Response to reviewers

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

This document provides response to Editor and Reviewers for the above manuscript.

The comments from Editor/Reviewers are quoted in blue, italic text. Previous responses from the online discussion are formatted in normal text. New responses after revision are in red.

Response to Editor

Editor Decision: Publish subject to revisions (further review by editor and referees) (03 Nov 2019) by Louise Slater

Dear Authors,

Thank you for the detailed responses to the two reviews on your paper.

Overall, both reviewers were quite positive about the study, and I agree with many of their comments. They made some valuable suggestions regarding important topics such as: identifying the research gap and clarifying the motivation; providing an outline (or schematic) of the methodology; attempting to relate the results to the model characteristics (where feasible); clarifying the terminology to avoid misinterpretation; providing further discussion of results and interpretation of findings; and making some of the data available for reproducibility.

I find your proposed modifications appropriate and would like to invite you to go ahead and submit a revised manuscript, which will be returned to the reviewers for further review.

Sincerely,

Louise Slater

We would like to thank the Editor for her encouraging evaluation of our manuscript. Our revision has incorporated the suggestions from Reviewers, including:

- Substantial revisions of the introduction to better highlight the research motivation.
- A new figure has been added to Methodology section to summarise the research outline.
- Significant changes across the manuscript to ensure the results are presented/interpreted in a concise manner, and avoid potential misinterpretation. Major changes are: (i) new material in supplementary to better highlight the difference across models; (ii) significant revision of Section 2 and Section 3 to communicate the results more precisely; (iii) adding clarifications when needed; and (iv) restructuring Section 4 to make the key findings more appealing.
- We also provided csv file contains historical trends for MAX7 index across 3666 locations derived from GSIM dataset and all model simulations.

We would like to draw the Editor attention that the manuscript has an additional author (Camelia-Eliza Telteu), and has corrected the name order of another author (Julien Eric Stanislas Boulange). Camelia has synthesized the key differences and similarities across GHMs (supplementary Section 1), and contributed substantially in the revision to improve the quality of this manuscript.

Response to Reviewer #1

We would like to thank Reviewer #1 for the constructive comments which will help to improve the quality of our manuscript. Each of the reviewer's comments is quoted in *blue, italic text*, followed by our reply formatted in normal text. Additional responses (after online discussion) after revisions are in red.

General comments

Not content related: the authors list reads like a "manel". Sad to see so little inclusion of other genders than male. Maybe something to think about for future studies?

We thanks the reviewer for their note about the broader diversity in the author panel in such global scale initiative. We will aim for improved inclusion of gender diversity in future investigations.

I do miss the overarching research motivation of this study. Even though the introduction contains a good amount of background literature and describes the three research objectives set, it does not become clear what the research gap is that the present study intends to fill. Or in short, how are the objectives derived and what is the (societal) relevance of the study? Please ensure this is clearly and concisely explained in the revised version of the manuscript.

Thank you for your note that the research motivation should be presented more prominently. We will revise the introduction and consider revisions focusing on: (i) highlighting the contribution of this study to address existing research gap in the field, (ii) including the motivation from the perspective of global hydrological model users (e.g. decision makers).

We have revised the Introduction substantially to make our motivation clearer. The objectives have also been rewritten into bullet points for improved readability.

The models used in the study differ in quite some characteristics/schematization. An overview of these differences would be very useful (also possible as supplement). Besides, none of the deviations between results obtained with different GHMs is related to those different characteristics/schematization. I would guess that they do play a big role in explaining the obtained results and thus I recommend extending the manuscript with such an analysis (or a clear and convincing statement why not).

We appreciate the reviewer's suggestion about exploring the role of model conceptualization in simulating trends in floods. We will explore the possibility of having a supplementary table outlining the differences among GHMs in the revision.

A more detailed description of GHMs as well as key difference across models (and between versions of the same model) have been added in Supplementary.

We note, however, that a detailed documentation of differences in model characteristics does not necessarily enable evidence of why the models produce different outputs. From our understanding, previous studies have related the impacts on peak flows of some specific processes such as routing

scheme or reservoir algorithms (Zhao et al., 2017;Masaki et al., 2017), but it does not mean that a similar impact is presented in simulated trends (also highlighted in our manuscript at Line 85). This is also a motivation of this study, as we want to compare the trends of a high flow indicator simulated by different GHMs.

To nevertheless incorporate the reviewer's comment, we will make the main objective of this study (i.e. a comparison of model capacity in simulating trends in a flood indicator and an assessment of uncertainty of projected changes in floods) more prominent in the introduction. We will also mention the reviewer's suggestion (i.e. to assess the impact of model schematization on changes in flood hazard) in the conclusion as a potential research direction.

The Introduction has been revised to better communicate key motivations and objectives of the study. We also added, in the Conclusion, a potential research direction to explore the reasons leading to discrepancies in trends simulated by different models.

The manuscript reads bit lengthy sometimes. To some extent, this is the result of reporting a lot of numbers, which are partially also provided within tables. My recommendation would be to let the tables and figures speak for themselves and to check which reported numbers can be neglected as they are not directly needed for understanding the methodology or results. Plus a check whether more concise wording could be used.

The manuscript will be carefully revisited to focus more on the key findings, reduce any redundant information and present the results in a more concise manner. We will focus especially on Section 3.1 where we will remove any descriptions that have already available in the tables or figures.

We have revised the manuscript substantially to reduce redundancy, and to improve readability wherever possible (the vast majority of these changes are in Section 3). We also reorganized Section 4 to improve readability and better highlight the key findings.

This study is very much focussed on data and their analysis. Extending the implication of the findings of the study to the societal dimension would greatly benefit the manuscript to make the results more tangible and applicable.

We thank the reviewer for the suggestions on extending the study findings to societal dimension. We note that the key motivation of this study is to explore the level of consistency of trends detected from streamflow observations and model simulations, which is currently underrepresented in the literature. As a result, we would like to focus on this research pathway and will refine the introduction to highlight this objective better.

The Introduction is now revised substantially to better communicate our research pathway (to explore the uncertainty of model-based inferences on changes in floods).

I would very much welcome it if at least the data pairs for the observation stations are provided via a supplement. This would be in the spirit of FAIR hydrologic modelling, thus increasing reproducibility of your findings.

We will upload the observed and modelled trends at each station as a supplementary (csv files) together with the revised manuscript.

The csv file containing geographical coordinates and historical trends (calculated from all datasets) across 3,666 locations has been uploaded as Supplementary.

Specific comments

• P1/L29+L30: An agreement of 12-25 % can hardly be named "moderate agreement", can it?

The abstract will be revised to ensure appropriate terminologies are used, potentially by changing "moderate" to "low-to-moderate".

• P1/L31 "significant differences": please specify what kind of differences you refer to.

We will clarify that the characteristics of trends (trend mean, trend standard deviation) simulated by GHMs forced with historical climate is significantly different to that simulated by GHMs forced by bias corrected climate model output.

The abstract was revised as outlined in the two responses above.

• P2/L52 "specific regions": specific regions such as? Please specify.

We will clarify that this the statement may only applicable where rainfall plays the dominant role in flood occurrence.

Clarification added.

• P2/L53 "recent evidence": what evidence? Please provide name, source, etc.

The following sentences (Lines 54-56) in fact have extended our discussion and provided some evidence for this statement. We noted that this may be unclear and will revisit this paragraph to ensure the statement is justified.

We have added two references to justify this statement.

• P2/L52-L66: what about the role humans play in changing flood hazard? Please add this dimension to the paragraph.

We will add the impact of human activities to changes in flood hazard at the end of this paragraph.

Impact of human activities added.

• P3/L77 "factorial experiments": what are you referring to with this term? Please explain or use more common terminology.

"Factorial experiments" indicate 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 interactions among the factors on the response variable. In the context of hydrological modelling, the impact of atmospheric forcing, land use change and other drivers of change on streamflow trends could be "turn on/off" to provide a full "factorial experiment design". We will rewrite this statement to improve readability in the revision.

The statement was rewritten and "factorial experiments" is no longer used.

• P3/L87-L100: The description of the research objectives could profit from using bullet points

We thank the reviewer for their suggestion. We will consider using bullet points to separate the research objectives.

The Introduction has been revised, and the objectives are now presented in bullet points.

• P4/L105-L108: why are these models used? Why are some others available within ISIMIP not used? Please clarify.

There is no model selection in this study. We actually used all GHMs that have provided discharge data within Phase 2a and 2b simulations at the time this study was initiated (June 2018). In the revision, we will highlight this fact in a transparent manner.

We have added clarification of model choice in section 2.1.

• P4/L122-L126: As mentioned in the general comments, a (technical) description of the models is needed. A particular focus should be on how the models changed between ISIMIP2a and ISIMIP2b and whether those changes may have influence results (or not). Possible changes in e.g. functionality, spatial resolution, etc. may have had a great impact on results and thus affecting the comparison performed in your study. I thus strongly disagree that checking this is outside the context of the study.

Similar to our response to the general comment from the reviewer, a precise conclusion about the impact of changes in models (e.g. functionality, spatial resolution) on trends in floods should be based on a full multi-model experiment (i.e. to compare trends simulated by different versions of the same GHM), which is unfortunately not readily available. Although we are aware that changes and bug-fixes done in MPI-HM affect only the human impact simulations (and the influence is insignificantly), it is not straightforward to generalize this conclusion. We will aim for some extended discussion, but would like to keep our statement as-is (i.e. checking the affects is outside the context of the study).

We also note that the issues raised here and in the earlier comment show the need for the next step of model inter-comparisons which should focus on diagnosing the reasons for differences across models. We have mentioned about this need in our manuscript (Lines 542-547) and will consider make this call more prominent in the revision.

We have added a new sub-section in Supplementary (section 3.3) to illustrate how modifications in GHMs could influence the results. As only WaterGAP simulations are available for this analysis (the

impact is minor), it is not possible to draw a common conclusion across all models, we have noted in both the manuscript and Supplementary that the potential effects of technical discrepancies cannot be checked in the context of this study.

• P6/L144-L147: what was the reason to not only use the un-routed runoff for all catchments? Wouldn't this increase comparability between results as it removes (unnecessary) transformation of results and units?

The reason is that for large catchments, observed discharge and unrouted runoff are not comparable. In some very large basins, it takes one to three months for upstream runoff to reach river mouth through the channels (and be measured as discharge here by some observing gauges). The same magnitude of basin total runoff, depending on its spatial distribution (i.e., evenly distributed versus concentrated in the downstream), could generate rather different discharge after routing. Therefore, we adopt different procedures for large and small basins to achieve maximum consistency in model-observation comparisons.

• P6/L149 "catchment area": which area estimates did you use? For all models the same? Per model based on catchment delineation? Please clarify to avoid that data was used inconsistently.

We only used the reported catchment area of each stream-gauge in this calculation. We will clarify about this technical aspect in the revision.

Clarification added.

• P7/L177-L185: great you are pointing out the differences in methodology!

Thank you for your encouragement.

• P11/L228-L230: these lines read as if they should not be part of the methodology, rather of the results/discussion section. The fact that that there may be 'hot-spots' of future flood hazard should be discussed in more detail and thus deserves a more prominent location in the manuscript.

We appreciate the reviewer's suggestion. We will revisit this paragraph and consider highlighting the "hot-spots" aspect more in the revision.

Section 2.4.2 has been revised to clarify the methodology. However, we decided to not extend our discussion about "hot-spots" in flood hazards – considering the high uncertainty presented in the GCM-GHM ensemble. From our perspective, the current discussion about "high-risk" locations are under-sampled is appropriate in the context of this investigation.

• P11/L231 "each grid-cell": each grid-cell or only those paired with a GSIM-location? Does not become very clear from reading.

This analysis was conducted for each grid-cell across the globe, regardless there is stream-gauge or not. We will clarify about this to avoid confusion.

We revised the description to "each grid-cell available in the discharge simulation grid" for clarity.

• P11/L235-L237: why was this done? What does it add?

This step was included to assess whether the locations that robustly projected with increasing/decreasing trends in flood hazard (i.e. the magnitude of MAX7 index increases/decreases significantly during the 2006-2099 period) has been observed adequately by the current streamflow observation system.

We will revisit this section in the revision to improve clarity.

Section 2.4.2 has been revised substantially for improved clarity.

• P11/L240-L247: You describe the observation and simulations, but you do not mention the reasons behind it. Why are certain areas experiencing increases and others decreases? Is it all hydrology or not? Can we say something about the driving factors behind it? Please add.

We acknowledge the reviewer's perspective about the importance of attributing changes in flood hazards to hydrological or climatic mechanisms factors. We noted, however, that the key objective of the present study is to assess model capacity rather than exploring the mechanisms driving changes in flood hazard. The reviewer also noted that the manuscript has been quite complex in its current state already. As a result, we propose to not include these discussions in the revision. Instead, we will make our objectives clearer, and will clarify that the paper does not focus on explaining the mechanisms driving changes in floods.

The Introduction was revised to further highlight the motivation and objectives of our investigation.

• P12/L255+L256: What is the implication of this finding?

We discussed about the implication of this finding at line 290, in which we suggested that averaging will reduce the magnitude of trends and thus ensemble average should not be used as a sole ground to infer change in floods. This is also a motivation for us to provide the range of trend characteristics across all ensemble members.

We will revisit our discussion to communicate this implication clearer.

P12/L261+L262: So, if it is not visible through Figure 2, how can it be an alternative explanation? This sounds contradictory to me – either it's possible based on your results or your results say it's not a thing. Please clarify.

The intention of this statement was to indicate that we would explore GHM's performance in more detail (i.e. through the next paragraphs/sections) because Figure 2 alone was insufficient to explain such feature.

We found this statement may be confusing and will revise it to improve clarity.

Section 3.1 has been revised substantially to incorporate the two suggestions above from the reviewer.

Figure 2: A bit bigger figure (maybe with subplots of USA, EU) would help seeing the differences between differences between historical trends.

We note that the Supplementary has included sub-region plots for this figure. This figure is also useful to highlight the "white spaces" over many regions, which was then linked to our call for more streamflow observation.

We will explore the options to improve graphical quality of this figure in the revision. Some possible options are to include the vector graphic or use another colour pallet.

Figure 2 has been revised, but the focus was not on highlighting trends in sub-regions. Instead, we added scatter-plots to highlight the difference in simulated trends of GSWP3 and GCMHIND. The figure will be provided as vector graphic to ensure high quality of the final image.

• P14/L285-L287 and Table 3: what are possible reasons for the different model results? Model structure, processes simulated per model, spatial resolution, routing schemes applied or something else? Would be great if you could elaborate a bit on this.

From our perspective, this question is not straightforward to answer due to the number of participating models (six) and the many factors involved (e.g. the individual and collective effects of differences in model conceptualization, spatial resolution and routing scheme). These aspects (i.e. possible reasons for different trends simulated by different GHMs) in fact is s till under-represented in the literature. Even when model differences are documented extensively, it is still challenging to precisely attribute output discrepancies to a specific (or a set of) factor(s) without supports from another set of GHM simulations (e.g. checking the sensitivity of simulated trends corresponding to changes in a specific factor). Nevertheless, we will consider to elaborate about potential sources of differences in model outputs in the revision. We will also highlight this in the conclusion as a potential research pathway.

We have expanded our supplementary to cover the key features of different models (supplementary Section 1). As the impact of model structure differences to outputs cannot be explicitly identified (supplementary section 1.2), we emphasized in Conclusion section the call for more investigations to explore the reasons of output discrepancies (despite having a common climate forcing as input).

• P14/L290+L291: if it should not be used as "sole ground", what other measures would you (like to) use to infer changes in floods?

We will clarify in the revision that the range of all ensemble members should be used to illustrate the spread of simulated trends (e.g. the information showed in Table 4).

We have revised this section to provide a better narrative for these ideas.

• P23/L463+L464: Does that mean your results are not usable to help inform flood management practices in less well-observed areas? What would be the implications? Please elaborate briefly on the consequences of your results.

The intention of this statement is to set the stage for our next analysis which shows the regions projected with increasing flood hazards are under-sampled, and ultimately leads to our call for more attention to improved streamflow observations. We will revise our discussion to improve the narrative of these ideas.

Section 3.3 has been revised substantially to improve clarity.

• P24/L475-L478: What would be ways forward to reduce the dependency on not evenly spatially distributed observation systems? Which opportunities do, for instance, remotely sensed data products bring? Please put your findings into context, here and/or the conclusions section.

This statement was used as a ground for our call (at the conclusion) for more FAIR streamflow observations to support hydrological research. We acknowledge that the narrative may need improvement and will revisit the paper, potentially including some of the reviewer's suggestions.

We have added a call for using remotely sensed data products and runoff reanalysis to offset observation scarcity at the end of this section.

P26/L533-L537: It should be added that also the routing schemes of GHMs should improve, not only the runoff. Well timed runoff with right magnitude can still result if inaccurate streamflow if the routing scheme is too simplistic. Vice versa, higher-order routing schemes cannot perform at their best if input runoff is not accurate. Relevant literature: Hoch et al., 2019 (https://doi.org/10.5194/nhess-19-1723-2019) and Zhao et al., 2017 (https://doi.org/10.1088/1748-9326/aa7250).

We will extend our discussion to include the importance of the routing scheme on GHMs' performance and the need to improve this important feature in future GHM generations.

Our conclusion has been extended to incorporate the reviewer's comment.

P27/L550-L559: I very much agree with this, well written!

We thank the reviewer for their encouraging comment.

Response 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. Additional responses after revisions are in red.

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.

New figure added.

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.

We have revised Section 4 substantially to improved clarity and readability.

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").

We have carefully revisited the manuscript to avoid misinterpretation of our findings.

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".

In line with our response above, the abstract, introduction and conclusion have been revised to emphasize that the findings were based on analysing MAX7 index. As MAX7 index can also be used as a proxy of flood magnitude, we proposed to keep the title as-is (i.e., "Historical and future changes in global flood magnitude - evidence from a model-observation investigation") to attract the broad readership of HESS.

*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.

We have clarified our definition of "risk" at the Methodology, Results, and Conclusion sections to avoid misinterpretations.

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.

To address the reviewer's concerns, we have provided a summary of model characteristics and mention the impacts of technical discrepancies wherever possible (section 2.1, supplementary section 1.2, and supplementary section 3.3). As an explicit statement of technical discrepancy impacts on simulated trends was not available through our investigation (also discussed in our responses to Reviewer#1), we have included a call (in the Conclusion) for future research to explore the reasons behind this feature.

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.

We have clarified model choices in the revision.

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.

To further incorporate the reviewer's suggestion, we have reorganize the final section (Summary and conclusion) to emphasize that the results of our findings are consistent to previous studies (the first point of the summary).

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.

We have added the above discussion at the end of Section 3.1 to clarify that GCM outputs generally have lower capacity to simulate the spatial structure of weather extremes, thus the lower capacity of GCMHIND trends in MAX7 index is somewhat expected.

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.

In the revision, we have revised the manuscript and supplementary substantially to make the key findings better come through. We note, however, that we decided to keep the headings as-is.

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".

The abstract has been revised to address the two concerns above from the reviewer.

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.

We have added two references to justify our statement.

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.

We have revised this statement.

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.

This paragraph has been revised (the above response was added) to incorporate the reviewer's comment.

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.

In line with our response to Reviewer#1, we have extended our discussions, and included additional information in supplementary to elaborate on the effects of technical discrepancies.

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.

We have revised this statement to elaborate the impacts of differences in the number of coastal grid-cells.

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.

We have added a brief explanation for this methodological choice.

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).

This figure has been revised.

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.

Clarification added.

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.

We have revised this statement to incorporate the reviewer's suggestion.

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.

We have added a note that the technique is briefly explained in Table 2.

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.

Clarification added.

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.

In line with our response to the reviewer's general comments, we have revisited the manuscript and clarify our definition of "risk" in the revision.

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.

We have fixed this mistake.

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.

To incorporate the reviewer's suggestion, we have revised section 3.1 to highlight the fact that multi model average may potentially mask out individual trends. As a result, the full range of the simulation ensemble was reported to reflect the uncertainty underlying the results.

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.

We have revised the concluding remark of this paragraph into "it is therefore crucial to improve **not only models' capacity**, but also data accessibility and expand streamflow observational networks …"

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.

We extended this statement to cover parts of the reviewer's comment.

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.

We have revised our conclusion to incorporate the reviewer's comment.

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.

Clarification added.

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.

This paragraph (and the acknowledgement) has been revised.

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.

The figure has been revised.

