Replies to reviewers – Submission titled "Historical and future changes in global flood magnitude – evidence from a model-observation investigation" [HESS-2019-388]

Replies to Reviewer #2

We would like to thank Reviewer #2 for the constructive comments that will help us to improve the quality of the manuscript. For clarity, we formatted reviewer's comments in *blue, italic text*, while our responses are formatted in normal text.

This manuscript describes the work of an extensive model study and in its final version will be for sure appreciated by the readers of HESS.

We thank the reviewer for the encouraging evaluation.

General Comments: As the study presented is quite extensive, it is sometimes difficult to follow study setup and all the analysis steps. Therefore, the authors should provide a detailed schematic, showing the main building blocks of their study and the different steps of analysis (preferably showing the section numbers in the schematic as well) to allow the reader to have a complete 'picture' of the study design, before embarking on the details in the main text.

We will include an additional figure at the start of Section 2 to provide a complete picture of the analyses presented in this study.

Additionally, due to the complexity of the study and details provided in the result section, I think a summary table or bullet points at the end of the study would be helpful for the reader to get a better overview of the key results obtained.

Thank you for pointing out this issue from the reader's perspective. In our revision, we will carefully revisit the manuscript to improve readability and simplify the contents where relevant. We will also consider your suggestions to format the main findings in bullet points to communicate the key points better.

Another important point is that the study uses 7-day annual maximum as a surrogate for 'food'. This fact needs to be made more explicit throughout the study to avoid misunderstandings from the general perception of flood, which would shorter (e.g. often 1-day). This is of importance, as the results might be quite different. I.e. a single day peak value trend study will show different results, not only in terms of magnitude of change, but also in terms of the flood hydrograph shape. E.g. if floods would become flashier in some location in future, it might look as if the trend of a 7-day maximum might not change at all or get smaller, but the peak day could be of much higher magnitude. The authors need to make sure they call the variable under investigation for what it is, i.e. not calling it 'flood', 'peak discharge' or 'streamflow maximum' to avoid misunderstanding of the results.

We thank the reviewer for the suggestion. We would like to note that in our preliminary analysis, we also analysed the annual maximum values of daily streamflow (i.e. 1-day flood time series, MAX index). Although trend at specific site may vary between these two indices, we found that the key conclusions (e.g. the regional pattern of increasing/decreasing trends, the consistency between trends exhibited from observed data and that obtained from simulated data) are quite similar, regardless which index

being used (i.e. MAX or MAX7 index). To address your concerns, we will revise our manuscript substantially to ensure: (i) there is sufficient information about the consistency between results introduced by MAX and MAX7 index, and (ii) the terminologies are used appropriately (e.g. replace "streamflow maximum" by "7-day streamflow maximum").

Along this line, I also think that the title '. . . changes in global flood magnitude . . . ' is also misleading. The study shows rather an 'global assessment of the 7-day annual maximum average value'. Please consider changing the title to better represent the content of the study.

We acknowledge the reviewer's concern and will consider a change in the title. A possible option is to replace "magnitude" by "indicator"(i.e. "Historical and future changes in an indicator of global flood hazard - evidence from a model-observation investigation"), due to the fact that MAX7 can also be used as an indicator for floods, and using MAX index has generally led to comparable results to that of MAX7 index (discussed above). From our perspective, this proposed title is not misleading, and can potentially reach a broader readership than using a too technical title such as "Global changes in 7-day annual maximum average value".

*Additionally, to avoid misinterpretations of your results please avoid using the term 'hazard' in its current form in the manuscript, as hazard means: hazard=risk*exposure (which is not the correct terminology here). The same also applies to the term 'risk' which is related to 'probability and consequences'.*

We noted that this study was developed from the perspective that 'hazard' (e.g. flood magnitude, frequency or inundation) is a component of 'risk' (i.e. Flood Risk = Hazard x Values x Vulnerability; Kron, 2005). From this point of view, we judge 'hazard' the appropriate terminology to refer to the MAX7 streamflow index. We acknowledge the reviewer's concern about the use of the term 'risk' and will carefully evaluate the manuscript and clarify/make changes where relevant to ensure the appropriateness of each terminology.

In this manuscript I feel that the GHM are used by the authors as 'black-box' that give some output. However, for this study to be valuable, it would be important that the authors would try to relate the observed differences/deviations in the outputs to the actual differences in the hydrological model setup. The authors just state ". . . there are potential effects of technical discrepancies to the findings which cannot be checked in the context of this study" (L 126).

