Replies to Referee #1

We are grateful for all the comments. In this discussion forum, we will briefly reply to the comments point by point. The detailed revisions will be shown in our manuscript and the final replies to reviewers.

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 this paper is discussing the model sensitivity to different climate and model conditions. So, we will reframe the structures and revise the contents in the following aspects.

1. Adding discussions in the Introduction part how the uncertainty/sensitivity is discussed and addressed in the previous literature.
2. Revise the introduction in order not to mislead the readers that we are validating our model.
3. The comparisons with two other methods will be removed from the results since they are not "validation" and they help little to sensitivity analysis.

- A single model with often extremely limited geographic scope

Re: Yes, 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 can be added, we can further investigate the flood extent sensitivity to model choice. However, this may need collaboration with different teams.

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

Re: Yes, we are trying to use different variables in the fitting process because we think different hydrological variables may 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.

- The contextual relevance of distribution goodness-of-fit

Re: This is lacking in the current version. We will try to find the relevance of AIC with the results and probably add a few other metrics for an overall evaluation.

- 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 especially in the developing countries. And the two datasets (JRC and GAR) used in this study are not appropriate. Therefore, we will first remove the current section 4.5 (comparison over the lower Mekong region). Then since we decided to focus on the sensitivity analysis at the global scale, we will not use any benchmark data.

****
Specific comments:

The introduction is mostly good, but could contain a richer discussion on why unstitch- ing 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: Additional literature and new discussions will be added in the introduction. All the text throughout the paper will be checked again and improved by professional English editing service.

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 to make our goal more clearly – focusing on the sensitivity rather than validation. The manuscript will be revised accordingly especially the introduction and results.

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 subse- quent 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: It is difficult to find observational values of the historical-level flood (e.g., 100 years return period). This is why we introduced the results from JRC and GAR to show how CaMa-Flood is compared with modelled (JRC) and more observation-based (GAR) results. (But this part will be removed in the new manuscript because we will more focus on the sensitivity analysis). There are various uncertainties which can lead to diversities in flood extent estimates while in this study we will first discuss a few while assume others are certain.

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. Com- mon application of FFA is to discharge, yet this is not done here. 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? 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$\neq$ 100-year streamflow$\neq$ 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: 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.

I can not see evidence that WAK is the best distribution because of it having 5 param- eters. 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: In this study, we tested six different fitting distributions (e.g., WAK, GEV, etc.). AIC is used to evaluate the fitting performance. However, AIC or other metrics (e.g., RMSE, bias, etc.,) cannot indicate if we should reject the distribution or not. Though, we do not aim to identify the distribution in this study but we mainly want to test how much of the variation can be caused if we use different fitting distributions.

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 will be removed in the new version. Regarding the middle one, we have one pre-described assumption that the users don’t know which runoff is the best. In this case, the users tend to use the ensemble 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 will be removed from the current 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: It is worthy trying to analyze the point uncertainty for all global grid cell. We will think how we can better illustrate/present the results. However, we doubt if it is necessary to work on the normalized results since they will not show the difference caused by biases in the mean value. The current point is selected randomly, so we need to explore whether this is representative at a global scale. We will also explore if AIC can explain part the diversities in this Figure.

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: Thanks. It is good to investigate the uncertainties in different climates and locations. This will be added in Section 5 with the global analysis of the uncertainties.

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 will remove section 4.5. While we may not add analysis in the US or western Europe because observations are still not available. Deleting this part will not affect the uncertainty analysis planned for this study.

Section 5 is strong, but will benefit from drawing on some of the above points. Gen- erally, 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.

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 in high quality, so it can be enlarged. 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.