References

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.
Zhao, F., Veldkamp, T. I., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., and Portmann, F. T.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environmental Research Letters, 12, 075003, 2017.

Historical and future changes in global flood magnitude – evidence from a model-observation investigation

Hong Xuan Do[(1)(2)(3)(*)], Fang Zhao[(4)(5)(*)], Seth Westra[(1)], Michael Leonard[(1)], Lukas Gudmundsson[(6)], Julien Eric Stanislas Boulange[(7),] Jinfeng Chang[(~~7~~8)], Philippe Ciais[(~~7~~8)], Dieter Gerten[(5)(~~8~~9)], Simon N. Gosling[(~~9~~10)], Hannes Müller Schmied[(~~10~~)(11)(~~12~~)], Tobias Stacke[(~~12~~)~~, Boulange Julien Eric Stanislas~~[(13)], Camelia-Eliza Telteu[(11)], Yoshihide Wada[(14)].

(1) School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, Australia.
(2) Faculty of Environment and Natural Resources, Nong Lam University, Ho Chi Minh City, Vietnam.
(3) School for Environment and Sustainability, University of Michigan, Ann Arbor, Michigan, United States.
(4) School of Geographical Sciences, East China Normal University, Shanghai, China.
(5) Potsdam Institute for Climate Impact Research, Potsdam, Germany.
(6) Institute for Atmospheric and Climate Science, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland.
(~~7~~7) Center for Global Environmental Research, Japan.
(8) Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ/IPSL, Université Paris Saclay, 91191 Gif sur Yvette, France.
(~~8~~9) Geography Dept., Humboldt-Universität zu Berlin, Berlin, Germany.
(~~9~~10) School of Geography, University of Nottingham, Nottingham, United Kingdom.
(~~10~~11) Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany.
(~~11~~12) Senckenberg ~~Leibnitz~~Leibniz Biodiversity and Climate Research Centre (SBiK-F), Frankfurt am Main, Germany.
(~~12) Max Planck~~13) Institute ~~for Meteorology, Hamburg~~of Coastal Research, Helmholtz-Zentrum Geesthacht (HZG), Geesthacht, Germany.
~~(13) Center for Global Environmental Research, Japan.~~
(14) International Institute for Applied Systems Analysis, Laxenburg, Austria.

(*) *Corresponding authors:* Hong Xuan Do (hong.do@adelaide.edu.au) *and* Fang Zhao (fangzhao@pik-potsdam.de)

Abstract. To improve the understanding of trends in extreme flows related to flood events at the global scale, historical and future changes of annual maximum of 7-day streamflow are investigated, using a comprehensive streamflow archive and six global hydrological models. The models' capacity to characterise trends in annual maximum of 7-day streamflow at the continental and global scale is evaluated across 3,666 river gauge locations over the period from 1971 to 2005, focusing on four aspects of trends: (i) mean, (ii) standard deviation, (iii) percentage of locations showing significant trends and (iv) spatial pattern. Compared to observed trends, simulated trends driven by observed climate forcing generally have a higher mean, lower spread, and a similar percentage of locations showing significant trends. Models show a low-to-moderate capacity to simulate spatial patterns of historical trends, with approximately only 12-25% of the spatial variance of observed trends across all gauge stations accounted for by the simulations. Interestingly, there are statistically significant differences between trends simulated by GHMs forced with ~~historical~~observational climate and forced by bias corrected climate model output during the historical period, suggesting the important role of the stochastic natural (decadal, inter-annual) climate variability. Significant differences were found in simulated flood trends when averaged only at gauged locations compared to when averaged across all simulated grid cells, highlighting the potential for bias toward well-observed regions in the state-of-understanding of changes in floods. Future climate projections (simulated under RCP2.6 and RCP6.0 greenhouse gas concentration scenario) suggest a potentially high level of change in individual regions, with up to 35% of cells showing a statistically significant trend (increase or decrease; at 10% significance level) and greater changes indicated for the higher

concentration pathway. Importantly, the observed streamflow database under-samples the percentage of ~~high risk~~ locations consistently projected with increased flood hazards under RCP6.0 greenhouse gas concentration scenario by more than an order of magnitude (0.9% compared to 11.7%). This finding indicates a highly uncertain future for both flood-prone communities and decision makers in the context of climate change.

1 Introduction

Global hydrological models (GHMs) are critical tools for diagnosing factors of rising trends in flood risk (Munich Re, 2015;Swiss Re, 2015;Miao, 2018;Smith, 2003;Guha-Sapir et al., 2015;CRED, 2015), and can help identify the contribution of changing flood hazard characteristics relative to the changing exposure of human assets to floods. GHMs are also used to project future changes in flood hazard, owing to their ability to simulate streamflow under projected atmospheric forcing. Using GHM simulations, several studies have found more regions showing increasing trends than decreasing trends in flood hazards at the global scale, and have attributed these changes to anthropogenic climate change ~~(Dankers et al., 2014;Arnell and Gosling, 2014;Alfieri et al., 2015;Kettner et al., 2018;Willner et al., 2018;Asadieh and Krakauer, 2017).~~(Dankers et al., 2014;Arnell and Gosling, 2014;Alfieri et al., 2015;Kettner et al., 2018;Willner et al., 2018;Asadieh and Krakauer, 2017). The pattern of increasing trends obtained from GHM simulations is consistent with observations of increases in precipitation extremes ~~(Westra et al., 2013;Westra et al., 2014;Donat et al., 2013;Guerreiro et al., 2018)~~(Westra et al., 2013;Westra et al., 2014;Donat et al., 2013;Guerreiro et al., 2018) that have been used by a number of studies as a proxy to suggest that flood hazard may increase as a result of climate change (Alfieri et al., 2017;Pall et al., 2011;IPCC, 2012;Forzieri et al., 2016).

The inference of changes in flood hazard following the same direction as extreme precipitation may be appropriate over ~~specific regions~~regions where rainfall plays the dominant role in flood occurrence (Hoegh-Guldberg et al., 2018;Mallakpour and Villarini, 2015;Mangini et al., 2018), but recent evidence based on instrumental trends in flood hazard suggests it is not necessarily globally applicable~~.~~ (Ivancic and Shaw, 2015;Blöschl et al., 2019). This is due to a 'dichotomous relationship' between trends exhibited in extreme precipitation and extreme streamflow ~~(Sharma et al., 2018)~~(Sharma et al., 2018), highlighted in recent observation-based studies of trends in streamflow magnitudes (Wasko and Sharma, 2017;Do et al., 2017;Hodgkins et al., 2017;Gudmundsson et al., 2019). The hypothesised reason for this potentially inconsistent relationship is the complexity of the drivers of flood risk (Johnson et al., 2016;Blöschl et al., 2017;Do et al., 2019;Berghuijs et al., 2016), with the implication that historical and future changes to flood hazard at the global scale are unlikely to be reflected by changes to a single proxy variable alone, such as annual maximum rainfall. For example, even though trends in extreme flows are highly correlated to changes in extreme rainfall when rainfall plays the

dominant role (Mallakpour and Villarini, 2015;Blöschl et al., 2017), snowmelt-related flood magnitude has been found to decrease in a warmer climate, potentially due to a shift in snowmelt timing (Burn and Whitfield, 2016;Cunderlik and Ouarda, 2009). The sign of change is also unclear for locations where antecedence soil moisture plays an important role (Woldemeskel and Sharma, 2016;Sharma et al., 2018), owing to the combined influences of seasonal/annual precipitation, potential evaporation and extreme precipitation (Bennett et al., 2018;Ivancic and Shaw, 2015;Leonard et al., 2008;Wasko and Nathan, 2019). The sensitivity of changes in streamflow to anthropogenic influences such as urbanization, dams and reservoir operations, or river morphology (FitzHugh and Vogel, 2011;Slater et al., 2015) further suggests that it is not possible to use trends in extreme precipitation alone to infer changes in flood hazards.

To better understand historical and future trends in streamflow, the emphasis has therefore moved to analysing trends directly in streamflow measurements. Investigations using streamflow observations at global, continental and regional scales (see ~~Do et al. (2017)~~Do et al. (2017) and references therein) have generally detected a mixed pattern of trends, with some global-scale studies finding more stations having decreasing trends than increasing trends (Do et al., 2017;Hodgkins et al., 2017;Kundzewicz et al., 2004). These conclusions appear *prima facie* to be inconsistent with model-based evidence, which generally suggests the opposite (more locations showing increasing trends). However, varying sampling strategies, statistical techniques and reference periods make it difficult to derive a common perspective of trends in global flood hazards from a composite of observational and modelling studies. In addition, data coverage limitations (Hannah et al., 2011;Gupta et al., 2014;Do et al., ~~2018b~~2018a) remain a barrier to reliably benchmarking trends over some areas such as the flood-prone regions of South and East Asia.

GHMs, with the advantage of better spatial coverage, remain an important line of evidence about historical and future trends. GHMs also enable ~~'factorial' experiments~~the possibility to explore the individual roles of atmospheric forcing, land use change and other drivers of change on streamflow trends~~.~~ by including or excluding a specific factor from simulation setting. However, ~~unlike climate models, for which the performance in terms of reproducing trends of extreme precipitation~~no study has ~~been~~ evaluated ~~substantially (Kiktev et al., 2003;Kiktev et al., 2007;Kumar et al., 2013;Sakaguchi et al., 2012),~~ the performance of GHMs ~~has~~in terms of reproducing trends of streamflow indices, including flood indicators. To date, GHMs have been assessed ~~mostly~~extensively on their capacity to represent physical features of the hydrological regime, such as streamflow percentiles, the seasonal cycle ~~or the timing of peak discharge (Gudmundsson et al., 2012a;Zaherpour et al., 2018;Beck et al., 2017;Zhao et al., 2017;Veldkamp et al., 2018;Pokhrel et al., 2012;Biemans et al., 2011;Giuntoli et al., 2018). Streamflow~~, or the timing of peak discharge (Gudmundsson et al.,

2012a;Zaherpour et al., 2018;Beck et al., 2017;Zhao et al., 2017;Veldkamp et al., 2018;Pokhrel et al., 2012;Biemans et al., 2011;Giuntoli et al., 2018). Nevertheless, streamflow variability can be subject not only to long-term changes in atmospheric forcing, but also to climate variability (e.g. inter-annual, inter-decadal) as well as human activities across the drainage basin (Zhang et al., 2015;Zhan et al., 2012). Thus, the GHMs' capacity to represent physical features of a hydrological regime is not necessarily sufficient to determine their performance in simulating characteristics of trends ~~in extremes.~~ . The absence of a holistic understanding of GHMs' capacity to simulate trends implies that model-based inferences on changes in flood hazards are highly uncertain (Dankers et al., 2014), limiting the usefulness of GHMs in developing flood adaptation policy in a warming climate.

To ~~better understand~~address this limitation and further improve GHMs' applicability, this study provides the first comprehensive evaluation of GHMs' capacity ~~of GHMs~~ in simulating historical trends ~~in extreme~~ of a flood hazard indicator. This study also explores the uncertainty in developing projected changes in flood hazards using GCMs-GHMs ensemble. Specifically, we used the Global Streamflow Indices and Metadata (GSIM) archive (Do et al., 2018b;Gudmundsson et al., 2018), to-date the largest possible global streamflow ~~and potential implications for the development of projections, this study focusses on three research objectives. The first objective is to evaluate the capacity of GHMs~~database, to identify observed changes in annual maximum of 7-day streamflow (MAX7 index) over the 1971-2005 period. Streamflow simulations, available ~~at http://www.isimip.org~~ through the Inter-Sectoral Impact Model Intercomparison Project ISIMIP phase 2a and 2b ~~(Warszawski et al., 2014), to simulate trends in observed streamflow extremes during the 1971-2005 historical period.~~ (Warszawski et al., 2014), were used to derive historical (1971-2005) and projected (2006-2099) changes in MAX7 index simulated by GHMs. Observed and simulated trends were then analysed to achieve three research objectives.

- Objective 1: to evaluate the capacity of GHMs to reproduce observed trends of an indicator of flood hazard (MAX7). The particular interest is in reconciling ~~observed~~model- and ~~simulated trends in~~observation-based inferences of historical ~~streamflow extremes~~changes in flood hazard at the global and continental scale ~~using the Global Streamflow Indices and Metadata (GSIM) archive (Do et al., 2018a;Gudmundsson et al., 2018b), to-date the largest possible streamflow observations database. GSIM has been used in recent global scale investigations and is also an important source for the production of GRUN, a data-driven century long runoff reconstruction (Ghiggi et al., 2019). The second objective is.~~
- Objective 2: to determine the representativeness of observation locations (streamflow gauges) in GHM simulations ~~by comparing trends simulated at these locations to trends simulated across all land grid points of GHMs.~~. This objective is motivated by the sparse coverage of streamflow observations over several regions (e.g.

South and East Asia), which could lead to biased inferences of observation-based studies over large spatial domains wherever gauges are not a representative sample. ~~The third and final objective is~~

- Objective 3: to assess the implication of model uncertainty for projections of flood hazard, focusing on the uncertainty of the mean/spread of trends together with the spatial pattern of trends in annual maximum streamflow. We are also curious of whether the regions consistently projected with an increase in flood have been adequately observed by the global observation networks.

2 Data and methods

This section summarizes the workflow to achieve three objectives of this study (Figure 1). Observed and simulated streamflow (section 2.1) were used to estimate the magnitude and significance of changes in an indicator of flood hazards (section 2.3). To enable an observation-model comparison, a procedure was developed to extract streamflow for a subset of observed catchments that meet data quality criteria (section 2.2). A range of statistical techniques were then applied to trends of an indicator of flood magnitude (section 2.4) to assess (i) the capacity of GHMs to reproduce characteristics of observed trends, (ii) the representativeness of observation locations in GHM simulations, and (iii) the implication of simulation uncertainty on projected trends (results are discussed in sections 3.1, 3.2, and 3.3).

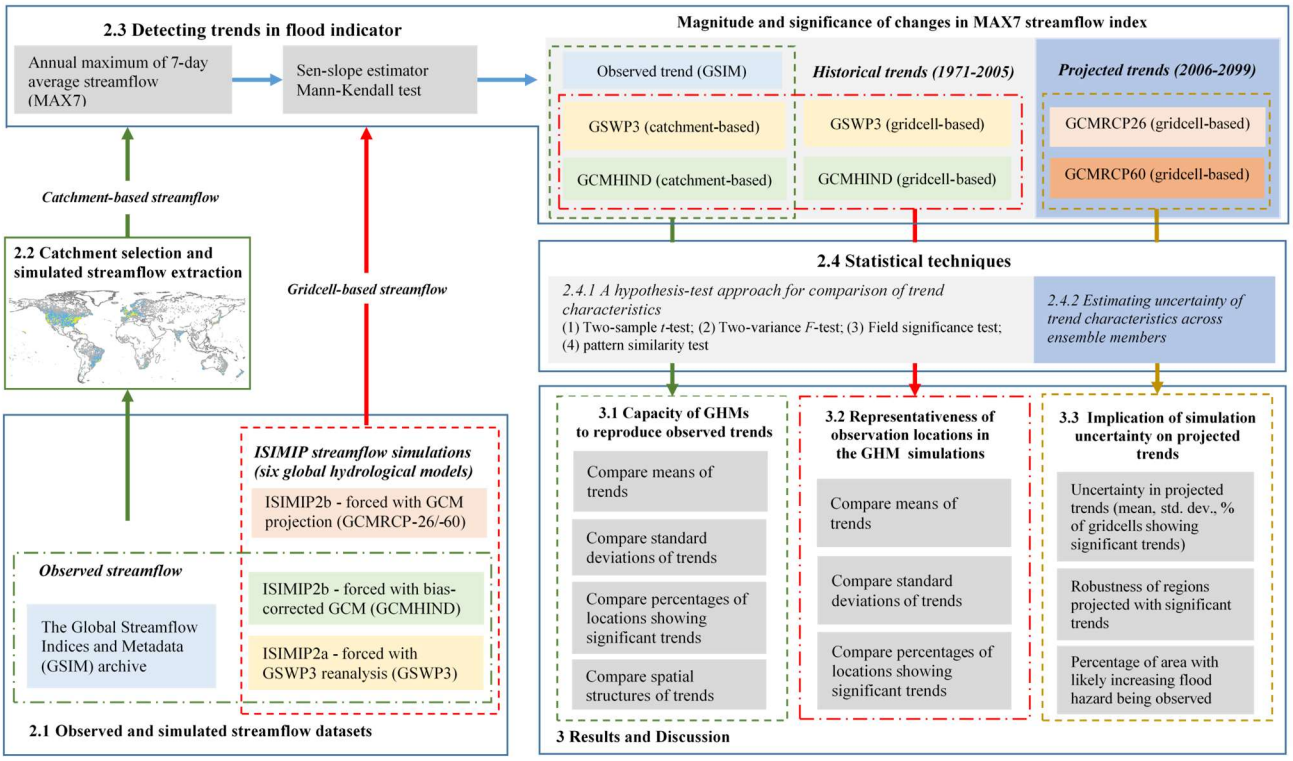


Figure 1. Flowchart of the datasets and methodologies used to achieve three research objectives of this study.

2.1 Observed and simulated streamflow datasets

The GSIM archive is used as daily observational discharge for this analysis. Daily streamflow simulations available through the ISIMIP are used, with historical simulations (~~ISIMIP2a~~forced with observational climate in ISIMIP2a and bias-corrected climate model outputs in ISIMIP2b) spanning from 1971 to 2005 (Gosling et al., 2019) and future simulations (ISIMIP2b) covering 2006-2099 period (Frieler et al., 2017). Six GHMs are considered: H08 ~~(Hanasaki et al., 2008a, b), LPJmL (Schaphoff et al., 2013)~~(Hanasaki et al., 2008b, a), LPJmL (Schaphoff et al., 2013), MPI-HM (Stacke and Hagemann, 2012), ORCHIDEE (Guimberteau et al., 2014;Guimberteau et al., 2018), PCR-GLOBWB (Wada et al., 2014;Sutanudjaja et al., 2018), and ~~WaterGAP~~WaterGAP2 (Müller Schmied et al., 2014;Mueller Schmied et al., 2016). These models were selected as they have provided discharge data within phases 2a and 2b of ISIMIP at the time this study began (June 2018). A summary of the similarities and differences across participated GHMs is provided in supplementary section 1.2.

To assess the model structural uncertainty across GHMs, trends in streamflow extremes simulated under observational atmospheric forcing, available through the Global Soil Wetness Project Phase 3 (GSWP3) reanalysis (Kim, 2017), were compared to observed trends. The influence of the ~~acknowledged~~ high uncertainty in climate models ~~(Kumar et al., 2013;Kiktev et al., 2003)~~(Kumar et al., 2013;Kiktev et al., 2003) on streamflow simulations was assessed by comparing observed trends and trends simulated when using atmospheric forcing from four General Circulation Models (GCMs) for the historical period ('hindcast' simulations~~). These GCM~~; hereafter referred to GCMHIND atmospheric forcing). These GCMs were bias corrected but their simulations have different sub-monthly, inter-annual and decadal variability and thus the hindcast simulations reflect both GHM and GCM uncertainty. To quantify the implication of model uncertainty for future projections of flood hazard, trends simulated under projected climate change by the end of this century (using the same four GCMs) were also assessed~~.~~ for two greenhouse gas concentration scenario RCP2.6 (hereafter referred to GCMRCP2.6 atmospheric forcing) and RCP6.0 (hereafter referred to GCMRCP6.0 atmospheric forcing). As a result, four simulation settings were used in this study, denoted by the atmospheric forcing; an overview is given in Table 1. These settings comprise two historical runs (GSWP3 and GCMHIND runs), and two future runs (GCMRCP2.6 and GCMRCP6.0), collectively amounting to a total of 69 simulations (see Table ~~S2~~S3 in supplementary with full list of simulations).