We agree with the reviewer that the relationship between the model's structure and model's capacity in simulating trends in floods is an important aspect. However, addressing this comment is not straightforward, as there are a total of six models with many factors (e.g. routing schemes, spatial resolution, and parameterisations) that could individually or collectively lead to output discrepancies. These aspects (i.e. possible reasons for different trends simulated by different GHMs) in fact is still under-represented in the literature. From our perspective, this type of investigation deserves a separate paper by itself as the work involved should be tremendous, and potentially involved another set of simulations (e.g. to check the sensitivity of simulated trends corresponding to changes in a specific factor).

In the revision, we will refine the introduction to clarify the key objectives of our study, which is to compare trends observed by different models and the uncertainty in projected trends rather than to explore the mechanisms driving discrepancies in model outputs. We will also highlight the reviewer's comment in the conclusion as a potential research direction.

However, I think based on the model selection, the authors should have a notion of why they selected certain models and what the key differences are. Hence, the authors should at least try to come up (also based on past literature) with some sort of reasoning for model selection and also more importantly an interpretation of their findings. . .

In this study, we did not make any model selections. Specifically, we used all hydrological models that have produced discharge outputs for both Phase 2a and 2b at the time this study was initiated (June 2018). In the revision, we will highlight this fact better to avoid confusion.

For example, are the changes the models are giving as an output considered in line with the current understanding of the effects of climate change on floods or are there surprising results? I think this could be done in a separate paragraph discussing/comparing with previous literature.

Our manuscript has highlighted that the historical trends obtained in the present study are consistent with what has been reported in the literature (Lines 240-247). The reviewer suggested that simulated trends should also be linked to the current understanding of the effects of climate change. Although this aspect is important, we intend to not cover it as our objectives are not to attribute change in flood hazard to climate change or human activities. For historical trends (1971-2005; or ISIMIP2a), the focus was to compare model capacity in reproducing observed trends and compare the performance of simulations driven by observed (GSWP3 simulations) and modelled atmospheric forcing (GCMHIND simulations). The ultimate goal is to show the uncertainty of trends in the MAX7 index detected from the current GHM-GCM ensemble. For future trends (2006-2099), the focus was on the robustness of projected trends introduced by the ensemble members.

In addition, there is another ISIMIP investigation dedicating on river flow changes attribution, and thus we decided to exclude this aspect from this manuscript to avoid overlap.

In several instances in the manuscript, the authors are highlighting the 'substantial influence of the atmospheric forcing in driving the spatial structure of the simulated trend'. I think this is another important point that needs to be discussed in more detail in the discussion section, i.e. why to the hydrological models have little influence...

The hindcast simulations of the global climate model are forced by historical CO_2 (Katragkou et al., 2015), and so the timing of wet/dry periods or the spatial distribution of precipitation will be different from what has been observed in the past. As precipitation is arguably one of the most important inputs for streamflow simulation, it is expectable that GCMHIND trends will have a more prominent impact on the spatial patterns of simulated trends relative to model structure. We will consider including this justification in the revision.

Overall, I think a new separate discussion section of the results of such a complex analysis would be beneficial, as this would free up the room for a better refined summary and conclusion section, that focused on the key results and the overall implications of the results not just for the scientific world but also for the 'end-users', such as decision makers etc.

Thank you for suggesting this potential improvement. We will revisit the whole paper to better discuss the findings and improve the paper readability. Some opportunities for improvement have been identified, which we believe will help the paper streamlined better:

- Revisit our introduction to clearly state the research objectives and narrate the analyses.

- Include an additional figure (in line with our previous response) to show the overall framework of the study and how does it address the research questions.

- Simplify the contents where relevant, potentially in Section 3.1, to exclude redundant information and make the analyses more focus.

We will also consider your suggestion (i.e. having a separate discussion section) during our revision.

Specific Comments:

L37: For clarity, please provide significance level used in this study in parentheses.

We will add the level of significance (10% two-sided) in the revision for clarity.

L38: replace the term 'high-risk location'.

Thanks for noting, we will evaluate the terminologies across the manuscript, potentially using "locations robustly projected with increasing flood hazards".

L54: Please provide reverence to this statement

The following sentences (Lines 54-62) in fact has extended our discussion and provided some evidence and references for this statement. We noted that this may be unclear and will revisit this paragraph to ensure the statement is justified.

