find large variations over the deltas, this might because we haven't introduced other flood models. With the same flood model, the deviation among different inputs/fitting functions can be small over the deltas.

Visualizing the global results is difficult, we therefore provide detailed maps for specific river basins. The results for the lower Mekong River Basin are presented in the main text, while results for five other river basins (i.e., Amazon, Yangtze, Mississippi, Lena, Nile) are provided in the supporting material.

I do not see any value to section 4.5. I have little doubt the CaMa-Flood 100-year map is more accurate than the GAR and JRC maps: it is an uninformative comparison, and certainly

not "validation". You only have to look at the stripes of JRC's map in Figure 12b to know that that is not a model you should aspire to resemble! I appreciate finding suitable validation data is difficult, but it is difficult to understand the relevance of the authors' conclusions without some. Perhaps running this analysis in the US or western Europe where high-quality models exist and comparing to those would be a good idea.

Re: We have removed this section. And we didn't add analysis in the US or western Europe because observations are still not available. Excluding this part will not affect the current scope of this study.

Section 5 is strong, but will benefit from drawing on some of the above points. Generally, the manuscript is quite long, and so the impact from section 5 is dampened by unnecessarily long analysis in 4.2-4.4. Throughout the paper, I would ensure each test is a worthwhile inclusion for the conclusions drawn. At present, there are many analyses which offer little additional information which I would consider cutting.

Re: Thanks. Yes, by removing some analysis (e.g., section 4.5) and shortening sections 4.2-4.4, we can explore more on the global maps where the uncertainties are higher and why this happens. The detailed analysis can be found as section 3.2.

Figures are generally good quality, but most need to be larger. I would change the colour scheme of some figures (e.g. 4-6) where colour scales are used for variables which are not ordinal (no reason to go from blue to red, when the distributions are in no order).

Re: The Figures are prepared with high quality, so it can be displayed in a large size. Regarding the color in Figure 4-6, it is actually near random. But I may change the colorbar which seems that the colors are sequent (but the mentioned figures are removed from the text.)

Replies to Referee #2

General comments

This paper is based on large scale hydrologic-hydrodynamic simulations to investigate different sources of uncertainty in flood risk estimation, with the use of flood frequency analysis tools. The chosen topic deserves some interest, though the analysis is based on a specific configuration of a set of available hydrological model output (from the Earth2Observe project) and an in-house hydrodynamic model (CaMa). However, the focus on the global domain makes it of larger interest for a q wider community.

- Among the main limitations of the manuscript is the sub-optimal use of the english language, including both terminology, grammar, typos and structure of the sentences, which makes it hard to read and at times hampers the understanding of the content. I strongly suggest to work and improve it with the help of a native speaker.

Re: Thanks for the suggestion. The submitted revision has been checked with professional English editing service (https://www.proof-reading-service.com). A certificate is provided as a supporting material.

- Another important comment is related to the general framing of the analysis. In the current version a number of analyses are performed, focusing on different aspects, though in my opinion it lacks a consistent storyline and some reasoning behind why they were made and clear statements about what we learn following their results.

Re: Yes, we realized that we have included too much analysis from different aspects. In the revised manuscript, we deleted some of them and concentrated on the sensitivity analysis to various model inputs, fitting functions and the variable selected. Analysis will be conducted from the global scale to basin scale with emphasizing point analysis. The uncertainties in the inundation area and the potential impact on population and economy is also discussed with the uncertainties.

- The manuscript is too long compared to the information content it brings. I suggest shortening following the comments below. A number of figures should be removed, improved or put in the supplement material, for the reasons I explain below in the specific comments. In particular, I'm speaking about Figures 4 and 5 wrt the issues with fitting analytical functions with different degrees of freedom (comment #10), Fig. 6 (comment #18), Fig. 10, 12, and 14 (comments #24, #27, #29)

Re: Thanks, we have shortened the manuscript by considering your comments and comments from reviewer #1. For example, we deleted Figure 4 to Figure 6 because they are not relevant to the sensitivity analysis. We deleted Figure 9 and 10 since they bring little information (as #24). Figure 12 and 13 are removed as we will not discuss the validation in this study. Figure 14 is reorganized so that it can be better discussed. New Analysis on the sensitivity over many other river basins are added in the support materials to support our analysis. Analyses on the population and economic exposure to the floods are added.