For GSWP3 simulations, ~~naturalised runs (i.e. human water management not taken into account) were chosen, since this setting is available for more GHMs when compared to the human impact setting (i.e. human water management inputs~~

~~were used). A~~a preliminary analysis (see section 4 of supplementary material) shows that both 'naturalised runs' (i.e. human water management not taken into account) and 'human impact runs' (i.e. human water management inputs were used) exhibit similar characteristic of trends in ~~peak discharge.~~MAX7 index. Some potential reasons for negligible impacts of human water management are the spatial distribution of stream gauges (may be biased toward regions with insignificant changes in water management during the 1971-2005 period), or the inclusion of small catchments (more that 3,000 catchments with reported area less than 9,000 km^2), thus floods are more sensitive to changes in climate forcing relative to the accumulated basin-wide influence of human impacts. Naturalised runs were therefore chosen, since this setting is available for more GHMs (six) when compared to the human impact setting (four). Although significant efforts were made by ISIMIP to keep the setting across simulations as consistent as possible, there were some differences in model versions and input data (e.g~~. WaterGAP~~., WaterGAP2.2 (ISIMIP2a) was used in ISIMIP2a while WaterGAP2.2c was used in ISIMIP2b; ORCHIDEE (Guimberteau et al., 2014) was used in ISIMIP2a while ORCHIDEE-MICT (Guimberteau et al., 2018), with improvements on high latitude processes, was used in ISIMIP2b). ~~As a result, there are~~ Although the influence of versioning is minor for WaterGAP2, the potential effects of technical discrepancies ~~to the findings which~~ cannot be checked in the context of this study~~.~~, as not all required simulations are readily available (see our discussion in supplementary section 3.3). In addition, owing to technical requirements across GHMs, ~~the number of land grid cells with available data is also~~ different models do not have the same set of coastal cells, which may lead to some minor effect to the statistics when averaged across ~~simulations.~~ all simulation grid-cells.

Table 1. Summary of streamflow observation and simulation datasets used in this study. GSIM was used as the observed streamflow database. Streamflow simulations were obtained from six GHMs (H08, LJPmL, MPI-HM, ORCHIDEE, PCR-GLOBWB and ~~WaterGAP~~WaterGAP2). One observational atmospheric forcing dataset (GSWP3) and outputs of four GCMs were used as input for streamflow simulations.

Reference window	Streamflow obs./sim.	No. of GCM-GHM combination	Description	Note
Historical (1971-2005)	GSIM	-	Observational streamflow selected from GSIM archive.	Streamflow daily observations for 3,666 unique locations
	GSWP3	6	Historical simulation forced by observational atmospheric forcing.	Model did not use human water management input.

	(ISIMIP 2a)			
	GCMHIND (ISIMIP 2b)	21	Historical simulation using atmospheric forcing from four GCMs: GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR and MIROC5.	No HadGEM2-ES simulation for MPI-HM.
Projection (2006-2099)	GCMRCP2.6 (ISIMIP 2b)	21	Future simulation forced by projected atmospheric forcing under greenhouse gas concentration scenario RCP2.6. Four GCMs were used: GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR and MIROC5.	No HadGEM2-ES and MIROC5 simulations for ORCHIDEE.
	GCMRCP6.0 (ISIMIP 2b)	21	Future simulation forced by projected atmospheric forcing under greenhouse gas concentration scenario RCP6.0. Four GCMs were used: GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR and MIROC5.	

2.2 ~~Simulated~~Catchment selection and simulated streamflow extraction ~~and catchment selection~~ for observation-model comparison

To enable an observation-model comparison, simulated discharge needs to be extracted from gridded model output. Large-scale hydrological models, however, generally do not simulate discharge accurately over small-to-medium size catchments due to the coarse resolution of river network datasets in their routing schemes ~~(Hunger and Döll, 2008).~~(Hunger and Döll, 2008). To address this limitation, previous GHMs evaluations usually selected large catchments (a threshold of 9,000 km^2 was adopted, approximating the size of a one-degree longitude/latitude grid cell) and routed discharge (units: m^3/s) at the outlet of the catchment was used as simulated streamflow for a specific catchment ~~(Zhao et al., 2017;Veldkamp et al., 2018;Zaherpour et al., 2018;Liu et al., 2017;Zaherpour et al., 2019)~~(Zhao et al., 2017;Veldkamp et al., 2018;Zaherpour et al., 2018;Liu et al., 2017;Zaherpour et al., 2019). For evaluation studies that used relatively small catchments (e.g. area less than 9,000 km^2), the un-routed runoff simulation (units: mm/day) was extracted while observed discharge was converted to runoff using catchment area prior to comparison (Gudmundsson et al., 2012b;Beck et al., 2017). To increase the sample size for the model-observation comparison (the first objective), the present study used both daily (i) un-routed runoff for small catchments and (ii) routed discharge simulations for large

ones, and thus two extraction procedures were adopted. A summary of these extraction procedures is provided below while detailed technical descriptions are provided in section 2 of supplementary material.

- For catchments with area from 0 to 9,000 km^2: un-routed runoff (mm/day) was extracted and then converted into discharge (m^3/s) by multiplying averaged runoff with catchment area~~.~~ reported in station metadata. Specifically, catchment boundaries were superimposed on the GHM grid to obtain the weighted-area tables, which were then used to derive averaged runoff from the un-routed runoff simulation. To avoid double counting runoff from the same grid points, runoff for catchments that share similar weighted-area tables (i.e. similar simulated streamflow would be extracted – see supplementary section 2 for detail description) was averaged (using catchment areas as weights) and a single 'averaged time series' was used in place of the runoff from the component catchments.
- For catchments with area greater than 9,000 km^2: the 'discharge output' approach ~~(Zhao et al., 2017)~~(Zhao et al., 2017) was adopted to extract routed discharge (m^3/s) from the GHM cell corresponding to the outlet of each catchment.

To ensure sufficient data is available for historical trend analysis, only GSIM stations with at least 30 years of data available during the 1971-2005 period were considered (each year having at least 335 days of available records~~).~~, implying that annual maximum of a specific year is identified only when more than 90% of daily record is available). These relatively strict selection criteria also enable a comparison between this study and preceding observation-based investigations (Gudmundsson et al., 2019;Hodgkins et al., 2017). As catchment boundary shapefiles ~~(Do et al., 2018b)~~(Do et al., 2018a) were used to extract simulated streamflow for small catchments, stations were further filtered using two criteria: (i) availability of reported catchment area, and (ii) catchment boundary was accompanied with a "high" or "medium" quality flag (i.e. the discrepancy between reported and estimated catchment area is less than 10%).

A total of 4,595 stations satisfied the quality selection criteria, of which large catchments (i.e. area greater than 9,000 km^2) where no suitable grid cell could be identified were further removed (11 catchments). For cases of two or more small catchments (i.e. area less than or equal to 9,000km^2) having similar weighted-area tables, the 'averaged time series' (using catchment areas as weights) was calculated. A total number of 1,542 time series fell in this category and were aggregated into 624 'averaged time series'. Figure ~~1~~2 shows the spatial distribution of the final dataset for model-observation comparison, containing data for 3,666 locations (3,042 non-averaged time series and 624 averaged time series). The majority of available catchments are located in North America and Europe, with some regions over Asia, Oceania and South America are also covered.

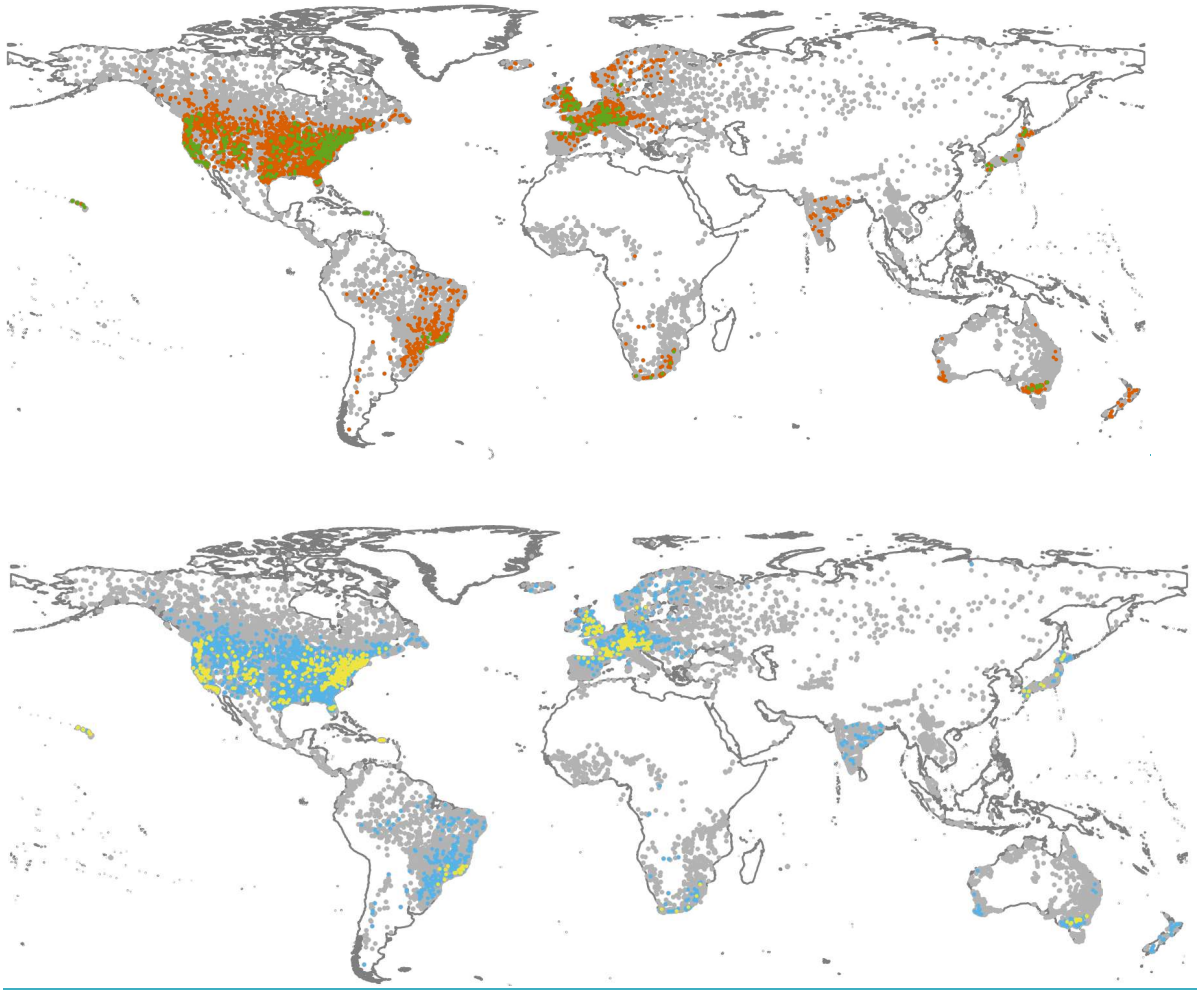


Figure ~~1~~2. Locations of 3,666 streamflow observations (~~brown~~blue dots: 3,024 non-averaged time series; ~~green~~yellow dots: 624 averaged time series, where geographical coordinates were averaged from all component gauging coordinates) selected from GSIM archive for the model-observation comparison. Grey dots indicate GSIM time series that were removed due to insufficient data availability or quality.

2.3 Detecting trends in annual maximum streamflow

For each streamflow dataset, daily discharge was smoothed to 7-day averages to reduce variability in simulated streamflow, which can arise from the coarse routing parameters of GHMs ~~(Dankers et al., 2014)~~(Dankers et al., 2014). The annual maximum time series of 7-day averaged discharge (labelled as the MAX7 index in the GSIM archive) was then derived to represent peak flow events. For gridded datasets, the 'centre averaged approach' (e.g. averaged streamflow of Jan 7th is the mean value of Jan 4 – 10th) was used (the common setting of the CDO software, freely available at https://code.mpimet.mpg.de/projects/cdo), and the MAX7 timeseries was therefore derived for each GSIM station using this same approach. As a result, the derived value of the MAX7 index is slightly different to the value

available in the online version of GSIM ~~(Gudmundsson et al., 2018a),~~ which applied a 'backward-moving average' technique (e.g. averaged streamflow of Jan 7th is the mean value of Jan 1 – 7th). Our preliminary analysis (not shown), however, indicated that this difference did not lead to substantial changes in the key findings~~.~~ (i.e., similar spatial composition between increasing and decreasing trends).

The magnitude of trends in the MAX7 index at a specific catchment or grid cell was quantified using the normalised Theil-Sen slope ~~(Gudmundsson et al., 2019;Stahl et al., 2010)~~(Gudmundsson et al., 2019;Stahl et al., 2010) and the results are expressed in % change per decade. The significance of the local trend was assessed using a Mann-Kendall test at the 10% two-sided significance level (Wilks, 2011). The null hypothesis (no trend) is rejected if the two-sided *p*-value of the test statistic (Kendall's τ) is lower than 0.1, while the direction of the trend (i.e. increasing or decreasing) was determined using the sign of τ.

2.4 Statistical techniques

~~To address the three identified objectives~~To explore GHMs' capacity to simulate observed trends and the implication of model uncertainty to projected trends, trends in streamflow extremes obtained from GSIM (observed trends) and ISIMIP simulations (simulated trends) are analysed. The observed trends were available for 3,666 observation locations. Simulated trends were available for all 59,033 GHM grid cells (estimated from routed discharge of each grid cell; Antarctica and Greenland were removed). To enable a model-observation comparison, we also extract a subset of simulated trends over the 3,666 observation locations (described in section 2.2).

2.4.1 A hypothesis-test approach for comparison of trend characteristics

A range of hypothesis tests (summarised in Table 2; GSWP3 simulations were used to assess GHM uncertainty while GCMHIND simulations were used to assess the combined GCM-GHM uncertainty) was applied to address the first two objectives, which require comparing trend characteristics exhibited from different streamflow datasets. Four characteristics of trends were assessed:

- Trend mean: The mean (% change per decade) of trends in streamflow extremes across all gauge-/cell-based time series over a spatial domain. A hypothesis test was adopted to assess whether the trend means exhibited from two specific streamflow datasets (e.g. model vs. observed) are significantly different from each other.
- Trend standard deviation: The standard deviation (% change per decade) of trends in streamflow extremes across all gauge-/cell-based time series over a spatial domain. A hypothesis test was adopted to assess whether the trend

~~of~~ standard deviations exhibited from two specific streamflow datasets are significantly different from each other.

- Percentage of significant trends (%): The percentage of trends in a domain that are statistically significant, with gauge- or cell-based significance calculated using the Mann-Kendall test at the 10% significance level. To assess whether the percentage of significant (increasing/decreasing) trends exhibited from a specific streamflow dataset is produced by random chance, a field significance test ~~(Do et al., 2017)~~(Do et al., 2017) was adopted~~.~~ (described in Table 2).
- Trend spatial pattern: The spatial distribution of trends in streamflow extremes over a spatial domain. Pearson's ~~(spatial) correlation between trends of two datasets was used as a measure of similarity in the trend spatial structure.~~correlation (*r* statistic) (Galton, 1886;Kiktev et al., 2003) between trends of MAX7 index obtained from two datasets was used as a measure of similarity in the trend spatial structure. The hypothesis test (pattern similarity test) was adopted to assess whether: (i) the correlation between simulated trends introduced by GHMs and observed trends is significantly higher than zero; and (ii) the correlation between trends simulated under hindcast atmospheric forcing and observed trends is significantly lower than that between trends simulated under observational atmospheric forcing and observed trends.

294 **Table 2.** Hypothesis tests conducted to address the first two objectives.

Objective	Null-Hypotheses	Streamflow dataset	Statistical tests
Objective 1: Capacity of GHMs to reproduce observed trends in flood hazards	Hypothesis 1: Trend means obtained from two streamflow datasets over observation locations were not statistically different from each other.	(i) Observed discharge across 3,666 observation locations (ii) Simulated discharge across 3,666 observation locations (extraction processes outlined in Section 2.2)	Two-sample *t*-test at the 10% two-sided significance level
	Hypothesis 2: Trend standard deviations obtained from two streamflow datasets over observation locations were not statistically different from each other.		Two-variance *F*-test at the 10% two-sided significance level
	Hypothesis 3: Percentage of significant trends obtained from all observation locations of a specific streamflow dataset was not produced by random chance.		Field significance test similar to that presented in ~~Do et al. (2017)~~Do et al. (2017) was adopted. A moving-block-bootstrap (block-length $L = 2$) was used to derive a null-hypothesis distribution of the change that occurred due to random chance. The null hypothesis is rejected at 5% one-sided significance level when the true percentage falls on the right-hand side of the 95th percentile of the resampled distributions.
	Hypothesis 4: The correlation between trends obtained from two streamflow datasets was not significantly higher than '0' (i.e. zero pattern similarity).		'Zero pattern similarity' was compared to the probability distribution function (PDF) of pairwise correlation between simulated and observed trends, drawn from a bootstrap procedure similar to that proposed by Kiktev et al. (2003). The null hypothesis is rejected at 5% one-sided significance level when zero correlation falls on the left-hand side of the 5th percentile of the resampled distributions.

	Hypothesis 5: The correlation between GCMHIND simulated trends and observed trends was not significantly lower than the correlation between GSWP3 simulated trends and observed trends		The actual pairwise correlation between GCMHIND simulated trends and observed trends (denoted by $r_{GCMHIND}$) was compared to the bootstrapped PDF of correlation exhibited from GSWP3 simulated trends (denoted by r^*_{GSW}). If $r_{GCMHIND}$ falls on the left-hand side of the 5th percentile r^*_{GSWP} , there is evidence to reject the null-hypothesis at the 5% one-sided significance level.
Objective 2: The representativeness of observation locations in the GHM simulations	Hypothesis 6: Trend mean obtained from observation locations was not statistically different to that obtained from all grid cells.	(i) Simulated discharge across 3,666 observation locations (extraction processes outlined in Section 2.2)	Two-sample *t*-test at the 10% two-sided significance level
	Hypothesis 7: Trend standard deviation obtained from observation locations was not statistically different to that obtained from all grid cells.		Two-variance *F*-test at the 10% two-sided significance level
	Hypothesis 8: Percentage of significant trends obtained from all grid cells of a specific streamflow dataset was not produced by random chance.	(ii) Routed discharge across all landmass grid cells (59,033 cells)	Field significance test similar to that presented in Hypothesis 3 but trends obtained from all grid cells were the subject of the assessment.

295

2.4.2 Estimating uncertainty of trend characteristics across ensemble members

The third and final objective, which focused on the implications of GCM-GHM uncertainty on projected changes in flood hazard, was addressed by quantifying the spread of trend characteristics (i.e. trend mean, trend standard deviation, and percentage of significant trends) exhibited from routed discharge projections under two representative concentration pathways.

The spatial uncertainty of projected trends (GCMRCP2.6 and GCMRCP6.0) was also quantified by calculating intra-/inter-model correlation of the trend patterns across all ensemble members available under the two projections. Intra-model correlation represents spatial uncertainty introduced by the GCM and was calculated from simulated trends introduced by the same GHM (using different simulated atmospheric forcing). Inter-model correlation represents the combined GCM-GHM spatial uncertainty, and was calculated for each pair of simulated trends that were: (i) introduced by the different GHMs; and (ii) forced with different projected atmospheric forcing. ~~This assessment also identified regions that were consistently detected with a significant increasing trend across at least 11 simulations, which can be used as an indication of potential 'hot-spots' of future flood hazard.~~

To assess the robustness of GHMs in projecting changes in flood hazard, each grid-cell ~~of~~available in the discharge simulation grid was then categorised into one of the five 'flood-risk' (here "flood-risk" level is defined as the number of ensemble members projecting significant increasing trends) groups based on the number of GCMRCP2.6/GCMRCP6.0 simulation members projecting a significant increasing trend (Group 1: no members, Group 2: from 1 to 5 members, Group 3: from 6 to 10 members, Group 4: from 11 to 15 members and Group 5: from 16 to 18 members). ~~Each GSIM gauge was also~~

Finally, to assess whether locations projected with an increasing trend by the majority simulations are adequately monitored, each GSIM gauge was allocated into one of these five groups based on the gauge's geographical coordinates. The allocation of gauges into these groups was then analysed to determine whether the most comprehensive global database of daily streamflow records to-date was evenly distributed across the five 'flood risk regions'. An inadequately coverage of stream-gauge networks over high risk regions indicate potentially high vulnerability to future changes in flood hazards, as insufficient data is available to inform decision makers.