L77: What is 'factorial evidence' in this regard? Please elaborate.

"Factorial experiments" refer to studies analysing the effect of different factors (e.g. land use change) on the response variable (e.g. changes in floods), as well as the effects of the interactions among the factors on the response variable. In the context of hydrological modelling, the impact of atmospheric forcing, land use change and human water management on streamflow trends could be "turn on/off" to provide a full "factorial experiment design". In the revision, we will revise this statement to improve clarity.

L121-122: Please elaborate why the authors think that the 'naturalised runs and the human impact runs exhibit similar characteristics of trend' Would one not expect considerable differences?

We thank the reviewer for the suggestion, which will be incorporated in our revision. Some potential reasons are the spatial distribution of stream gauges, which may be biased toward regions with insignificant changes in human intervention within the reference period (1971-2005), or the inclusions of small catchments (more that 3000 catchments with area less than 9000km^2), and floods are more sensitive to changes in extreme precipitation relative to the accumulated basin-wide influence of human impacts.

L126: What are the 'potential effects'. Can you briefly elaborate.

The most pronounced effect comes from the difference in the versions of GHMs that were used in ISIMIP2a and ISIMIP2b. Specifically, ISIMIP2a was designed as an evaluation framework to improve the models for the projection phase isimip2b. As a result, the assessment using historical simulation (from 1971-2005) may not reflect the "true" model capacity in simulating trends in floods during the future period (2006-2099). We will elaborate on this fact and the potential effects of different model versions

in the revision. However, as mentioned in our response to Reviewer#1, a solid conclusion about these effects may not be available.

L127: Please also elaborate what the effects/impacts of this on the results are.

Thanks for your comment. We will revise the manuscript to elaborate on the potential effects/impacts of technical differences across GHMs, potentially including:

- Different drainage direction maps across different models could lead to gauging stations (in some rare cases) that do not lie on the river network (Masaki et al., 2017).

- Different models do not have the same set of coastal cells which may lead to some minor effect to the statistics when averaged across all simulation grid-cells.

- ORCHIDEE runs on 1-degree resolution but is routed at 0.5-degree resolution and thus influenced by a stronger spatial averaging that could lead to more flatten discharge time series.

L158: What is the rationale of 335 days. Please explain briefly.

The rationale of this choice is every single year must have at least 90% of streamflow data available. This criterion is a common data filtering condition in large-scale observation-based investigation (Do et al., 2017;Mallakpour and Villarini, 2015). This data criterion was chosen to fit the purpose of a hybrid observation-simulation study. We will consider clarify this methodological choice in the revision.

L172: Fig1: These colours are not 'safe' for colour-blind readers. Please use different colour combination

We will revise this figure in the revision to address this concern. Specifically, we will consider the use of an eight-color discrete palette that is colorblind safe (available in ggthemes R package at https://rdrr.io/cran/ggthemes/man/colorblind.html).

L184: 'Our preliminary analysis. . . did not lead to substantial changes'. So what were the 'not so substantial changes' one is wondering?

The preliminary assessment showed that the regional patterns of changes detected from MAX and MAX7 indices are generally consistent. We will clarify this point in our revision.

L192: Can you please name the 'three identified objectives' again as it is quite difficult to keep up with this extensive work.

We will consider provide the identified objectives as bullet points to remind readers about the focus of this study.

L210: To spare the reader from having to go to the original reference, please name the field significance test used and elaborate briefly what exactly is evaluated.

We will incorporate your suggestions in the revision by adding a brief explanation about the bootstrapping technique that was used.

L211: What 'Pearson's (spatial) correlation' was used? Reference? What variables are correlated?

Here we computed the Pearson's correlation *r* metric (Kiktev et al., 2003;Galton, 1886) to represent the spatial consistency between two sets of trends in MAX7 index. We will clarify this statistical technique to improve clarity in the revision.

L220: Please replace the term 'flood hazard' with something more appropriate to what has been done. This also applies to the subsequent usage, as well as the term 'floodrisk' later used in the manuscript.

As mentioned in our previous response, we think flood hazard is the appropriate term to refer to the magnitude of MAX7 index. Nevertheless, we will carefully evaluate the manuscript to ensure the most appropriate terminologies are used.

L245 & 493:to me it does not look like norther Europe has increasing trends. Scandinavia etc looks decreasing. . .. Please check.