Specific comments

1. p2, l8-9: acronyms should be defined with “full name (acronym)”, e.g., Global Runoff Data Centre (GRDC). Same for p3, l5 and l26.

Re: Thanks, the same errors are fixed in the text.

2. p2, l14: Pearson type III

Re: Revised.

3. p3, l1: suggested “connected” –> “analyzed the relation between ...”

Re: Revised.

4. p3, l3-5: Sentence not clear. Please rephrase.

Re: The Introduction has been almost rewritten.

5. P4, l3: please define the acronym SAR

Re: Synthetic-aperture radar (SAR). Added in the manuscript.

6. p4, l10: “various runoff inputs” is too general. Please add details here or a reference to the details included in Sect. 2.2 wrt the inputs used.

Re: Yes, the various runoff inputs are listed in section 2.1 now. We added a reference to the previous section to explain the “various runoff inputs”.

7. P4, l13: I suggest adding an introductory sentence here to give more details about the experiment itself, before jumping to the uncertainties to investigate.

Re: We reorganized the method part. One new paragraph is added before introducing the experiments. (This is reflected in first paragraph in the section 2.1)

8. P4, l14-16: please improve this part. Also, I find the variable names V1_(rivdph) and V2_(sto2dph) not very intuitive. Why not simply calling them depth and storage? Especially sto2dph creates confusion on whether it is a storage or a depth.

Re: Thanks. As suggested, in the new manuscript, we use river water depth and water storage instead of “rivdph” or “sto2dph”. Explanations of the differences between using water level and water storage, as well as the reason for not using discharge are added in the manuscript (see the second paragraph in section 2.1). The explanations also can be found in replies to the fourth question of Reviewer #1.

9. Table 1: I suggest removing “Various” in the caption.

Re: Revised.

10. P5, l12: Note that the Gumbel and the Gamma distributions have 2 parameters. In fact, results in Figure 5 seems to me the natural consequence of fitting a series of points with mathematical functions with different degrees of freedom, where the 5 parameter distribution

is able to fit the data more skillfully (though it doesn't mean it will be more skillful in predictive mode for future floods), Then the 3 parameter distributions and the 2-parameter Gamma and Gumbel as the least skillful. One would obtain similar results when fitting the series of data with polynomials of grade 5,3 and 2, because higher grade polynomials can fit better the input data.

Re: Thanks. The degree of freedom is the cause for the diversities of final results using different fitting distributions. We have added this explanation to the revised manuscript in section 3.1.

11. P5, l13: I suggest renaming this section (e.g., "Fitting performance" or similar)

Re: Revised. The subtitles for other sections are also checked.

12. p5, l15: calculated

Re: revised.

13. p5, l19-20: This should be expressed more clearly. E.g:" Smaller aic denote higher fitting performance" or similar, which is actually better written in p6, l23-24

Re: we also added just after the equation that the smaller aic denotes higher fitting performance in section 2.3.

14. p6, l24-26: Use active rather than passive form (e.g., "we compare")

Re: We revised the above and sentences in similar locations. Before resubmitting the revision, the manuscript has been sent for professional English correction to minimize such problems.

15. p7, l6-7: Is the normalization the real reason? Also, I suggest giving more details on how to weigh the aic values. What is the optimum? What are normally considered good or bad values? It is not intuitive for those who have never used it.

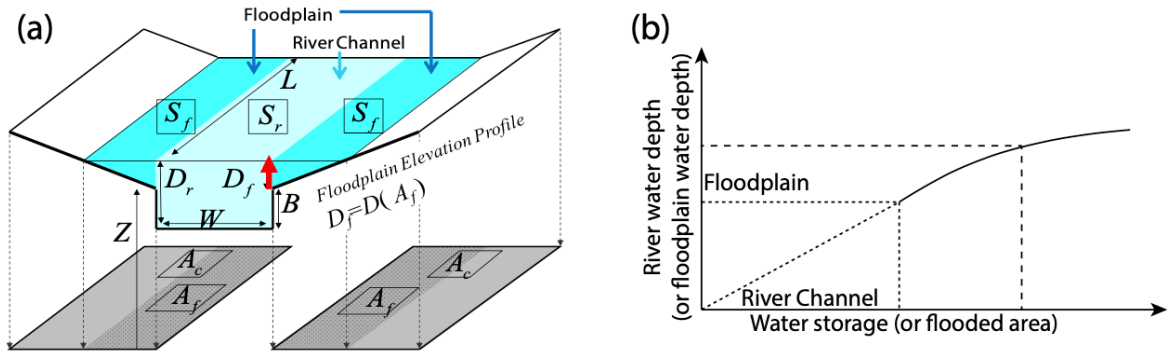
Re: Although, there are various performance metrics to measure the goodness-of-fit, the aic is better in our study because it will enlarge the small difference between samples and estimations. We only have 35 samples and these are sorted, therefore the fitting performance should be very high and fitting results will not have large differences. Then optimum of aic is negative infinity. Same as RMSE, the lower AIC, the performance is better.