3 Results and Discussion

3.1 Capacity of GHMs to reproduce observed trends in flood hazards

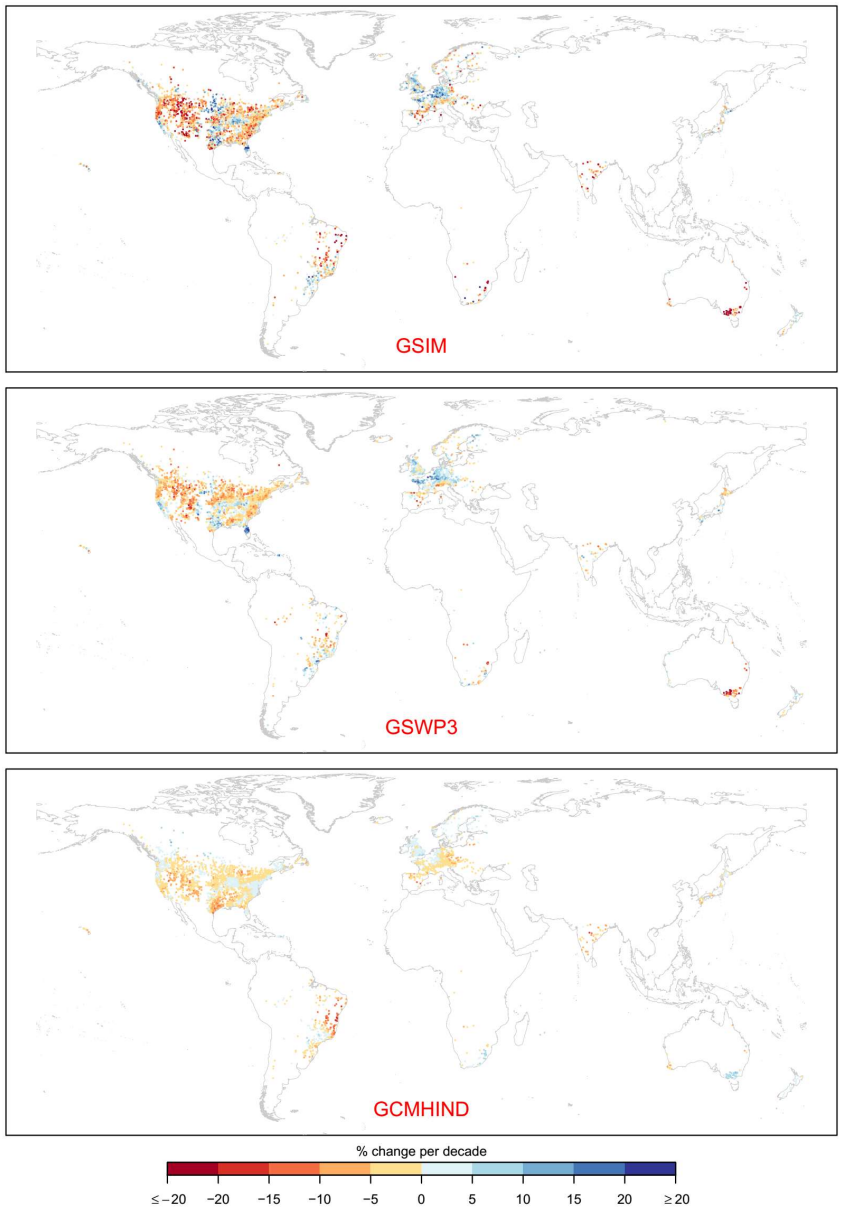
Visual inspection of the normalised Theil-Sen slope across the GSIM time series (top panel of Figure ~~2~~3; regional maps provided in Supplementary Figure S4) shows a spatial pattern that is consistent with recent findings on trends in

observed flood magnitude ~~(Mangini et al., 2018;Do et al., 2017;Mallakpour and Villarini, 2015;Gudmundsson et al., 2019;Burn and Whitfield, 2018;Ishak et al., 2013).~~(Mangini et al., 2018;Do et al., 2017;Mallakpour and Villarini, 2015;Gudmundsson et al., 2019;Burn and Whitfield, 2018;Ishak et al., 2013). Specifically, decreasing trends tend to dominate Asia (most stations located in Japan and India), Australia, the Mediterranean, western/north-eastern US and northern Brazil, while increasing trends appear mostly over central North America, southern Brazil and the northern part of Western Europe (including the UK). Note that the observation locations are not evenly distributed (86% in North America and Europe), and thus the confidence of this assessment varies substantially across continents.

The multi-model average of GSWP3 simulated trends (trends simulated under observational atmospheric forcing; middle ~~panel~~panels of Figure ~~2~~3) has generally good capacity to reproduce spatial patterns of observed trends. The multi-model average of GCMHIND simulated trends (trends simulated under hindcast atmospheric forcing; lower ~~panel~~panels of Figure ~~2~~3), however, could not reproduce some spatial agglomerations of trends in streamflow maxima (e.g. the decreasing trends in south-eastern Australia, increasing trends over north-eastern Europe). This feature indicates the inconsistent climate variability between GCMs and the real world, suggesting GCM climate forcing cannot account for observed trends at sub-continental scale. In addition, GCMs uncertainty can potentially contribute to this inconsitency. Interestingly, the multi-model average of both GSWP3 and GCMHIND simulations generally exhibits a lower magnitude of changes (i.e. closer to 'zero change') compared to the observed trends. This feature is more prominent in GCMHIND (21 simulations available) compared to GSWP3 (six simulations available), and can be explained by two possibilities. The first possible explanation is the nature of averaging, which tends to smooth out variability in trend magnitude across ensemble members, leading to a relatively 'close to zero' change across the globe (given that each GCMs has stochastic decadal climate variability, so that averaging results forced by GCMs tends to cancel trends). An alternative explanation is that individual simulations also exhibit a lower magnitude of change relative to observation~~, which is not visible through Figure 2.~~

~~To further explore GHMs' performance~~. As Figure 3 is not sufficient to evaluate the latter possibility, a more detailed comparative analysis between observed trends and individual simulated trends using both historical climate forcings (via GSWP3) and GCM hindcasts was conducted. Specifically, four characteristics of trends in extreme flows (i.e. trend mean, trend standard deviation, percentage of significant trends and trend spatial structure) were assessed for individual simulations and the results are reported in following sections. At the global scale, GSIM observed trends exhibit a mean and standard deviation of -2.4% and 9.9% change per decade over the 1971-2005 historical period. Furthermore, there are 7.5% (12.1%) stations showing significant increasing (decreasing) trends (detected by the

Mann-Kendall test at the 10% significance level). These numbers, however, are not statistically significant at the global scale.



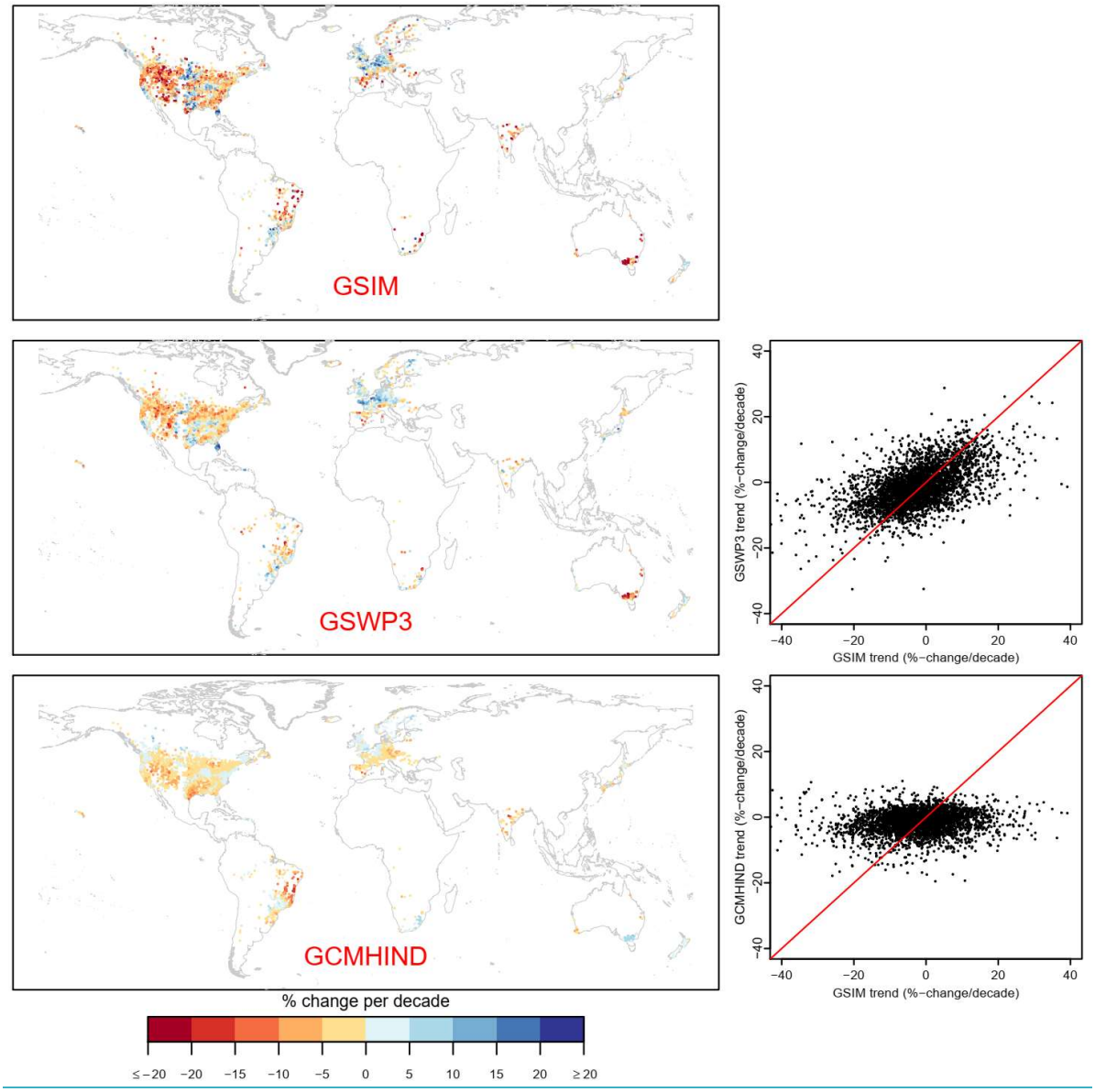


Figure 23. Normalised Theil-Sen slope for historical trends in flood magnitude (MAX7 index) exhibited over 3,666 locations across three streamflow datasets (top left: GSIM; middle left: GSWP3; bottom left: GCMHIND). Multi-model average is shown for simulated trends. Trend is expressed in % change per decade. Scatter plot between trends obtained from GSIM and GSWP3/GCMHIND simulated streamflow are provided in the right panels.

Table 3 shows the results of the global model-observation comparison using GSWP3 simulated trends across the six GHMs. Compared to observed trends, most simulated trends have a significantly higher global trend mean at the observed locations ~~(ranging from -2.2% to 0.1% change per decade)~~ and lower trend standard deviation ~~(ranging from 7.1% to 8.7% change per decade).~~. The percentage of locations showing significant trends varies substantially across simulations, but the values were not statistically significant. All GHMs demonstrate low-to-moderate capacity in

simulating the spatial pattern of trends (spatial correlation coefficients range from 0.35 to 0.50, indicating that GSWP3 simulated trends account for between 12%-25% of the cross-location variability in the observed trend signal). There is, however, a notable difference in terms of the overall sign of trends simulated by each ~~different~~ GHM. This feature indicates that using different GHMs can lead to different interpretations about the overall change in flood hazard at the global scale, despite having a common boundary forcing. ~~For example, PCR-GLOBWB suggests there are more locations showing significant increasing trends (9.6%) than decreasing trends (6.1%) while LPJmL shows the opposite pattern (4.5% and 7.3% of locations showing significant increasing and decreasing trends respectively). The variation of trends characteristics exhibited by different GHMs also indicates that~~Therefore, the 'closer to zero' trends of ensemble averages (illustrated in Figure ~~2~~3) likely reflects the implication of averaging rather than a systematic bias of GHMs toward a low magnitude of change. As an implication, ensemble averages even though useful, should not be used as a sole ground to infer ~~change~~changes in floods, as ~~this~~it may undermine the actual magnitude of simulated trends. As a result, the following analyses will report the full range (and mean) of each trend characteristic estimated across all ensemble members to communicate the uncertainty underlying the results.

Table 3. Characteristics of trends in the MAX7 index over the 1971-2005 period across 3,666 locations for GSIM observed trends and GSWP3 simulated trends (six GHMs available). Trend mean and trend standard deviation are expressed in % change per decade. Correlation was obtained from GSIM observed trends and GSWP3 simulated trends for each GHM. Boldface texts represent values that reject the null-hypotheses outlined in Table 2 (hypothesis 1 to 4).

GHM	Trend mean	Trend stand. dev.	% of sig. inc. trends	% of sig. dec. trends	Corr. obs. trend
H08	**-1.9**	**8.3**	4.8	6.7	**0.42**
LPJmL	-2.2	**7.1**	4.5	7.3	**0.37**
PCR-GLOBWB	**0.1**	**7.7**	9.6	6.1	**0.46**
WaterGAP2	**-0.3**	**8.2**	8.5	4.2	**0.49**
MPI-HM	-2.1	**8.7**	5.6	7.5	**0.50**
ORCHIDEE	**-1.4**	**8.6**	7	8.2	**0.35**
GSIM (observation)	-2.4	9.9	7.5	12.1	-

Table 4 provides the results of the model-observation comparison using GCMHIND simulated trends (intra-model averages are shown while results of individual simulations are reported in section 4 of supplementary material). Similar to GSWP3 trends, intra-model averages (i.e. calculated from simulations of one GHM) of GCMHIND trends tend to have a higher global mean ~~(ranging from -2.3% to -0.4% change per decade with 19 out of 21 simulations suggesting a significantly different trend mean)~~ and lower trend standard deviation ~~(ranging from 7.4% to 8.7%~~

~~change per decade, with all simulations suggesting a significantly different trend standard deviation)~~ than observed. The composition between the percentages of locations showing significant trends varies substantially across simulations ~~(ranging from 2.2%/4.1% to 12.2%/17.3% for significant increasing/decreasing trends)~~ and statistical significance was found only for decreasing trends over three out of 21 simulations (two LPJmL simulations and one MPI-HM simulation). The multi-model ranges encapsulate the observed trend mean and percentage of significant trends, while the observed trend standard deviation is clearly above the range exhibited from all GCMHIND simulations. The significantly lower simulated trend standard deviation can be partially attributable to the coarse resolution of GHMs' atmospheric/land surface inputs, which may not sufficiently reflect the variation of hydrological processes across small-to-medium size catchments.

Table 4. Characteristics of trends in the MAX7 index over the 1971-2005 period across 3,666 locations for GCMHIND simulated trends. Trend mean and trend standard deviation are expressed in % change per decade. Intra-model averages of trend characteristics are shown for each GHM. Values in the parentheses show the number of simulations rejecting the null hypothesis (from 1 to 4) outlined in Table 2 (out of four GCMs). Multi-model min/max/average values together with those exhibited from GSIM are also provided.

GHM	Trend mean	Trend stand. dev.	% of sig. inc. trends	% of sig. dec. trends	Corr. obs. trend
H08	-1.7 (4)	8.5 (4)	4.9 (0)	8.8 (0)	0.03 (2)
LPJmL	-2.3 (4)	7.9 (4)	4.2 (0)	12.6 (2)	0.09 (3)
PCR-GLOBWB	-1.1 (2)	7.4 (4)	7.5 (0)	9.4 (0)	0.06 (3)
WaterGAP2	-1.3 (4)	8.4 (4)	5.4 (0)	8.0 (0)	0.02 (2)
MPI-HM	-1.8 (3)	8.7 (3)	5.7 (0)	9.9 (1)	0.05 (2)
ORCHIDEE	-0.4 (2)	8.6 (2)	6.9 (0)	7.0 (0)	0.04 (1)
Multi-model min	-4.2	7.0	2.2	4.1	-0.06
Multi-model max	0.6	9.5	12.2	17.3	0.18
Multi-model average	-1.5	8.2	5.6	9.5	0.05
GSIM (observation)	-2.4	9.9	7.5	12.1	-

Among 21 GCMHIND simulations, the 'zero similarity' hypothesis (hypothesis 5) was rejected over 13 simulations, indicating that GCM-GHM ensemble members possess some capacity to simulate the spatial structure of observed trends in streamflow extremes. The correlation between GCMHIND simulated trends and GSIM observed trends ~~(ranging from -0.06 to 0.18),~~ however, is significantly lower than that exhibited from GSWP3 simulated trends across all GHMs (reported at Table 3). The results of the similarity assessment are illustrated for a single GHM (H08; as the results were similar for other GHMs) in Figure ~~3~~4, where the correlation between observed trends and GSWP3 simulated trends is significantly different from zero. In contrast, the correlation between observed trends and each of

the simulated trends under hindcast atmospheric forcing (GCMHIND simulations) is much lower, with two of the four not being statistically higher than zero. These results confirm the substantial influence of atmospheric forcing on the simulated trend pattern relative to GHMs structure.

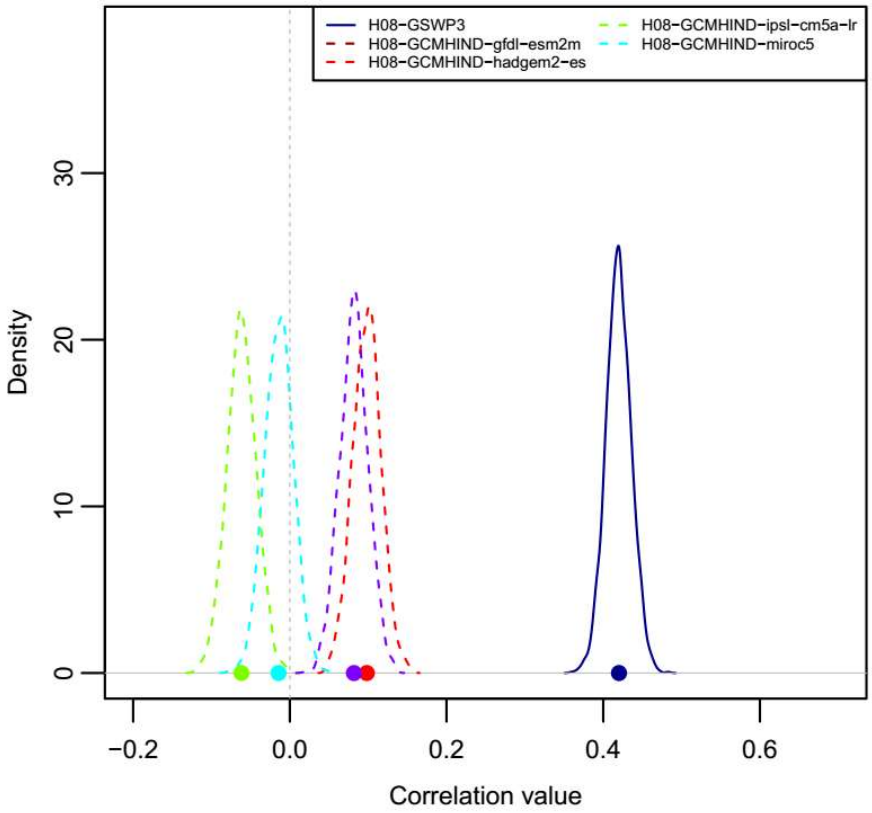


Figure ~~3~~4. Model-observation correlation between observed trends and simulated trends across all simulations (GSWP3 and four GCMHIND simulations) of a single model (H08; similar results for other GHMs). Coloured dots indicate actual correlation between a specific simulated trend pattern and observed trend pattern across 3,666 locations. Colour lines represent the PDFs of correlation between simulated trend pattern and observed trend pattern obtained through a bootstrap resampling procedure (B = 2000).

To further quantify changes at the regional scale, a model-observation comparison (identical to that at the global scale) was conducted over six continents and the results are summarised in Table 5 (multi-model averages are shown). The trend mean exhibited from GSIM ranges from -10.7% (Oceania) to 2.4% change per decade (Europe) while trend standard deviation ranges from 8.3% (Europe) to 15.8% change per decade (Oceania). The percentage of significant increasing (decreasing) trends exhibited from GSIM ranges from 3.2% to 22.6% (from 6.3% to 29.1%) and the composition of significant trends across the six continents is consistent to a previous investigation ~~(Do et al., 2017)~~(Do et al., 2017). The observed percentage of significant trends is found to be above random chance for Europe

(increasing flood magnitude) and Australia (decreasing flood magnitude) and this feature is captured quite well by GSWP3 simulated trends, with at least half of the simulations confirming field significances detected from GSIM.