We thank the reviewer for noting out this mistake – which should be "the northern part of Western Europe". We will revise the manuscript to ensure correct description is presented.

L258: I agree, very much with this point. The study analyses 'extremes (i.e. floods) but then model 'averages' are provided. His is counter intuitive. This can lead to strong underestimation of the actual changes. The usage of averages vs individual models that show extremes should be better discussed in the discussion section. Hence, I also agree with L 419.

Many thanks for your encouraging comment. To address the shortcoming of using model average, the subsequent analyses have therefore used the multi-model min/max/average of trends to communicate the results. We also discussed in our manuscript that "ensemble averages should not be used as a sole ground to infer changes in floods, as this may undermine the actual magnitude of simulated trends" (Line 291).

Considering the key objective of this study (i.e. to compare GHMs capacity in simulating floods and the uncertainty in projected trends) and the complexity of the manuscript in its current form (also noted by the reviewer), we propose to not focus on this aspect in the revision. However, we will make this methodological choice and associated rationale more prominent in the revision.

L281: is this really 'the spatial pattern of trends' that is evaluated or is it a cell by cell comparison? Please elaborate and have in mind that although a correlation is it can still mean that the overall spatial pattern (i.e. approximate location of increasing and decreasing trends) might still be correct.

We assume that the reviewer means that "the overall spatial pattern of increasing and decreasing regions might still be correct even when the correlation value is low". During our investigation, we have conducted some visual inspections which confirmed that a low correlation value usually reflect the inconsistency in the spatial pattern between two specific set of trends (an example was provided in the Supplementary). This metric was also used extensively in the climate literature (Kumar et al., 2013;Kiktev et al., 2003;Kiktev et al., 2007) to assess the spatial consistency of trends introduced by different gridded products.

L 370-384: The authors mention 'a significant difference between trend characteristics from all model grid cells compared to those obtained from the observation locations' and the conclude that '' that trends exhibited from observation locations are not a representative sample of trends obtained from all simulation grid cells'' (L379-380) And then call ''to improve data accessibility and expand streamflow

observational networks''. However, if there are such ''significant difference even in data rich regions, how can one justify expanding the network based on the previous finding? Instead to me this reasoning would rather require the need to improve our models instead (notwithstanding the fact that I agree with the data needs mentioned by the authors.)

We thank the reviewer for noting this out. We will carefully revise our discussion to incorporate this suggestion. Potential changes are (i) to elaborate more on model performance in data-rich regions, and (ii) highlight the need for improved capacity of GHMs in reproducing trends at the Conclusion.

L 460: Maye the authors can elaborate a little more what an 'flexible adaptation strategy' entails in terms of flood mitigation. Any suggestion on how this can be achieved under tight budgets. Can we as scientists not provide any guidance than just saying 'stay flexible' to those who have to take decisions know?

We will consider extending our discussion to include feasible strategies and guidance to address high uncertainty in projections of changes in flood hazards.

L531 & 534: Along the lines of improved GHM: It is not only important that the spatial patterns are being reproduced correctly but also that the timing of the high-flows/floods are being modeled correctly. I.e. 'the flood seasonality patterns can be used as ´ an additional metric to test large-scale hydro-logical models for their ability to reproduce the spatial and temporal flood characteristics.' (Hall and Bloschl, 2018, HESS). ´ As this would give more confidence that the models actually get the flood generation processes correctly.

We thank the reviewer for this constructive comment. We agree that the timing of flood is a useful metric. This statistic should also be considered in the assessment of model capacity in terms of reproducing flood characteristics at the global and continental scale. We will extend our conclusion and include some corner-stone references (Hall and Blöschl, 2018;Blöschl et al., 2017;Dettinger and Diaz, 2000) to incorporate your suggestion.

L 538: What does 'constraining ' entail? Please briefly elaborate. Would this prevent the model to adjust to changes in the flood generating processes, as one would expect to happen in some regions of the world. E.g. from snow-melt floods to rainfallgenerated floods?

This term (i.e. "constraining") refers to the process of using observations to constrain multi-model projections and is commonly used in the climate literature (Padrón et al., 2017;Allen and Ingram, 2002). The purpose of this process is to prevent climate models projecting an unrealistic state of the future climate system (Flato et al., 2013). The constraints are usually the global average values of variables that model developers judge to be important (e.g. the global mean top of the atmosphere energy balance, cloud feedbacks). From our understanding, this process will not violate the fundamental physical processes of the hydrological cycle. We will clarify this terminology in the revision.