16. P8, l8: "The later peak" – > "the latter"

Re: Revised.

17. Figure 3: Interesting to see how the pdfs of gamma and gumbel have similar peaks to the other distributions only for the storage, but not for the river depth. Indeed it is clearly fisible also in Fig. 3c. Would be interesting to investigate and motivate the reasons. Now it is only mentioned but no justification is given.

Re: Thanks. Investigating the difference in fitting the water level and the storage is interesting. As shown in the following illustration, in CaMa-Flood the water level is calculated by allocating the water storage to the river channel and floodplain from the bottom to the top. In the channel, the relation between water level and water storage is linear because we assume the river profile as rectangle, while it changes to nonlinear in the floodplains. So, if the maximum water level for the different years locates both in the river channel and the floodplain, fitting the water level becomes more difficult especially for GAM and GUM since they only have 2-parameters. While, because the storage is not affected by the channel shape, fitting the water storage with different fitting functions will not make large differences. (This has been reflected in the section 3.1).



18. Figure 6: How does this analysis relate to the FFA and to the rest of the paper in general? I'm not sure of the value of these maps, given the little information the readers have on the 7 runoff inputs, and also because there is no clear patter identified. Perhaps the main information one can obtain is that anu and univu tends to be on the lower side, while cnrs and univk on the higher side. Yet, this doesn't say anything about the skills of these estimates, which would imply validation with gauge data at a number of stations.

Re: We decided to remove this subsection (and Figure 4-6) because it is not very relevant to the rest of the paper now.

19. P13, l2: after – > downstream

Re: revised.

20. Note that Figure 8 is referenced before Figure 7

Re: Thanks, we have re-ordered the texts and the figures, so this problem is fixed.

21. Sect 4.1 refers to return periods in Fig.7, hence in Fig.7 I advise to show return periods in place of frequencies. In any case, to be correct you should refer to those as annual frequencies of occurrence, to avoid confusion. Also, in Figure 7c, why not all distributions are shown?

Re: Thanks. We have revised the x-axis label and ticks to return period. We didn't show the results from Gamma and Gumbel because the fitting performance for these two fitting

distributions are the lowest among all the six distributions. But we have added them in the new figures in revised manuscript.

22. P14, l2: please give some details and possibly a reference on the downscaling procedure.

Re: Ok, the downscaling procedure is added to the Methods. A reference (Zhou et al., 2020) is also added for details about the downscaling method.

Zhou, X., Prigent, C., and Yamazaki, D.: Reasonable agreements and mismatches between land-surface-water-area estimates based on a global river model and Landsat data, Earth and Space Science Open Archive, p. 31, https://doi.org/10.1002/essoar.10504917.1, https://doi.org/10.1002/essoar.10504917.1, 2020.

23. P14, l3-6: To aid the assessment of water depths I suggest showing in Fig.8 a map or contour of the permanent water bodies. Clearly it is normal to have higher water depths in rivers and lakes, compared to areas normally dry. Also, I cannot find information about the terrain model, in particular whether it represents the river bed or some reference water level. This is important for this analysis.

Re: Thanks. We use Multi-Error-Removed Improved-Terrain DEM (MERIT DEM) as the terrain model (this is added in section 2.4 and the figure caption). Sorry that we decided not to add the permanent water bodies because of the following two reasons. First, the accuracy of the estimated permanent water still needs validation, while it is not appropriate to use permanent water from satellite (e.g., Landsat, Pekel et al., 2016) since it may not match the model estimation because limitation of satellites as well. Second, the permanent water for the deltas can be very small compared to the flooded area at the 100-yr return period. It may not help a lot to our analysis.