~~Similar to the assessment at the global scale, most GSWP3 simulations generally exhibit a higher trend mean compared to the observed trend at the continental scale (see also Section 3.1 of the supplementary). Over data-covered regions, a general lower trend standard deviation was also exhibited across all simulations~~Trend characteristics simulated by GHMs at continental scale confirms some important findings from global scale assessments, suggesting substantial uncertainty of trends in streamflow extremes introduced by GHMs at the continental scale~~. The~~:

- Both GSWP3 and GCMHIND simulations generally exhibit a higher trend mean and lower trend standard deviation compared to the observed trend at the continental scale (see also Section 3.1 of the supplementary).
- GCMHIND simulations generally exhibit lower capacity to reproduce trend characteristics relative to GSWP3 simulations due to the combined GHM-GCM uncertainty.

For GSWP3 simulations, the spatial correlation is weakest in Asia, as no simulation rejects the null-hypothesis of 'zero similarity', while the spatial correlation is strongest in Oceania (mainly southern Australia; correlation of 0.63). Oceania, however, exhibits the highest model-observation discrepancy in trend mean and trend standard deviation, indicating the capacity of a given GHM in terms of the trend spatial structure is not necessarily consistent with its performance in terms of the mean and spread of trends.

~~GCMHIND simulations generally exhibit lower capacity in terms of reproducing trends. The majority of GCMHIND simulated trends tends to not capture the continental trend mean and trend standard deviation exhibited in the observed (see also Section 3.1 of the supplementary).~~ GCMHIND trends also suggest the opposite composition between percentages of significant trends compared to GSIM trends (e.g. simulated trends suggest more locations showing significant increasing trends while observed trends suggest the opposite). ~~Finally, the spatial correlation is also significantly lower than GSWP3 correlation (except for Asia and South America).~~ Among six continents, GCMHIND trends exhibited the lowest correlation (-0.14) in Oceania, whereas GSWP3 suggested the strongest correlation in this continent. This assessment further indicates the substantial impact of atmospheric forcing relative to GHM model structure on the simulated trends in high flow events. It is informative to note that this result is expected, as GCMs (although have been bias-corrected) generally have low capacity in reproducing the timing of wet/dry periods or the spatial distribution of climate extremes (Kiktev et al., 2007), and GHMs are likely to inherit these limitations when using GCMs' outputs as climate forcing data.

Table 5. Characteristics of trends exhibited from GSIM/GSWP3/GCMHIND streamflow dataset at the continental scale (each observation location of 3,666 sites was allocated into one of the six continents). For simulated trends, only the multi-model average is shown for each region. Trend mean and trend standard deviation are expressed in % change per decade. Values in the parentheses show the number of simulations rejecting the null-hypothesis described in Table 2 (up to six for GSWP3 simulations and 21 for GCMHIND simulations). For GSIM, field significance of increasing/decreasing trends was highlighted by boldface texts.

	No. of	**Trend mean**			**Trend Stand. Dev.**			**% sig. inc. trends**			**% sig. dec. trends**			**Corr. obs. trends**	
Region	**loc.**	*GSIM*	*GSWP3*	*GCMHIND*	*GSIM*	*GSWP3*	*GCMHIND*	*GSIM*	*GSWP3*	*GCMHIND*	*GSIM*	*GSWP3*	*GCMHIND*	*GSWP3*	*GCMHIND*
Asia	96	-3.1	-1.2 (4)	-2.7 (6)	8.8	6.6 (5)	7.2 (15)	4.2	4.2 (0)	2.2 (0)	15.6	10.3 (1)	9.7 (2)	0.07 (0)	0.11 (11)
N. America	2441	-3.5	-2.4 (3)	-1.6 (18)	9.4	7.9 (6)	8.0 (19)	3.2	2.8 (0)	5.3 (0)	13.4	7.5 (0)	9.3 (3)	0.38 (6)	0.03 (12)
Europe	730	2.4	2.6 (6)	-0.7 (17)	8.3	7.1 (5)	5.9 (21)	**22.6**	20.2 (3)	7.3 (1)	6.3	2.1 (0)	10.1 (4)	0.43 (6)	0.10 (13)
Africa	48	-2.5	-1.3 (0)	1.5 (12)	14.8	9.8 (5)	8.0 (20)	6.3	2.8 (0)	9.6 (2)	10.4	10.4 (0)	3.3 (0)	0.46 (6)	0.07 (6)
S. America	265	-2.0	-0.2 (5)	-3.6 (14)	10.1	7.6 (6)	10.0 (20)	7.9	7.2 (0)	3.4 (1)	10.2	4.4 (0)	13.4 (5)	0.26 (6)	0.18 (17)
Oceania	86	-10.7	-6.1 (4)	2.4 (21)	15.8	10.9 (6)	8.4 (21)	4.7	3.7 (0)	11 (2)	**29.1**	22.1 (4)	1.9 (0)	0.63 (6)	-0.14 (2)

3.2 Determining the representativeness of observation locations in the GHM simulations

To assess the representativeness of observations locations in GHM grid cells, trend characteristics obtained from all simulated grid cells were compared to those estimated from the observation locations (3,666 sites globally). For GSWP3 simulations, the results suggest a significant difference between trend characteristics from all model grid cells compared to those obtained from the observation locations (Table 6; multi-model averages shown). This feature is consistent at both global and continental scales, including North America and Europe – the continents with the best stream-gauge density. Specifically, the trend mean tends to get closer to zero, while the trend standard deviation obtained from all grid cells tends to be higher than that over observation locations. The difference between the percentages of significant increasing/decreasing trends across all grid cells also gets smaller. For instance, the percentage of observation locations showing significant increasing (decreasing) trends over Oceania is 3.7% (22.1%) for GSWP3 multi-model averages (reported in Table 5), while the corresponding values are 10.7% (15.1%) when all grid cells are considered (reported in Table 6). Additionally, field significance for increasing (decreasing) trends is detected in two (four) out of six simulations over Oceania, while the same feature could not be detected over the observation locations. These findings confirm that trends exhibited from observation locations are not a representative sample of trends obtained from all simulation grid cells, which has also been suggested through Figure ~~1~~2. As a result, a common model-observation picture of changes in global flood hazard remains elusive. To enable a holistic perspective of changes in extreme flows, it is therefore crucial to improve not only models' capacity, but also data accessibility and expand streamflow observational networks to ensure unbiased samples are available for large scale investigations.

The findings using GCMHIND simulations are similar in terms of the trend mean (closer to zero) and trend standard deviation (higher) across all grid cells relative to the observation locations. Across all land areas, the composition of the percentages of land mass showing significant trends exhibited by GCMHIND simulations contradicts that obtained from the GSWP3 simulations for many continents. For example, GSWP3 simulations suggest more land areas showing significant decreasing trends than increasing trends over Asia and Oceania while GCMHIND simulations indicate an overall increasing change in extreme flows over the same continents. This feature further confirms the importance of uncertainty in atmospheric forcing in driving the spatial structure of the simulated trends, which will be explored further in the next section.

Table 6. Characteristics of simulated trends across all grid cells at both continental and global scales (multi-model averages are showed). For each simulation, cell-based trend mean/trend standard deviation was compared to that of gauge-based trends (reported in Table 4). Values in parentheses represent the number of simulations reject the null-hypothesis described in Table 2 (up to six simulations for GSWP3 and 21 simulations for GCMHIND). GSIM results are also provided for reference.

	Trend mean			**Trend Stand. Dev.**			**% sig. inc. trends**			**% sig. dec. trends**		
Region	*GSIM*	*GSWP3*	*GCMHIND*	*GSIM*	*GSWP3*	*GCMHIND*	*GSIM*	*GSWP3*	*GCMHIND*	*GSIM*	*GSWP3*	*GCMHIND*
Asia	-3.1	-0.7 (3)	0.4 (16)	8.8	10.3 (6)	9.0 (15)	4.2	7.7 (0)	9.6 (7)	15.6	9.9 (3)	7.7 (4)
N. America	-3.5	-1.8 (4)	0.4 (19)	9.4	10.3 (6)	8.3 (17)	3.2	6.9 (1)	8.2 (4)	13.4	12.3 (5)	6.6 (0)
Europe	2.4	1.1 (5)	0.2 (16)	8.3	8.5 (5)	8.4 (20)	**22.6**	11.5 (2)	9.1 (5)	6.3	4.5 (0)	7.9 (3)
Africa	-2.5	0.7 (2)	-1.7 (15)	14.8	11.0 (3)	10.1 (12)	6.3	10.9 (1)	8.5 (6)	10.4	11.2 (2)	15.5 (11)
S. America	-2.0	-2.0 (6)	-0.7 (19)	10.1	8.7 (3)	9.1 (17)	7.9	4.9 (0)	5.0 (0)	10.2	8.6 (0)	8.2 (1)
Oceania	-10.7	-1.0 (6)	0.5 (17)	15.8	11.3 (4)	10.4 (17)	4.7	10.7 (0)	10.3 (3)	**29.1**	15.1 (1)	9.6 (6)
Global	-2.4	-0.6 (6)	-0.1 (20)	9.9	10.3 (6)	9.4 (19)	7.5	8.3 (1)	8.6 (6)	12.1	10.2 (4)	9.0 (6)

3.3 The implication of simulation uncertainty on the projection of trends in flood hazard

This section focuses on the uncertainty in simulated trends under projected climate forcing at the global scale. For MPI-HM (no simulation for HadGEM2-ES forcing), streamflow was only simulated across the main stream-network (approximately 45% of the global land grid cells), and thus three simulations of this GHM were removed from the analysis. As a result, only 18 ensemble members were used to explore the uncertainty in projected trends (GCMRCP2.6 and GCMRCP6.0 – trends estimated for the 2006-2099 period and all cells were considered).

Table 7 shows a relatively low spread of the global trend mean (ranging from -1.3% to 0.8% change per decade; multi-model average of 0.0% change per decade for both GCMRCP2.6 and GCMRCP6.0) and trend standard deviation (ranging from 1.8% to 4.1% change per decade) across ensemble members. LPJmL and ORCHIDEE generally suggest a decreasing trend at the global scale, evident through the negative global mean and more grid cells showing significant decreasing trends. The standard deviation of trends in future simulations ~~(multi-model average of 2.3% and 3.2% change per decade for GCMRCP2.6 and GCMRCP6.0 respectively)~~ is substantially lower than the historical run (~~multi-model average of 9.4% change per decade as~~ reported in Table 6). This feature is potentially due to the capacity of longer time series in capturing the inter-decadal variability of the streamflow regimes, with both dry and wet periods being considered ~~(Hall et al., 2014)~~(Hall et al., 2014). Projected trends under the RCP2.6 scenario generally have closer to zero mean and lower standard deviation compared to those introduced by the RCP6.0 scenario, reflecting the nature of an ambitious 'low-end warming' scenario, when anthropogenic climate change reaches its peak at the middle of the 21st century followed by a generally stable condition.

Interestingly, although most models suggest relatively moderate changes in the global trend mean, the composition between percentages of grid cells showing significant trends varies substantially, ranging from 7.5% (7.1%) to 30.1% (35.0%) for significant increasing (decreasing) trends at the 10% level, with RCP6.0 generally exhibits higher values. This finding indicates that inferences of changes focusing on global averages may mask significant regional trends, as there was a substantially high percentage of locations exhibiting significant increasing and decreasing trends exhibited in individual models.

Uncertainty in the spatial structure of trends in streamflow extremes is further investigated using both intra-model (to reflect GCM uncertainty) and inter-model correlations (to reflect the combined GCM-GHM uncertainty). A more robust spatial pattern of projected trends under RCP6.0 was found, indicated through generally higher intra-/inter-model correlation ~~(multi-model averages of 0.34/0.04)~~ compared to those exhibited from trends simulated under RCP2.6~~; multi-model averages of 0.08/0.01)~~ across all GHMs. This feature potentially reflects the less contrasted regional climate change of RCP2.6 relative to RCP6.0. The inter-model correlation ~~(ranging from -0.18 to 0.21)~~ is

consistently lower than intra-model correlation ~~(ranging from -0.03 to 0.48)~~ due to the combined uncertainty of both GHMs and GCMs.

Table 7. The uncertainty in the characteristics of projected trends (GCMRCP2.6 and GCMRCP6.0) across 18 members at the global scale (five GHMs). Trend mean and trend standard deviation have unit of %-change per decade. At-site significance of trend was identified using Mann-Kendall test at 10% level and the percentage of grid cells showing significant increasing/decreasing trends was reported (no field significance test was conducted). Intra-model average value of each metric across is shown for each GHM (numbers of simulations are provided in the first column).

Model	No. of sim	Trend mean		Trend standard deviation		% of sig. inc. trends		% of sig. dec. trends		Intra-model correlation		Inter-model correlation	
		GCM RCP2.6	*GCM RCP6.0*	*GCM RCP2.6*	*GCM RCP6.0*	*GCM RCP2.6*	*GCM RCP6.0*	*GCM RCP2.6*	*GCM RCP6.0*	*GCM RCP2.6*	*GCM RCP6.0*	*GCM RCP2.6*	*GCM RCP6.0*
H08	4	0.1	0.3	2.5	3.4	14.2	22.1	11.6	19.3	0.17	0.41	0.02	0.21
LPJmL	4	-0.1	-0.2	2.1	3.0	10.0	19.1	9.4	19.7	0.04	0.41	0.01	0.18
ORCHIDEE	2	-0.5	-0.8	2.6	3.6	9.1	14.4	17.6	28.1	0.07	0.34	0.03	0.11
PCR-GLOBWB	4	0.1	0.0	2.4	3.4	15.1	22.7	11.6	20.2	0.07	0.30	0.02	0.18
WaterGAP2	4	0.2	0.5	2.3	3.0	13.0	25.9	8.0	11.8	0.03	0.25	0.01	0.17
Multi-model min	-	-0.6	-1.3	1.8	2.6	7.5	12.3	7.1	9.6	-0.03	0.12	-0.11	-0.18
Multi-model max	-	0.4	0.8	2.9	4.1	18.0	30.1	21.2	35.0	0.30	0.48	0.21	0.21
Multi-model average	-	0.0	0.0	2.3	3.2	12.6	21.6	11.0	18.9	0.08	0.34	0.01	0.04

To quantity the robustness in terms of regions with significant trends in streamflow extremes, the number of simulations showing significant increasing/decreasing trends was counted for each grid cell (value ranging from 0 to 18). As shown in Figure 45, the projections under RCP2.6 (top panels) do not suggest many regions with an increasing trend for most ensemble members, but consistently suggest decreasing trends over the majority of Africa, Australia and the western America. Although both scenarios suggested a similar spatial pattern, projections under the RCP6.0 scenario (lower panels) show a substantially higher robustness in terms of regions with significant changes over time in streamflow extremes. For instance, significant increasing trends are projected consistently over southern and south-eastern Asia, eastern Africa, and Siberia, while high agreement of decreasing trends is found over southern Australia, north-eastern Europe, the Mediterranean and north-western North America. These findings share some similarity with a previous investigation that used the ISIMIP Fast Track simulations (published before the ISIMIP2a and 2b simulations used here) to identify regions projected with an increasing magnitude of 30-year return level of river flow ~~(Dankers et al., 2014)~~(Dankers et al., 2014). Specifically, both studies suggest overall: (1) an increasing trend over Siberia and South-East Asia; and (2) a decreasing trend over north-eastern Europe and north-western North America. The present study, however, additionally highlights a dominant decreasing trend over Australia, which was not shown previously. The different numbers of ensemble members (45 in ~~Dankers et al. (2014)~~Dankers et al. (2014) and 18 in the present study) and greenhouse gas concentration scenario (RCP8.5 in ~~Dankers et al. (2014)~~Dankers et

al. (2014) and RCP2.6/RCP6.0 in the present study) between two studies indicate that the choice of GCM-GHM ensemble and greenhouse gas concentration scenarios could lead to substantially different projections of changes in flood hazard at the regional scale.

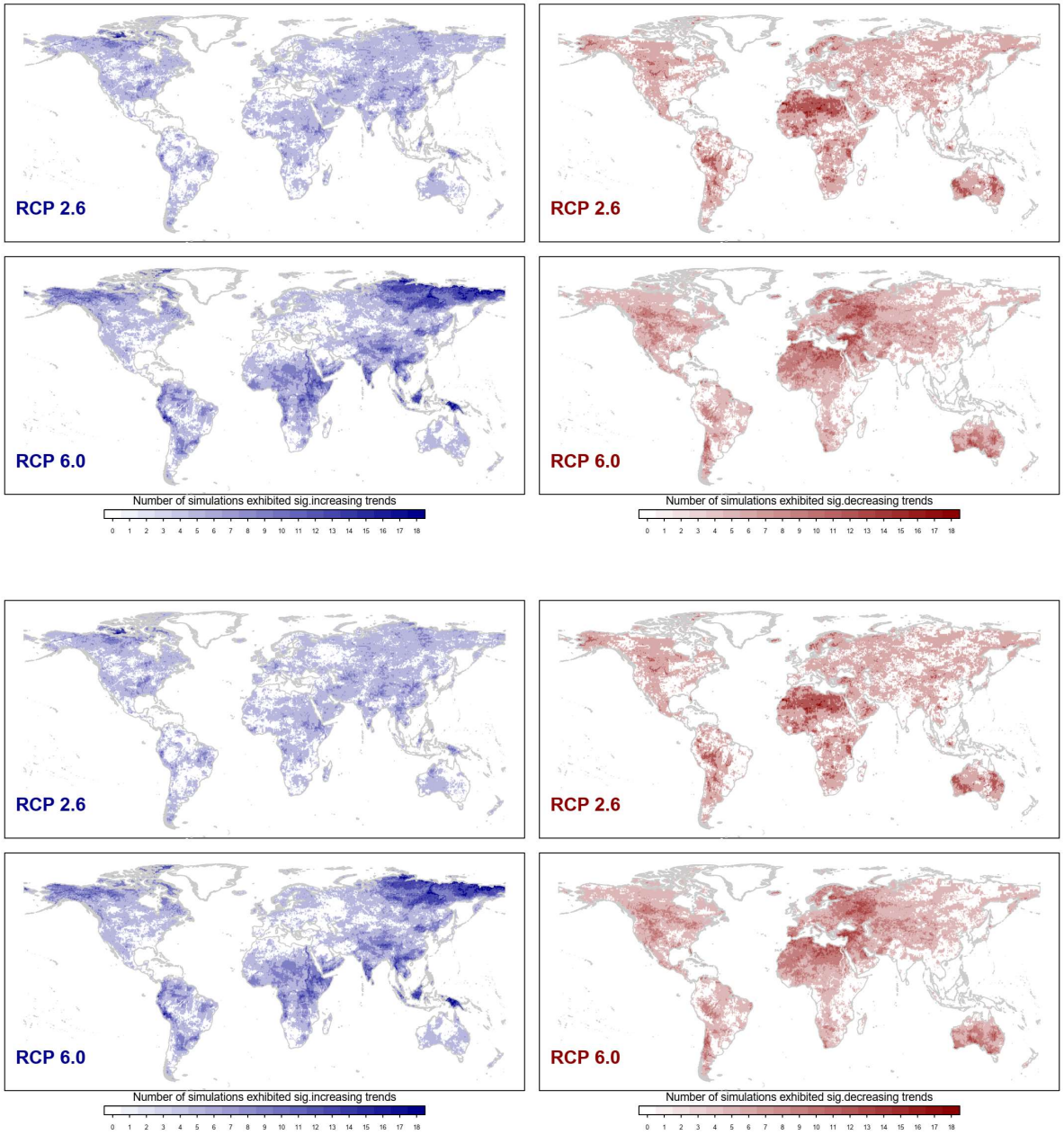
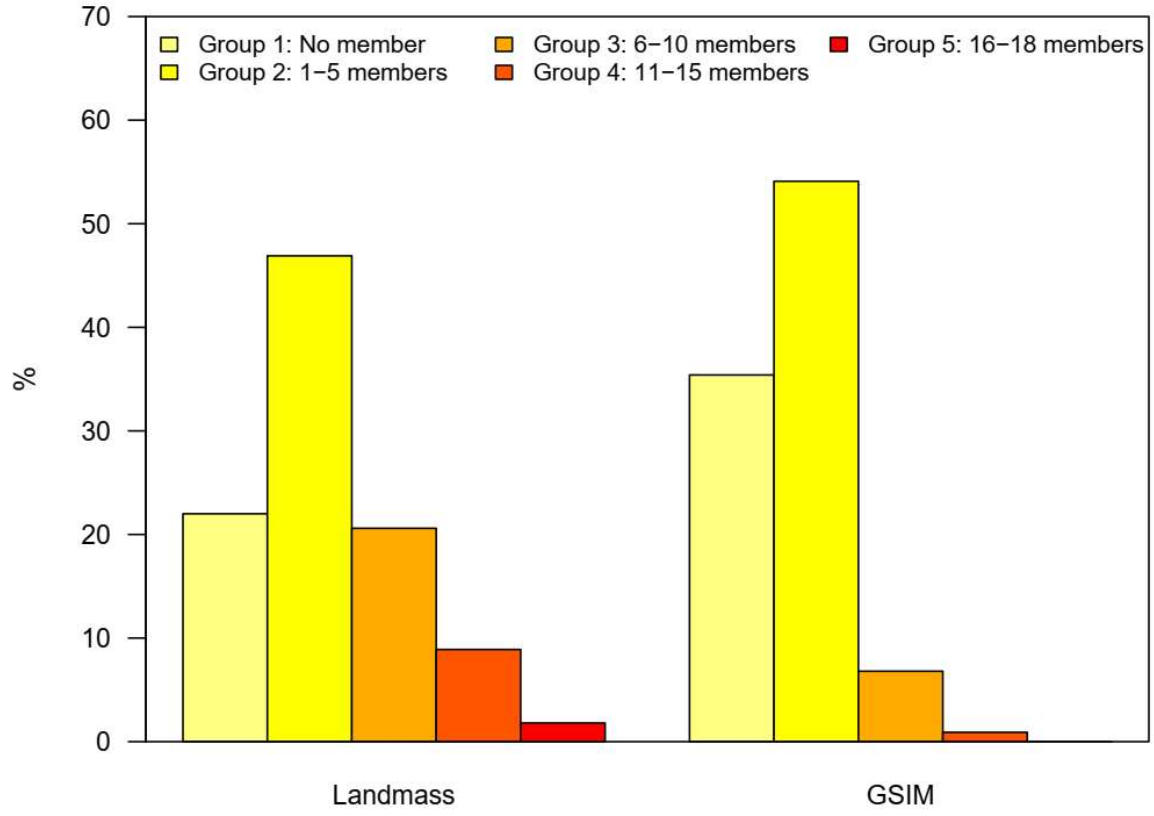


Figure 45. Number of simulations showing statistically significant trends at the 10% level at each grid cell. The left panels show results for the assessment of increasing trends, while the right panels show results for significant decreasing trends. Top: results of GCMRCP2.6 simulations; Bottom: results of GCMRCP6.0 simulations.