L 550-559: I agree with this call, as this is very important. However, one needs to keep in mind that in many countries maintaining monitoring networks and data curation is/is considered too expensive. Hence it needs to be made clear to decision makers that such data is of importance. However, I know of cases where countries/agencies have been or are currently considering discontinuing their data networks, as they don't see the benefit or don't see their data being used (partly lack of proper citation of the (often freely available) original data source). This implication needs to be kept in mind when large

datasets of observational data are being compiled and subsequently only credit is given to the compiled data. . . This hides to the funding/responsible agencies the usage of their data (i.e. the original data source) and might lead to the misconception that their data is not being needed/downloaded and hence the data network can be discontinued and to allocate funds to more (perceived) useful sectors. . .

We thank the reviewer for the comment. We agree this is very important to make national data authorities aware of the importance of their works. We will specifically emphasize the role of data "end-user" in making streamflow data more FAIR by properly acknowledging the efforts and merits that data providers deserve.

Fig S5: Suggest using same y-axis scale for all panels on the left/right to be able to compare the regions better with each another.

We will revise the figure in our revision to ensure a consistent scale on the y-axis is used.

Reference

Allen, M. R., and Ingram, W. J.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419, 224-232, 2002.

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., Aronica, G. T., Bilibashi, A., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Fiala, K., Frolova, N., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Rogger, M., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Viglione, A., Volpi, E., Wilson, D., Zaimi, K., and Živković, N.: Changing climate shifts timing of European floods, Science, 357, 588, 2017.

Dettinger, M. D., and Diaz, H. F.: Global Characteristics of Stream Flow Seasonality and Variability, Journal of Hydrometeorology, 1, 289-310, 10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2, 2000.

Do, H. X., Westra, S., and Michael, L.: A global-scale investigation of trends in annual maximum streamflow, Journal of Hydrology, 10.1016/j.jhydrol.2017.06.015, 2017.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of climate models, in: Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change edited by: Stocker, T. F., Cambridge University Press, Cambridge, United Kingdom and New York, NY, 741-866, 2013.

Galton, F.: Regression towards mediocrity in hereditary stature, The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263, 1886.

Hall, J., and Blöschl, G.: Spatial patterns and characteristics of flood seasonality in Europe, Hydrol. Earth Syst. Sci., 22, 3883-3901, 10.5194/hess-22-3883-2018, 2018.

Katragkou, E., García Díez, M., Vautard, R., Sobolowski, S. P., Zanis, P., Alexandri, G., Cardoso, R. M., Colette, A., Fernández Fernández, J., and Gobiet, A.: Regional climate hindcast simulations within EURO-CORDEX: evaluation of a WRF multi-physics ensemble, 2015.

Kiktev, D., Sexton, D. M., Alexander, L., and Folland, C. K.: Comparison of modeled and observed trends in indices of daily climate extremes, Journal of Climate, 16, 3560-3571, 2003.

Kiktev, D., Caesar, J., Alexander, L. V., Shiogama, H., and Collier, M.: Comparison of observed and multimodeled trends in annual extremes of temperature and precipitation, Geophysical research letters, 34, 2007.

Kron, W.: Flood Risk = Hazard • Values • Vulnerability, Water International, 30, 58-68, 10.1080/02508060508691837, 2005.
Kumar, S., Merwade, V., Kinter III, J. L., and Niyogi, D.: Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations, Journal of Climate, 26, 4168-4185, 2013.
Mallakpour, I., and Villarini, G.: The changing nature of flooding across the central United States, Nature Clim. Change, 5, 250-254, 10.1038/nclimate2516, 2015.
Masaki, Y., Hanasaki, N., Biemans, H., Schmied, H. M., Tang, Q., Wada, Y., Gosling, S. N., Takahashi, K., and Hijioka, Y.: Intercomparison of global river discharge simulations focusing on dam operation—multiple models analysis in two case-study river basins, Missouri–Mississippi and Green–Colorado, Environmental Research Letters, 12, 055002, 2017.
Padrón, R. S., Gudmundsson, L., Greve, P., and Seneviratne, S. I.: Large-Scale Controls of the Surface Water Balance Over Land: Insights From a Systematic Review and Meta-Analysis, Water Resources Research, 53, 9659-9678, doi:10.1002/2017WR021215, 2017.