24. Figure 10: results shown in this figure are rather obvious. I suggest removing this figure as it brings little information. Over large inundation depths it is normal to have good agreement on whether there's inundation or not, as having poor agreement would mean huge differences in the results of the model used (hence very poor skills for some models).

Re: Thanks, this Figure and relevant texts have been deleted in the new manuscript.

25. P17, l11: return periods should not be expressed as percentage

Re: Thanks. We have corrected it.

26. p18, l10 and Figure 12: Is this the mean inundation of the 7 models? Clarify

Re: Yes, the map represents the mean values among all different experiments, with different runoff inputs, fitting distributions and two selected variables. However, this has been removed from the manuscript.

27. I find the analysis in Figure 12 of limited use, being a qualitative visual comparison with two other publicly available maps, but also resulting from modeling exercise with limited calibration. Similarly, the comments in p19, l14-18 are partly speculative. More rigorous validation with observed flooded areas would give much more strength to the paper.

Re: Thanks, the comparison of CaMa-Flood result to the other two sources (Figure 12 and subsection 4.5) has been removed in the revised manuscript.

We are working on the calibration/validation of the CaMa-Flood with observed flooded areas, but that work is out of scope of this study. So, we decided not to mention the validation in this paper.

28. P 21,l6: for flood impact assessment it is more interesting to know (even smaller) inundation depths in areas where people live or where economic assets are, rather than the inundation in the main channels, which has fewer fields of application.

Re: Thanks for the comments! The population exposure or GDP exposure to floods is one of key interests in flood damage assessment. In the new manuscript, we added analysis about the population exposure and economic exposure to floods in different continents. There are some interesting conclusions.

1. Asia will suffer the largest flood extent, population exposure and economic exposure among all the continents. While the proportion of the inundation area in North America is the largest, although the population and economic exposure ratio are very low in North America.

2. Africa has a lower ratio of inundation area, but relative high exposure of population and economy. This indicates that Africa replies very heavily the water and flood will cause significant damages.

3. The uncertainties in Africa is the largest, indicating that models are not consistent in Africa. This can be caused by the complexity of the topography and climate zones in Africa.

Detailed discussions can be found in the section 3.3.3 in the manuscript.

29. Figure 14 is unreadable and of limited use in the present form. It is impossible to get enough spatial details of a global inundation map at such small scales. Furthermore, the left and right column are almost indistinguishable. I suggest removing this figure and rather put it in the supplement, together with a number of inset panels zooming into some areas, especially those where the authors want to comment the results.

Re: The result for floods at 1-in-20 years return period has been removed from Figure 14. In order to see more details about the inundation, we added analysis on five other river basins (e.g., Amazon, Yangtze, Mississippi, Lena and Nile). Readers can find the details if checking the zoomed map for specific river basins. The maps are added in the supporting materials as Figure S1-S5.

30. Figure 15: What do you mean by the third (and fourth) row and the second row, in the caption? Is it related to the rows of Figure 14? If so it should be clearly stated.

Re: Yes. The captions links Figure 14. But in the new manuscript, the Figure 15 is deleted.

31. P23, l14-15: To be improved

Re: We have already rewritten the discussions in the new manuscript.

32. p24, l16: this is a model result for just one point in the entire world, hence it is completely irrelevant. Even more when looking at figure 6. Also (see lines 20-22), being in the middle of the 7 outputs doesn't mean it is more skillful. Validation with observed data is recommended.

Re: The point analysis helps analyze the uncertainties at different return period. It is not applicable to check all the points in the world, thus we selected one representative point (GRDC gauge near the basin outlet) for each river basin to be investigated. Beyond the point analysis, we also analyzed the inundation area, showing similar uncertainty changes for different return periods as the point results. We agree with the reviewer that one single model for one point is sufficient to the judgement of the best model. Therefore, in the new manuscript, we avoid justifying which model/fitting function is the best. But we focus on the uncertainty/sensitivity rather than the accuracy of the model. Validation of CaMa-Flood against river discharge has been provided in Hirabayashi et al., 2013. Model validation with observed data (e.g., inundation area, water altimetry) is now ongoing. However, we think the validation is out of the scope of this study, since we would like to discuss more on the sensitivities from different influential factors.