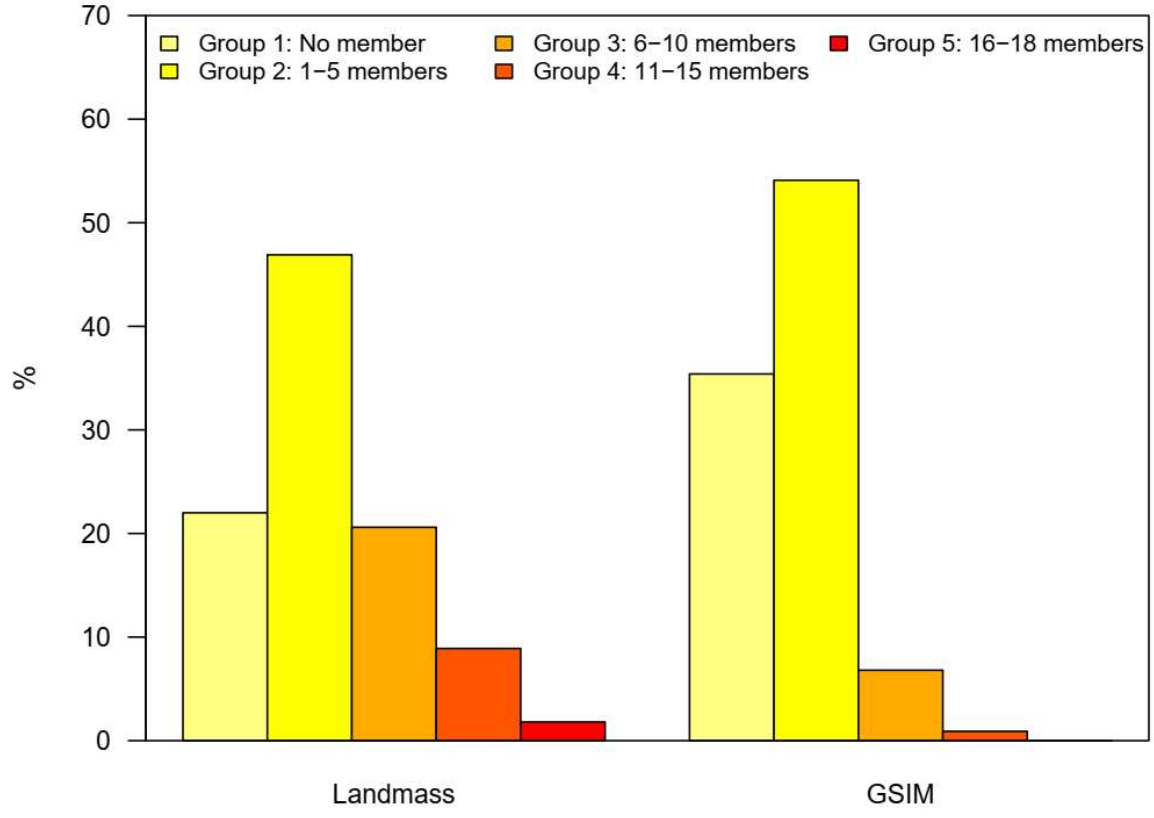
These results suggest the key role of GCM uncertainty in projections of changes in flood hazards, emphasising the importance of a flexible adaptation strategy at the regional scale that can take this uncertainty into account ~~(Dankers et al., 2014).~~(Dankers et al., 2014) such as increasing flexibility in reservoir operations, and focusing on improved infrastructure resilience, and safety to prepare for uncertain changes in the flood hazards. Such a strategy is achievable only through a reliable and robust understanding of the change in flood hazard. The assessment of the representativeness of streamflow observations (section 3.2), however, demonstrated that the observation locations selected for this assessment are not a representative sample of the entire land mass. As a result, inference of changes in flood hazard may be biased toward well-observed regions.

To further highlight the potential impact of limitations in observed streamflow datasets, the proportion of available stream gauges located in regions with different levels of projected 'flood risk' was assessed. We first categorised each ~~grid-cell into one of the five 'flood-risk' groups based on the number of simulations projecting a significant increasing trend.~~simulation grid-cell into one of the five 'flood-risk' groups. Note that in this analysis, "risk" is defined as the number of simulations projecting a significant increasing trend, rather than the prominent definition of risk as the combination of hazard, exposure and vulnerability (Kron, 2005). In this analysis, RCP6.0 scenario was chosen as it yielded a higher global 'risk' of flood hazard relative to RCP2.6 scenario.

Figure ~~5~~6 presents the percentage of all simulated grid cells (left panel) ~~and of the subset of GSIM station (right panel) falling~~categorized in each of the five groups~~.~~, and of GSIM stations located in each group (right panel). As can be seen, 11.7% of grid cells fell into the "high risk" groups (8.9% from Group 4 with 11-15 ensemble members, and 1.8% in Group 5 with 16-18 ensemble members), ~~compared to only 0.9% of stations available in GSIM archive (0.9% from Group 4 and no station located in Group 5). In contrast,~~while 68.9% of grid cells fell into the "low risk" groups (22.0% for Group 1 with no ensemble members, and 46.9% for Group 2 with 1-5 ensemble members~~),~~). Of all GSIM stations, only 0.9% are located in "high risk" grid cells (no station located in Group 5 grid cells) compared to 89.5% of stations ~~available~~located in ~~GSIM archive~~"low risk" grid cells (35.4% for Group 1 and 54.1% for Group 2). The uneven distribution of stream gauges indicates potential difficulties in using observational records to provide an assessment of global or regional changes in flood hazard, which in part arises from data caveats associated with the spatiotemporal coverage and quality of observed gauge records across the globe. This finding further suggests the urgent demand for ongoing efforts to make streamflow observation more accessible. In addition. new innovations in remote sensing (Gouweleeuw et al., 2018), or development of runoff reanalysis (Ghiggi et al., 2019) should also be supported to complement the understanding of changes in floods for locations that were not observed by stream gauges.


Group 1: No member
Group 2: 1−5 members
Group 3: 6−10 members
Group 4: 11−15 members
Group 5: 16−18 members
%
0
10
20
30
40
50
60
70
Landmass
GSIM

592


Group 1: No member
Group 2: 1−5 members
Group 3: 6−10 members
Group 4: 11−15 members
Group 5: 16−18 members
%
0
10
20
30
40
50
60
70
Landmass
GSIM

593

Figure 56. Percentage of grid-cell ("Landmass") grouped by the number of simulations projecting a significant increasing trend under RCP6.0 scenario; and the percentage of streamflow stations ("GSIM") assigned into each group. The range of possible simulations is from 0 to 18 and binned into five groups (Group 1: no members, Group 2: from 1 to 5 members, Group 3: from 6 to 10 members, Group 4: from 11 to 15 members and Group 5: from 16 to 18 members). To identify which group a specific station belongs to, the geographical coordinates of that station was superimposed on top of the global 'flood-risk' map.

4 Summary and conclusions

To reconcile observed and simulated trendsexplore the appropriateness of GHMs in historicalsimulating changes in flood hazards at the global and continental scale, this study evaluated the capacity of six GHMs to reproduce the characteristics of historical trends in 7-day annual maximum streamflow over the 1971-2005 period, using. The study also explored the implications of simulation uncertainty to projected changes in flood hazards over the 2006-2099 period. The findings of these investigations are summarized as follows.

1. Using observations from the Global Streamflow Indices and Metadata (GSIM) archive. The observed trends in annual maximum streamflow confirm, this study confirms previous findings about changes in flood hazard over data-covered regions (Do et al., 2017)(Do et al., 2017), in which significant decreasing trends were found mostly in Australia, the Mediterranean region, western US, eastern Brazil and Asia (Japan and southern India), while significant increasing trends were more common over central US, southern Brazil, and the northern part of Western Europe.
2. The ability ofTrends simulated by GHMs, when using an observational climate forcing, show moderate capacity to reproduce trends in streamflow maxima was assessed, focusing on fourthe characteristics of observed trends (i.e. the mean and standard deviation of trends, the percentage of stations showing significant increasing/decreasing trends, and the spatial structure of trends). Trends simulated by GHMs, when using an observational
2.3. Climate variability and climate forcing, show moderate capacity to reproduce the characteristics of observed trends. Climate forcingmodel uncertainty (i.e., the effect of using different GCMs to simulate the historical climate), however,) significantly reduced the extent to which the GHMs' captured the observed spatial structure of trends. This was evident through significantly lower spatial correlation between observed hydrological trends and simulated trends, when GCMs were used for the climate forcing, than when climate observations were used.

The simulated trends over observed areas inadequately represented spatially averaged trends simulated for wider spatial areas from all GHM grid cells at the continental and global scales. This was evident in most simulations for

trend mean and trend standard deviation, indicating a potential ~~mismatch between observation-based and model-based inferences about changes in flood hazard. As a result, alternatives for conventional approach in estimating change of streamflow extremes at the global and regional scale (i.e. unweighted mean across all grid points) should be investigated. For instance, the spatial weighted averages (e.g. using inverse distance relative to observed locations as weights) could be used to compute global means of changes.~~ ~~Regional analysis using homogenised regions as the basis of reporting spatial domains (Zaherpour et al., 2018;Gudmundsson et al., 2019) could be a potential alternative for continental scale assessment.~~

4. ~~Uncertainties of trends in streamflow extremes were analysed to assess their implication on the development of projected changes in flood hazard over the 2006-2099 period.~~ bias toward well-observed regions of observation-based inferences about changes in flood hazard.

5. Under both RCP2.6 and RCP6.0 greenhouse gas concentration scenarios, simulated trends in 7-day maximum streamflow across ensemble members have relatively low uncertainty in terms of the global trend mean (ranging from -1.3% to 0.8% change per decade) and trend standard deviation (ranging from 1.8% to 4.1% change per decade). ~~The~~

~~3.~~6. Projected trends have wide spread of the percentage of land mass showing significant ~~trends is high~~changes, ranging from 7.5% (7.1%) to 30.1% (35.0%) for significant increasing (decreasing) trends. This result indicates that limited changes to the global mean flood hazard could potentially mask out significant regional changes. ~~The spatial correlations across inter-model trend patterns are generally low (ranging from -0.18 to 0.21), further indicating high levels of uncertainty.~~

7. ~~In terms of regional planning to mitigate flood hazard, individual models may provide contradictory signals of changes in flood hazard for a specific region.~~Projected trends in flood hazards show low inter-model spatial correlations (ranging from -0.18 to 0.21), indicating high uncertainty in future changes in flood hazards at the global scale. Under RCP6.0 scenario, some regions, e.g. south-eastern Asia, eastern Africa, Siberia, were consistently projected with significant increasing trends, which has some similarity to previous findings that used ISIMIP Fast Track simulations ~~(Dankers et al., 2014). These 'high-risk' regions, however,~~(Dankers et al., 2014).

~~4.~~8. 'High-risk' regions (consistently projected with a significant increase in floods) of future changes in floods are sparsely sampled, covered by less than 1% of all available stream-gauges listed in the catalogue of GSIM. Data coverage, as a result, remains the key limitation of this study, which could potentially lead to an erroneous conclusion on the state-of-understanding of historical trends in flood hazard globally. Specifically, substantial changes, although having occurred, might not be captured by available streamflow records.

Our findings also show that individual models may provide contradictory signal of changes in flood hazards for a specific region, indicating high uncertainty in model-based inferences of changes in flood hazards. As a result, alternatives for the conventional approach in estimating changes in streamflow extremes at the global and regional scale (i.e. unweighted mean across all grid points) should be investigated. For instance, the spatial weighted averages (e.g. using inverse distance relative to observed locations as weights) could be used to compute global means of changes. Regional analysis using homogenised regions as the basis of reporting spatial domains (Zaherpour et al., 2018;Gudmundsson et al., 2019) could be a potential alternative for continental scale assessment.

The substantial discrepancy of trends simulated by different GHMs, despite having a common forcing boundary, represents another challenges in using GHM ensemble, as there are a wide range of factors that could contribute to these discrepancies. This study provides a (non-exhaustive) list of key differences across participated GHMs (supplementary Section 1) that could individually or collectively lead to different model outputs. Diagnosing the influence of these factors to models' capacity in simulating trends is still under-represented in the literature, and is an important research agenda for future investigations. For instance, the impact of different methods to simulate snow dynamic could be assessed by investigating model performances across catchments where snowmelt plays an (in)significant role in flood generations.

Improved performance of GHMs in terms of simulating changes in flood hazard, considering the many factors influencing model capacity, is achievable only through the combined efforts of many communities. The spread of trends in streamflow extremes (trend standard deviation) could be simulated more accurately by finer spatiotemporal resolution GHMs. Such an improvement in GHMs, however, is highly dependent on the quality of input datasets (e.g. dam operations, historical irrigation databases and land-use/land-cover, in addition to atmospheric forcing), which are driven by advances in other geophysical disciplines ~~(Bierkens et al., 2015;Wood et al., 2011)~~(Bierkens et al., 2015;Wood et al., 2011). The moderate capacity of GHMs in terms of simulating the spatial structure of trends in streamflow extremes indicates the need for improved representation of runoff generation at the global scale (e.g. to better reflect rainfall-runoff relationship and the contribution of snow-dynamics), which is also a focus of large-sample hydrology ~~(Gupta et al., 2014;Addor et al., 2017)~~(Gupta et al., 2014;Addor et al., 2017). Uncertainty in GCMs, a long-standing challenge for the climate community, should also be addressed to enable robust projections of flood hazard in a warmer climate. One possibility is through constraining model performance using historical observations~~,~~ (to prevent climate models projecting an unrealistic state of the future climate system such as atmosphere energy balance or cloud feedbacks), which could potentially reduce the uncertainties of atmospheric forcing projections (Greve et al., 2018;Lorenz et al., 2018;He and Soden, 2016;Padrón et al., 2019)~~.~~. In addition,

future development of GHMs should also pay attention to model's capacity to simulate flood timing, an important metric to represent flood generation processes (Blöschl et al., 2017;Hall and Blöschl, 2018;Do et al., 2019). Integrating more sophisticated and effective routing schemes into future generations of GHM should also be emphasized to ensure runoff is accurately converted into river discharge (Zhao et al., 2017).

This study presents a comprehensive investigation of historical and future changes in flood hazard using a hybrid model-observation approach. The results highlighted a substantial difference between trend characteristics simulated by GHMs and that obtained from GSIM archive~~, suggesting~~. Our findings, therefore, suggest more attention should be paid to investigating GHMs performance in the context of historical and future flood hazard~~.~~, which is important for not only the scientific community but also for stakeholders when using results of GHM simulations (Krysanova et al., 2018). This is particularly important to determine the appropriateness of GHMs in specific investigations, as model performance may vary substantially across different variables (e.g. moderate capacity in simulating spatial structure of trends may be accompanied by a low performance in terms of simulating trend mean).

Large-sample evaluations, however, are highly dependent on data availability, which ~~has been emphasised as~~is one of the key barriers to a holistic perspective of changes in floods. ~~Specifically~~In this study, the unevenly distributed GSIM stations, partially due to the constraint in data accessibility, do not provide representative samples at both global and continental scale. Sustained and collective efforts from the broad hydrology community~~,~~ (Addor et al., 2019), therefore, are required to make streamflow data becomes more FAIR (Findable, Accessible, Interoperable and Reusable; see Wilkinson et al., 2016), and ultimately complement our limited understanding of flood hazards. Data providers, considering their tremendous investments in maintaining and making streamflow observations publicly available ~~in the public domain~~, remain key agencies to enhance the evidence-base of the global terrestrial water cycle and changes in flood hazard. The important contribution of these agencies should be acknowledged appropriately when streamflow data being used. Centralised organisations such as GRDC or WMO should also push forward the movement of making streamflow data accessible to the research community. More initiatives based on citizen science (Paul et al., 2018) should be adopted, as this is a potential option to crowdsource water data and offset the limitation of traditional observation system. Finally, attention should also be paid to stream gauges maintenance, data housekeeping and data sharing to ensure ongoing flood monitoring is available to the present and future generations.

Acknowledgement

We thank two anonymous reviewers for their constructive comments that helped to improve the manuscript. Comments from Grabriele Villarini and Murray Peel to improve the manuscript are also appreciated. Hong Xuan Do

is currently funded by School for Environment and Sustainability, University of Michigan through grant number U064474. This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide and Flux HPC service at the University of Michigan. The daily streamflow data sets were made publicly available from many data provider, including: the Global Runoff Data Centre (GRDC), the ARCTICNET initiative; the China Hydrology Data Project; The GEWEX Asian Monsoon Experiment – Tropics project; USGS National Data Services; Environment Canada; Brazilian National Water Agency; Spanish Center for Hydrographic Studies; Japanese Ministry of Land, Infrastructure, Transport and Tourism; Australian Bureau of Meteorology; Indian Central Water Commission.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci. Discuss., 2017, 1-31, 10.5194/hess-2017-169, 2017.

Alfieri, L., Burek, P., Feyen, L., and Forzieri, G.: Global warming increases the frequency of river floods in Europe, Hydrol. Earth Syst. Sci., 19, 2247-2260, 10.5194/hess-19-2247-2015, 2015.

Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., Wyser, K., and Feyen, L.: Global projections of river flood risk in a warmer world, Earth's Future, n/a-n/a, 10.1002/2016EF000485, 2017.

Arnell, N., and Gosling, S.: The impacts of climate change on river flood risk at the global scale, Climatic Change, 1-15, 10.1007/s10584-014-1084-5, 2014.

Asadieh, B., and Krakauer, N. Y.: Global change in streamflow extremes under climate change over the 21st century, Hydrol. Earth Syst. Sci., 21, 5863-5874, 10.5194/hess-21-5863-2017, 2017.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881-2903, 10.5194/hess-21-2881-2017, 2017.

Bennett, B., Leonard, M., Deng, Y., and Westra, S.: An empirical investigation into the effect of antecedent precipitation on flood volume, Journal of Hydrology, https://doi.org/10.1016/j.jhydrol.2018.10.025, 2018.

Berghuijs, W. R., Woods, R. A., Hutton, C. J., and Sivapalan, M.: Dominant flood generating mechanisms across the United States, Geophysical Research Letters, 43, 4382-4390, 10.1002/2016GL068070, 2016.

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W. A., Heinke, J., von Bloh, W., and Gerten, D.: Impact of reservoirs on river discharge and irrigation water supply during the 20th century, Water Resources Research, 47, doi:10.1029/2009WR008929, 2011.

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., Sudicky, E., Sutanudjaja, E. H., van de Giesen, N., Winsemius, H., and Wood, E. F.: Hyper-resolution global hydrological modelling: what is next?, Hydrological Processes, 29, 310-320, 10.1002/hyp.10391, 2015.

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.

Burn, D. H., and Whitfield, P. H.: Changes in floods and flood regimes in Canada, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 41, 139-150, 10.1080/07011784.2015.1026844, 2016.

Burn, D. H., and Whitfield, P. H.: Changes in flood events inferred from centennial length streamflow data records, Advances in Water Resources, 121, 333-349, https://doi.org/10.1016/j.advwatres.2018.08.017, 2018.

CRED: The human cost of natural disasters: A global perspective, Centre for Research on the Epidemiology of Disasters, Brussels, 2015.

Cunderlik, J. M., and Ouarda, T. B. M. J.: Trends in the timing and magnitude of floods in Canada, Journal of Hydrology, 375, 471-480, http://dx.doi.org/10.1016/j.jhydrol.2009.06.050, 2009.

Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., Heinke, J., Kim, H., Masaki, Y., and Satoh, Y.: First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble, Proceedings of the National Academy of Sciences, 111, 3257-3261, 2014.

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.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata, Earth Syst. Sci. Data, 10, 765-785, 10.5194/essd-10-765-2018, 2018a.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive - Part 1: Station catalog and Catchment boundary, in, PANGAEA, 2018b.

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., Willett, K. M., Aguilar, E., Brunet, M., Caesar, J., Hewitson, B., Jack, C., Klein Tank, A. M. G., Kruger, A. C., Marengo, J., Peterson, T. C., Renom, M., Oria Rojas, C., Rusticucci, M., Salinger, J., Elrayah, A. S., Sekele, S. S., Srivastava, A. K., Trewin, B., Villarroel, C., Vincent, L. A., Zhai, P., Zhang, X., and Kitching, S.: Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, Journal of Geophysical Research: Atmospheres, 118, 2098-2118, 10.1002/jgrd.50150, 2013.

Forzieri, G., Feyen, L., Russo, S., Vousdoukas, M., Alfieri, L., Outten, S., Migliavacca, M., Bianchi, A., Rojas, R., and Cid, A.: Multi-hazard assessment in Europe under climate change, Climatic Change, 137, 105-119, 10.1007/s10584-016-1661-x, 2016.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., van Vliet, M., Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and Yamagata, Y.: Assessing the impacts of 1.5 °C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), Geosci. Model Dev., 10, 4321-4345, 10.5194/gmd-10-4321-2017, 2017.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: An observations-based global gridded runoff dataset from 1902 to 2014, Earth Syst. Sci. Data Discuss., 2019, 1-32, 10.5194/essd-2019-32, 2019.

Giuntoli, I., Villarini, G., Prudhomme, C., and Hannah, D. M. J. C. C.: Uncertainties in projected runoff over the conterminous United States, 150, 149-162, 10.1007/s10584-018-2280-5, 2018.

Gosling, S., Schmied, M. H., Betts, R., Chang, J., Ciais, P., Dankers, R., Döll, P., Eisner, S., Flörke, M., Gerten, D., Grillakis, M., Hanasaki, N., Hagemann, S., Huang, M., Huang, Z., Jerez, S., Kim, H., Koutroulis, A., Leng, G., Liu, X., Masaki, Y., Montavez, P., Morfopoulos, C., Oki, T., Papadimitriou, L., Pokhrel, Y., Portmann, F. T., Orth, R., Ostberg, S., Satoh, Y., Seneviratne, S., Sommer, P., Stacke, T., Tang, Q., Tsanis, I., Wada, Y., Zhou, T., Büchner, M., Schewe, J., and Zhao, F.: ISIMIP2a Simulation Data from Water (global) Sector (V. 1.1), in, GFZ Data Services, 2019.

Greve, P., Gudmundsson, L., and Seneviratne, S. I.: Regional scaling of annual mean precipitation and water availability with global temperature change, Earth Syst. Dynam., 9, 227-240, 10.5194/esd-9-227-2018, 2018.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, Journal of Hydrometeorology, 13, 604-620, 10.1175/JHM-D-11-083.1, 2012a.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resources Research, 48, n/a-n/a, 10.1029/2011WR010911, 2012b.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Time Series Indices and Homogeneity Assessment, in, PANGAEA, 2018a.
Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment, Earth Syst. Sci. Data, 10, 787-804, 10.5194/essd-10-787-2018, 2018b.
Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., and Seneviratne, S. I.: Observed Trends in Global Indicators of Mean and Extreme Streamflow, Geophysical Research Letters, 46, doi:10.1029/2018GL079725, 2019.
Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., Lewis, E., and Li, X.-F.: Detection of continental-scale intensification of hourly rainfall extremes, Nature Climate Change, 8, 803-807, 10.1038/s41558-018-0245-3, 2018.
Guha-Sapir, D., Hoyois, P., and Below, R.: Annual Disaster Statistical Review 2014: The numbers and trends, UCL, 2015.
Guimberteau, M., Ducharne, A., Ciais, P., Boisier, J. P., Peng, S., De Weirdt, M., and Verbeeck, H.: Testing conceptual and physically based soil hydrology schemes against observations for the Amazon Basin, Geoscientific Model Development, 7, 1115-1136, 2014.
Guimberteau, M., Zhu, D., Maignan, F., Huang, Y., Yue, C., Dantec-Nédélec, S., Ottlé, C., Jornet-Puig, A., Bastos, A., Laurent, P., Goll, D., Bowring, S., Chang, J., Guenet, B., Tifafi, M., Peng, S., Krinner, G., Ducharne, A., Wang, F., Wang, T., Wang, X., Wang, Y., Yin, Z., Lauerwald, R., Joetzjer, E., Qiu, C., Kim, H., and Ciais, P.: ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: model description and validation, Geosci. Model Dev., 11, 121-163, 10.5194/gmd-11-121-2018, 2018.
Gupta, H., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, p. 463-p. 477, 2014.
Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., Kjeldsen, T. R., Kriaučiūnienė, J., Kundzewicz, Z. W., Lang, M., Llasat, M. C., Macdonald, N., McIntyre, N., Mediero, L., Merz, B., Merz, R., Molnar, P., Montanari, A., Neuhold, C., Parajka, J., Perdigão, R. A. P., Plavcová, L., Rogger, M., Salinas, J. L., Sauquet, E., Schär, C., Szolgay, J., Viglione, A., and Blöschl, G.: Understanding flood regime changes in Europe: a state-of-the-art assessment, Hydrol. Earth Syst. Sci., 18, 2735-2772, 10.5194/hess-18-2735-2014, 2014.
Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources–Part 1: Model description and input meteorological forcing, Hydrology and Earth System Sciences, 12, 1007-1025, 2008a.
Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources–Part 2: Applications and assessments, Hydrology and Earth System Sciences, 12, 1027-1037, 2008b.
Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, Hydrological Processes, 25, 1191-1200, 10.1002/hyp.7794, 2011.
He, J., and Soden, B. J.: The impact of SST biases on projections of anthropogenic climate change: A greater role for atmosphere - only models?, Geophysical Research Letters, 43, 7745-7750, 2016.
Hodgkins, G. A., Whitfield, P. H., Burn, D. H., Hannaford, J., Renard, B., Stahl, K., Fleig, A. K., Madsen, H., Mediero, L., Korhonen, J., Murphy, C., and Wilson, D.: Climate-driven variability in the occurrence of major floods across North America and Europe, Journal of Hydrology, 552, 704-717, http://dx.doi.org/10.1016/j.jhydrol.2017.07.027, 2017.
Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A., Djalante, R., Ebi, K., and Engelbrecht, F.: Impacts of 1.5 ºC global warming on natural and human systems, 2018.
Hunger, M., and Döll, P.: Value of river discharge data for global-scale hydrological modeling, Hydrology and Earth System Sciences Discussions, 12, 841-861, 2008.
IPCC: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, Cambridge University Press, Cambridge, UK, and New York, NY, USA, 2012.
Ishak, E., Rahman, A., Westra, S., Sharma, A., and Kuczera, G.: Evaluating the non-stationarity of Australian annual maximum flood, Journal of Hydrology, 494, 134-145, 2013.
Ivancic, T., and Shaw, S.: Examining why trends in very heavy precipitation should not be mistaken for trends in very high river discharge, Climatic Change, 1-13, 10.1007/s10584-015-1476-1, 2015.

Johnson, F., White, C. J., van Dijk, A., Ekstrom, M., Evans, J. P., Jakob, D., Kiem, A. S., Leonard, M., Rouillard, A., and Westra, S.: Natural hazards in Australia: floods, Climatic Change, 1-15, 10.1007/s10584-016-1689-y, 2016.
Kettner, A. J., Cohen, S., Overeem, I., Fekete, B. M., Brakenridge, G. R., and Syvitski, J. P.: Estimating Change in Flooding for the 21st Century Under a Conservative RCP Forcing: A Global Hydrological Modeling Assessment, Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting, 157-167, 2018.
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.
Kim, H.: Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1), in, Data Integration and Analysis System (DIAS), 2017.
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.
Kundzewicz, Z. W., Graczyk, D., Maurer, T., Przymusińska, I., Radziejewski, M., Svensson, C., and Szwed, M.: Detection of change in world-wide hydrological time series of maximum annual flow, Global Runoff Date Centre, Koblenz, Germany, 2004.
Leonard, M., Metcalfe, A., and Lambert, M.: Frequency analysis of rainfall and streamflow extremes accounting for seasonal and climatic partitions, Journal of hydrology, 348, 135-147, 2008.
Liu, X., Tang, Q., Cui, H., Mu, M., Gerten, D., Gosling, S. N., Masaki, Y., Satoh, Y., and Wada, Y.: Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations, Environmental Research Letters, 12, 025009, 10.1088/1748-9326/aa5a3a, 2017.
Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America, Journal of Geophysical Research: Atmospheres, 123, 4509-4526, 2018.
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.
Mangini, W., Viglione, A., Hall, J., Hundecha, Y., Ceola, S., Montanari, A., Rogger, M., Salinas, J. L., Borzì, I., and Parajka, J.: Detection of trends in magnitude and frequency of flood peaks across Europe, Hydrological Sciences Journal, 63, 493-512, 10.1080/02626667.2018.1444766, 2018.
Miao, Q.: Are We Adapting to Floods? Evidence from Global Flooding Fatalities, Risk Analysis, 0, doi:10.1111/risa.13245, 2018.
Mueller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F. T., Reinecke, R., and Riedel, C.: Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use, Hydrology and Earth System Sciences, 20, 2877-2898, 2016.
Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., and Döll, P.: Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration, Hydrology and Earth System Sciences, 18, 3511-3538, 2014.
Munich Re: NatCatSERVICE: Loss events worldwide 1980 – 2014 Munich Re, Munich, 10, 2015.
Padrón, R. S., Gudmundsson, L., and Seneviratne, S. I.: Observational Constraints Reduce Likelihood of Extreme Changes in Multidecadal Land Water Availability, Geophysical Research Letters, 46, doi:10.1029/2018GL080521, 2019.
Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G. J., Lohmann, D., and Allen, M. R.: Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000, Nature, 470, 382-385, http://www.nature.com/nature/journal/v470/n7334/abs/10.1038-nature09762-unlocked.html#supplementary-information, 2011.
Paul, J. D., Buytaert, W., Allen, S., Ballesteros-Cánovas, J. A., Bhusal, J., Cieslik, K., Clark, J., Dugar, S., Hannah, D. M., Stoffel, M., Dewulf, A., Dhital, M. R., Liu, W., Nayaval, J. L., Neupane, B., Schiller, A., Smith, P. J., and Supper, R.: Citizen science for hydrological risk reduction and resilience building, 5, e1262, 10.1002/wat2.1262, 2018.
Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P. J.-F., Kim, H., Kanae, S., and Oki, T.: Incorporating Anthropogenic Water Regulation Modules into a Land Surface Model, Journal of Hydrometeorology, 13, 255-269, 10.1175/jhm-d-11-013.1, 2012.

Sakaguchi, K., Zeng, X., and Brunke, M. A.: Temporal-and spatial-scale dependence of three CMIP3 climate models in simulating the surface temperature trend in the twentieth century, Journal of Climate, 25, 2456-2470, 2012.
Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., and Lucht, W.: Contribution of permafrost soils to the global carbon budget, Environmental Research Letters, 8, 014026, 10.1088/1748-9326/8/1/014026, 2013.
Sharma, A., Wasko, C., and Lettenmaier, D. P.: If Precipitation Extremes Are Increasing, Why Aren't Floods?, Water Resources Research, 0, doi:10.1029/2018WR023749, 2018.
Smith, K.: Environmental hazards: assessing risk and reducing disaster, Routledge, 2003.
Stacke, T., and Hagemann, S.: Development and evaluation of a global dynamical wetlands extent scheme, Hydrology and Earth System Sciences, 16, 2915, 2012.
Stahl, K., Hisdal, H., Hannaford, J., Tallaksen, L., Van Lanen, H., Sauquet, E., Demuth, S., Fendekova, M., and Jordar, J.: Streamflow trends in Europe: evidence from a dataset of near-natural catchments, Hydrology and Earth System Sciences, 14, p. 2367-p. 2382, 2010.
Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geosci. Model Dev., 11, 2429-2453, 10.5194/gmd-11-2429-2018, 2018.
Swiss Re: Natural catastropes and man-made disaster in 2014, Swiss Reinsurance Company, Zurich, Switzerland, 52, 2015.
Veldkamp, T. I. E., Zhao, F., Ward, P. J., de Moel, H., Aerts, J. C., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel, Y., and Liu, X.: Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study, Environmental Research Letters, 13, 055008, 2018.
Wada, Y., Wisser, D., and Bierkens, M. F. P.: Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, Earth Syst. Dynam., 5, 15-40, 10.5194/esd-5-15-2014, 2014.
Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The inter-sectoral impact model intercomparison project (ISI–MIP): project framework, Proceedings of the National Academy of Sciences, 111, 3228-3232, 2014.
Wasko, C., and Sharma, A.: Global assessment of flood and storm extremes with increased temperatures, Scientific Reports, 7, 7945, 10.1038/s41598-017-08481-1, 2017.
Wasko, C., and Nathan, R.: Influence of changes in rainfall and soil moisture on trends in flooding, Journal of Hydrology, 575, 432-441, https://doi.org/10.1016/j.jhydrol.2019.05.054, 2019.
Westra, S., Alexander, L. V., and Zwiers, F. W.: Global Increasing Trends in Annual Maximum Daily Precipitation, Journal of Climate, 26, 15, 2013.
Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, Reviews of Geophysics, 52, 522-555, 10.1002/2014RG000464, 2014.
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 160018, 10.1038/sdata.2016.18, 2016.
Willner, S. N., Levermann, A., Zhao, F., and Frieler, K.: Adaptation required to preserve future high-end river flood risk at present levels, 4, eaao1914, 10.1126/sciadv.aao1914 %J Science Advances, 2018.
Woldemeskel, F., and Sharma, A.: Should flood regimes change in a warming climate? The role of antecedent moisture conditions, Geophysical Research Letters, 43, 7556-7563, 10.1002/2016GL069448, 2016.
Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., and Whitehead, P.:

~~Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resources Research, 47, n/a-n/a, 10.1029/2010WR010090, 2011.~~
~~Zaherpour, J., Gosling, S. N., Mount, N., Schmied, H. M., Veldkamp, T. I. E., Dankers, R., Eisner, S., Gerten, D., Gudmundsson, L., and Haddeland, I.: Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts, Environmental Research Letters, 2018.~~
~~Zaherpour, J., Mount, N., Gosling, S. N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Müller Schmied, H., Tang, Q., and Wada, Y.: Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models, Environmental Modelling & Software, 114, 112-128, https://doi.org/10.1016/j.envsoft.2019.01.003, 2019.~~
~~Zhan, C., Niu, C., Song, X., and Xu, C.: The impacts of climate variability and human activities on streamflow in Bai River basin, northern China, Hydrology Research, 44, 875-885, 10.2166/nh.2012.146, 2012.~~
~~Zhang, A., Zheng, C., Wang, S., and Yao, Y.: Analysis of streamflow variations in the Heihe River Basin, northwest China: Trends, abrupt changes, driving factors and ecological influences, Journal of Hydrology: Regional Studies, 3, 106-124, https://doi.org/10.1016/j.ejrh.2014.10.005, 2015.~~
~~Zhao, F., Veldkamp, T. I., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., and Portmann, F. T.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environmental Research Letters, 12, 075003, 2017.~~

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci. Discuss., 2017, 1-31, 10.5194/hess-2017-169, 2017.
Addor, N., Do, H. X., Alvarez-Garreto, C., Coxon, G., Fowler, K., and Mendoza, P.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, 2019.
Alfieri, L., Burek, P., Feyen, L., and Forzieri, G.: Global warming increases the frequency of river floods in Europe, Hydrol. Earth Syst. Sci., 19, 2247-2260, 10.5194/hess-19-2247-2015, 2015.
Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., Wyser, K., and Feyen, L.: Global projections of river flood risk in a warmer world, Earth's Future, n/a-n/a, 10.1002/2016EF000485, 2017.
Arnell, N. W., and Gosling, S. N.: The impacts of climate change on river flood risk at the global scale, Climatic Change, 1-15, 10.1007/s10584-014-1084-5, 2014.
Asadieh, B., and Krakauer, N. Y.: Global change in streamflow extremes under climate change over the 21st century, Hydrol. Earth Syst. Sci., 21, 5863-5874, 10.5194/hess-21-5863-2017, 2017.
Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881-2903, 10.5194/hess-21-2881-2017, 2017.
Bennett, B., Leonard, M., Deng, Y., and Westra, S.: An empirical investigation into the effect of antecedent precipitation on flood volume, Journal of Hydrology, https://doi.org/10.1016/j.jhydrol.2018.10.025, 2018.
Berghuijs, W. R., Woods, R. A., Hutton, C. J., and Sivapalan, M.: Dominant flood generating mechanisms across the United States, Geophysical Research Letters, 43, 4382-4390, 10.1002/2016GL068070, 2016.
Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W. A., Heinke, J., von Bloh, W., and Gerten, D.: Impact of reservoirs on river discharge and irrigation water supply during the 20th century, Water Resources Research, 47, doi:10.1029/2009WR008929, 2011.
Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., Sudicky, E., Sutanudjaja, E. H., van de Giesen, N., Winsemius, H., and Wood, E. F.: Hyper-resolution global hydrological modelling: what is next?, Hydrological Processes, 29, 310-320, 10.1002/hyp.10391, 2015.
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.
Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., and Bilibashi, A.: Changing climate both increases and decreases European river floods, Nature, 573, 108-111, 2019.
Burn, D. H., and Whitfield, P. H.: Changes in floods and flood regimes in Canada, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 41, 139-150, 10.1080/07011784.2015.1026844, 2016.
Burn, D. H., and Whitfield, P. H.: Changes in flood events inferred from centennial length streamflow data records, Advances in Water Resources, 121, 333-349, https://doi.org/10.1016/j.advwatres.2018.08.017, 2018.
CRED: The human cost of natural disasters: A global perspective, Centre for Research on the Epidemiology of Disasters, Brussels, 2015.
Cunderlik, J. M., and Ouarda, T. B. M. J.: Trends in the timing and magnitude of floods in Canada, Journal of Hydrology, 375, 471-480, http://dx.doi.org/10.1016/j.jhydrol.2009.06.050, 2009.
Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., Heinke, J., Kim, H., Masaki, Y., and Satoh, Y.: First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble, Proceedings of the National Academy of Sciences, 111, 3257-3261, 2014.
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.
Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive - Part 1: Station catalog and Catchment boundary, in, PANGAEA, 2018a.
Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata, Earth Syst. Sci. Data, 10, 765-785, 10.5194/essd-10-765-2018, 2018b.
Do, H. X., Westra, S., Leonard, M., and Gudmundsson, L.: Global-Scale Prediction of Flood Timing Using Atmospheric Reanalysis, Water Resources Research, 10.1029/2019wr024945, 2019.
Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., Willett, K. M., Aguilar, E., Brunet, M., Caesar, J., Hewitson, B., Jack, C., Klein Tank, A. M. G., Kruger, A. C., Marengo, J., Peterson, T. C., Renom, M., Oria Rojas, C., Rusticucci, M., Salinger, J., Elrayah, A. S., Sekele, S. S., Srivastava, A. K., Trewin, B., Villarroel, C., Vincent, L. A., Zhai, P., Zhang, X., and Kitching, S.: Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, Journal of Geophysical Research: Atmospheres, 118, 2098-2118, 10.1002/jgrd.50150, 2013.
FitzHugh, T. W., and Vogel, R. M.: The impact of dams on flood flows in the United States, River Research and Applications, 27, 1192-1215, 2011.
Forzieri, G., Feyen, L., Russo, S., Vousdoukas, M., Alfieri, L., Outten, S., Migliavacca, M., Bianchi, A., Rojas, R., and Cid, A.: Multi-hazard assessment in Europe under climate change, Climatic Change, 137, 105-119, 10.1007/s10584-016-1661-x, 2016.
Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., van Vliet, M., Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and Yamagata, Y.: Assessing the impacts of 1.5 °C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), Geosci. Model Dev., 10, 4321-4345, 10.5194/gmd-10-4321-2017, 2017.
Galton, F.: Regression towards mediocrity in hereditary stature, The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263, 1886.
Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth Syst. Sci. Data, 11, 1655-1674, 10.5194/essd-11-1655-2019, 2019.
Giuntoli, I., Villarini, G., Prudhomme, C., and Hannah, D. M. J. C. C.: Uncertainties in projected runoff over the conterminous United States, 150, 149-162, 10.1007/s10584-018-2280-5, 2018.
Gosling, S., Schmied, M. H., Betts, R., Chang, J., Ciais, P., Dankers, R., Döll, P., Eisner, S., Flörke, M., Gerten, D., Grillakis, M., Hanasaki, N., Hagemann, S., Huang, M., Huang, Z., Jerez, S., Kim, H.,

Koutroulis, A., Leng, G., Liu, X., Masaki, Y., Montavez, P., Morfopoulos, C., Oki, T., Papadimitriou, L., Pokhrel, Y., Portmann, F. T., Orth, R., Ostberg, S., Satoh, Y., Seneviratne, S., Sommer, P., Stacke, T., Tang, Q., Tsanis, I., Wada, Y., Zhou, T., Büchner, M., Schewe, J., and Zhao, F.: ISIMIP2a Simulation Data from Water (global) Sector (V. 1.1), in, GFZ Data Services, 2019.

Gouweleeuw, B. T., Kvas, A., Gruber, C., Gain, A. K., Mayer-Gürr, T., Flechtner, F., and Güntner, A.: Daily GRACE gravity field solutions track major flood events in the Ganges–Brahmaputra Delta, Hydrol. Earth Syst. Sci., 22, 2867-2880, 10.5194/hess-22-2867-2018, 2018.

Greve, P., Gudmundsson, L., and Seneviratne, S. I.: Regional scaling of annual mean precipitation and water availability with global temperature change, Earth Syst. Dynam., 9, 227-240, 10.5194/esd-9-227-2018, 2018.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, Journal of Hydrometeorology, 13, 604-620, 10.1175/JHM-D-11-083.1, 2012a.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resources Research, 48, n/a-n/a, 10.1029/2011WR010911, 2012b.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment, Earth Syst. Sci. Data, 10, 787-804, 10.5194/essd-10-787-2018, 2018.

Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., and Seneviratne, S. I.: Observed Trends in Global Indicators of Mean and Extreme Streamflow, Geophysical Research Letters, 46, doi:10.1029/2018GL079725, 2019.

Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., Lewis, E., and Li, X.-F.: Detection of continental-scale intensification of hourly rainfall extremes, Nature Climate Change, 8, 803-807, 10.1038/s41558-018-0245-3, 2018.

Guha-Sapir, D., Hoyois, P., and Below, R.: Annual Disaster Statistical Review 2014: The numbers and trends, UCL, 2015.

Guimberteau, M., Ducharne, A., Ciais, P., Boisier, J.-P., Peng, S., De Weirdt, M., and Verbeeck, H.: Testing conceptual and physically based soil hydrology schemes against observations for the Amazon Basin, Geoscientific Model Development, 7, 1115-1136, 2014.

Guimberteau, M., Zhu, D., Maignan, F., Huang, Y., Yue, C., Dantec-Nédélec, S., Ottlé, C., Jornet-Puig, A., Bastos, A., Laurent, P., Goll, D., Bowring, S., Chang, J., Guenet, B., Tifafi, M., Peng, S., Krinner, G., Ducharne, A., Wang, F., Wang, T., Wang, X., Wang, Y., Yin, Z., Lauerwald, R., Joetzjer, E., Qiu, C., Kim, H., and Ciais, P.: ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: model description and validation, Geosci. Model Dev., 11, 121-163, 10.5194/gmd-11-121-2018, 2018.

Gupta, H., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, p. 463-p. 477, 2014.

Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., Kjeldsen, T. R., Kriaučiūnienė, J., Kundzewicz, Z. W., Lang, M., Llasat, M. C., Macdonald, N., McIntyre, N., Mediero, L., Merz, B., Merz, R., Molnar, P., Montanari, A., Neuhold, C., Parajka, J., Perdigão, R. A. P., Plavcová, L., Rogger, M., Salinas, J. L., Sauquet, E., Schär, C., Szolgay, J., Viglione, A., and Blöschl, G.: Understanding flood regime changes in Europe: a state-of-the-art assessment, Hydrol. Earth Syst. Sci., 18, 2735-2772, 10.5194/hess-18-2735-2014, 2014.

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.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources–Part 2: Applications and assessments, Hydrology and Earth System Sciences, 12, 1027-1037, 2008a.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources–Part 1: Model description and input meteorological forcing, Hydrology and Earth System Sciences, 12, 1007-1025, 2008b.

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, Hydrological Processes, 25, 1191-1200, 10.1002/hyp.7794, 2011.

He, J., and Soden, B. J.: The impact of SST biases on projections of anthropogenic climate change: A greater role for atmosphere - only models?, Geophysical Research Letters, 43, 7745-7750, 2016.
Hodgkins, G. A., Whitfield, P. H., Burn, D. H., Hannaford, J., Renard, B., Stahl, K., Fleig, A. K., Madsen, H., Mediero, L., Korhonen, J., Murphy, C., and Wilson, D.: Climate-driven variability in the occurrence of major floods across North America and Europe, Journal of Hydrology, 552, 704-717, http://dx.doi.org/10.1016/j.jhydrol.2017.07.027, 2017.
Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A., Djalante, R., Ebi, K., and Engelbrecht, F.: Impacts of 1.5 ºC global warming on natural and human systems, 2018.
Hunger, M., and Döll, P.: Value of river discharge data for global-scale hydrological modeling, Hydrology and Earth System Sciences Discussions, 12, 841-861, 2008.
IPCC: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, Cambridge University Press, Cambridge, UK, and New York, NY, USA, 2012.
Ishak, E., Rahman, A., Westra, S., Sharma, A., and Kuczera, G.: Evaluating the non-stationarity of Australian annual maximum flood, Journal of Hydrology, 494, 134-145, 2013.
Ivancic, T., and Shaw, S.: Examining why trends in very heavy precipitation should not be mistaken for trends in very high river discharge, Climatic Change, 1-13, 10.1007/s10584-015-1476-1, 2015.
Johnson, F., White, C. J., van Dijk, A., Ekstrom, M., Evans, J. P., Jakob, D., Kiem, A. S., Leonard, M., Rouillard, A., and Westra, S.: Natural hazards in Australia: floods, Climatic Change, 1-15, 10.1007/s10584-016-1689-y, 2016.
Kettner, A. J., Cohen, S., Overeem, I., Fekete, B. M., Brakenridge, G. R., and Syvitski, J. P.: Estimating Change in Flooding for the 21st Century Under a Conservative RCP Forcing: A Global Hydrological Modeling Assessment, Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting, 157-167, 2018.
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.
Kim, H.: Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1), in, Data Integration and Analysis System (DIAS), 2017.
Kron, W.: Flood Risk = Hazard • Values • Vulnerability, Water International, 30, 58-68, 10.1080/02508060508691837, 2005.
Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., and Kundzewicz, Z. W.: How the performance of hydrological models relates to credibility of projections under climate change, Hydrological sciences journal, 63, 696-720, 2018.
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.
Kundzewicz, Z. W., Graczyk, D., Maurer, T., Przymusińska, I., Radziejewski, M., Svensson, C., and Szwed, M.: Detection of change in world-wide hydrological time series of maximum annual flow, Global Runoff Date Centre, Koblenz, Germany, 2004.
Leonard, M., Metcalfe, A., and Lambert, M.: Frequency analysis of rainfall and streamflow extremes accounting for seasonal and climatic partitions, Journal of hydrology, 348, 135-147, 2008.
Liu, X., Tang, Q., Cui, H., Mu, M., Gerten, D., Gosling, S. N., Masaki, Y., Satoh, Y., and Wada, Y.: Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations, Environmental Research Letters, 12, 025009, 10.1088/1748-9326/aa5a3a, 2017.
Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America, Journal of Geophysical Research: Atmospheres, 123, 4509-4526, 2018.
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.
Mangini, W., Viglione, A., Hall, J., Hundecha, Y., Ceola, S., Montanari, A., Rogger, M., Salinas, J. L., Borzì, I., and Parajka, J.: Detection of trends in magnitude and frequency of flood peaks across Europe, Hydrological Sciences Journal, 63, 493-512, 10.1080/02626667.2018.1444766, 2018.
Miao, Q.: Are We Adapting to Floods? Evidence from Global Flooding Fatalities, Risk Analysis, 0, doi:10.1111/risa.13245, 2018.

Mueller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F. T., Reinecke, R., and Riedel, C.: Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use, Hydrology and Earth System Sciences, 20, 2877-2898, 2016.
Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., and Döll, P.: Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration, Hydrology and Earth System Sciences, 18, 3511-3538, 2014.
Munich Re: NatCatSERVICE: Loss events worldwide 1980 – 2014 Munich Re, Munich, 10, 2015.
Padrón, R. S., Gudmundsson, L., and Seneviratne, S. I.: Observational Constraints Reduce Likelihood of Extreme Changes in Multidecadal Land Water Availability, Geophysical Research Letters, 46, doi:10.1029/2018GL080521, 2019.
Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G. J., Lohmann, D., and Allen, M. R.: Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000, Nature, 470, 382-385, http://www.nature.com/nature/journal/v470/n7334/abs/10.1038-nature09762-unlocked.html#supplementary-information, 2011.
Paul, J. D., Buytaert, W., Allen, S., Ballesteros-Cánovas, J. A., Bhusal, J., Cieslik, K., Clark, J., Dugar, S., Hannah, D. M., Stoffel, M., Dewulf, A., Dhital, M. R., Liu, W., Nayaval, J. L., Neupane, B., Schiller, A., Smith, P. J., and Supper, R.: Citizen science for hydrological risk reduction and resilience building, 5, e1262, 10.1002/wat2.1262, 2018.
Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P. J.-F., Kim, H., Kanae, S., and Oki, T.: Incorporating Anthropogenic Water Regulation Modules into a Land Surface Model, Journal of Hydrometeorology, 13, 255-269, 10.1175/jhm-d-11-013.1, 2012.
Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., and Lucht, W.: Contribution of permafrost soils to the global carbon budget, Environmental Research Letters, 8, 014026, 2013.
Sharma, A., Wasko, C., and Lettenmaier, D. P.: If Precipitation Extremes Are Increasing, Why Aren't Floods?, Water Resources Research, 0, doi:10.1029/2018WR023749, 2018.
Slater, L. J., Singer, M. B., and Kirchner, J. W.: Hydrologic versus geomorphic drivers of trends in flood hazard, Geophysical Research Letters, 42, 370-376, 10.1002/2014GL062482, 2015.
Smith, K.: Environmental hazards: assessing risk and reducing disaster, Routledge, 2003.
Stacke, T., and Hagemann, S.: Development and evaluation of a global dynamical wetlands extent scheme, Hydrology and Earth System Sciences, 16, 2915, 2012.
Stahl, K., Hisdal, H., Hannaford, J., Tallaksen, L., Van Lanen, H., Sauquet, E., Demuth, S., Fendekova, M., and Jordar, J.: Streamflow trends in Europe: evidence from a dataset of near-natural catchments, Hydrology and Earth System Sciences, 14, p. 2367-p. 2382, 2010.
Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geosci. Model Dev., 11, 2429-2453, 10.5194/gmd-11-2429-2018, 2018.
Swiss Re: Natural catastropes and man-made disaster in 2014, Swiss Reinsurance Company, Zurich, Switzerland, 52, 2015.
Veldkamp, T. I. E., Zhao, F., Ward, P. J., de Moel, H., Aerts, J. C., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel, Y., and Liu, X.: Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study, Environmental Research Letters, 13, 055008, 2018.
Wada, Y., Wisser, D., and Bierkens, M. F. P.: Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, Earth Syst. Dynam., 5, 15-40, 10.5194/esd-5-15-2014, 2014.
Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The inter-sectoral impact model intercomparison project (ISI–MIP): project framework, Proceedings of the National Academy of Sciences, 111, 3228-3232, 2014.
Wasko, C., and Sharma, A.: Global assessment of flood and storm extremes with increased temperatures, Scientific Reports, 7, 7945, 10.1038/s41598-017-08481-1, 2017.
Wasko, C., and Nathan, R.: Influence of changes in rainfall and soil moisture on trends in flooding, Journal of Hydrology, 575, 432-441, https://doi.org/10.1016/j.jhydrol.2019.05.054, 2019.
Westra, S., Alexander, L. A., and Zwiers, F. W.: Global Increasing Trends in Annual Maximum Daily Precipitation, Journal of Climate, 26, 15, 2013.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, Reviews of Geophysics, 52, 522-555, 10.1002/2014RG000464, 2014.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 160018, 10.1038/sdata.2016.18, 2016.

Willner, S. N., Levermann, A., Zhao, F., and Frieler, K.: Adaptation required to preserve future high-end river flood risk at present levels, 4, eaao1914, 10.1126/sciadv.aao1914 %J Science Advances, 2018.

Woldemeskel, F., and Sharma, A.: Should flood regimes change in a warming climate? The role of antecedent moisture conditions, Geophysical Research Letters, 43, 7556-7563, 10.1002/2016GL069448, 2016.

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., and Whitehead, P.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resources Research, 47, n/a-n/a, 10.1029/2010WR010090, 2011.

Zaherpour, J., Gosling, S. N., Mount, N., Schmied, H. M., Veldkamp, T. I. E., Dankers, R., Eisner, S., Gerten, D., Gudmundsson, L., and Haddeland, I.: Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts, Environmental Research Letters, 2018.

Zaherpour, J., Mount, N., Gosling, S. N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Müller Schmied, H., Tang, Q., and Wada, Y.: Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models, Environmental Modelling & Software, 114, 112-128, https://doi.org/10.1016/j.envsoft.2019.01.003, 2019.

Zhan, C., Niu, C., Song, X., and Xu, C.: The impacts of climate variability and human activities on streamflow in Bai River basin, northern China, Hydrology Research, 44, 875-885, 10.2166/nh.2012.146, 2012.

Zhang, A., Zheng, C., Wang, S., and Yao, Y.: Analysis of streamflow variations in the Heihe River Basin, northwest China: Trends, abrupt changes, driving factors and ecological influences, Journal of Hydrology: Regional Studies, 3, 106-124, https://doi.org/10.1016/j.ejrh.2014.10.005, 2015.

Zhao, F., Veldkamp, T. I., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., and Portmann, F. T.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environmental Research Letters, 12, 075003, 2017.

Supplementary Materials for

Historical and future changes in global flood magnitude – Evidence from a model-observation investigation

Hong Xuan Do[(1)(2)(3)(*)], Fang Zhao[(4)(5)(*)], Seth Westra[(1)], Michael Leonard[(1)], Lukas Gudmundsson[(6)], Julien Eric Stanislas Boulange[(7),] Jinfeng Chang[(~~7~~8)], Philippe Ciais[(~~7~~8)], Dieter Gerten[(5)(~~8~~9)], Simon N. Gosling[(~~9~~10)], Hannes Müller Schmied[(~~10~~)(11)(12)], Tobias Stacke[(~~12~~), Boulange Julien Eric Stanislas[(13)], Camelia-Eliza Telteu[(11)], Yoshihide Wada[(14)].

(1) School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, Australia.
(2) Faculty of Environment and Natural Resources, Nong Lam University, Ho Chi Minh City, Vietnam.
(3) School for Environment and Sustainability, University of Michigan, Ann Arbor, Michigan, United States.
(4) School of Geographical Sciences, East China Normal University, Shanghai, China.
(5) Potsdam Institute for Climate Impact Research, Potsdam, Germany.
(6) Institute for Atmospheric and Climate Science, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland.
(~~7~~(7) Center for Global Environmental Research, Japan.
(8) Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ/IPSL, Université Paris Saclay, 91191 Gif sur Yvette, France.
(~~8~~9) Geography Dept., Humboldt-Universität zu Berlin, Berlin, Germany.
(~~9~~10) School of Geography, University of Nottingham, Nottingham, United Kingdom.
(~~10~~11) Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany.
(~~11~~12) Senckenberg ~~Leibnitz~~Leibniz Biodiversity and Climate Research Centre (SBiK-F), Frankfurt am Main, Germany.
(~~12) Max Planck~~13) Institute ~~for Meteorology, Hamburg~~of Coastal Research, Helmholtz-Zentrum Geesthacht (HZG), Geesthacht, Germany.

~~(13) Center for Global Environmental Research, Japan.~~

(14) International Institute for Applied Systems Analysis, Laxenburg, Austria.

[(*)] *Corresponding authors:* Hong Xuan Do (hong.do@adelaide.edu.au) *and* Fang Zhao (fangzhao@pik-potsdam.de)

Contents:

1 Simulation information

1.1 Simulation setting

This section summarises the key simulation settings of each global hydrological model (GHM). Note that more detailed information is available in the protocols of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) available at https://www.isimip.org/protocol.

The following two input datasets were used for the GHM simulations, with specific model runs summarised in Table S1:

1. **Climate & CO_2 concentration scenarios (i.e. atmospheric forcing)**
   - GSWP3: observations-based dataset providing the climate forcing data.
   - RCP2.6: future climate and CO_2 concentration from RCP2.6
   - RCP6.0: future climate and CO_2 concentration from RCP6.0
   - HINDCAST: historical modelled climate and CO_2 concentration.
2. **Human influence and land-use scenarios**
   - nosoc: Naturalized runs (no human impact). No irrigation or man-made reservoirs. No population and GDP data prescribed.
   - varsoc: Varying historical land use and other human influences over historical period.
   - 2005soc: Fixed year-2005 land use and other human influences.

Note that GSWP3 was used as the sole observational atmospheric forcing dataset in this investigation. We also used modelled atmospheric forcing datasets introduced by four global climate models (GCM): GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR and MIROC5.

Table S1. Simulation set up of GHMs used in this investigation. 'Climate' represents atmospheric forcing dataset while 'human' represents human influence and land-use scenarios. Note that a more detailed inventory of available model runs is provided in Table S2.

<table>
<tr><th>Model</th><th>GSWP3_VARSOC</th><th>GSWP3_NOSOC</th><th>GCMHIND</th><th>GCMRCP2.6</th><th>GCMRCP6.0</th></tr>
<tr><td>H08</td><td rowspan="3">Climate: GSWP3
Human: varsoc</td><td rowspan="4">Climate: GSWP3
Human: nosoc</td><td>Climate: HINDCAST
Human: 2005soc</td><td rowspan="4">Climate: rcp26
Human: 2005soc</td><td rowspan="4">Climate: RCP6.0
Human: 2005soc</td></tr>
<tr><td>LPJmL</td><td rowspan="3">Climate: HINDCAST
Human: varsoc (except for ORCHIDEE using nosoc)</td></tr>
<tr><td>PCR-GLOBWB</td></tr>
<tr><td>WaterGAP2</td></tr>
<tr><td>MPI-HM</td><td rowspan="2">Simulations not available</td></tr>
<tr><td>ORCHIDEE</td></tr>
</table>

The results of preliminary assessment over 3666 observation locations suggest minor influence of human influence and land-use scenarios on the characteristics of trends in streamflow extremes (see section 4 of this supplementary material), and thus only GSWP3_NOSOC was used in the main text (denoted as GSWP3 in the main text).

1.2 Similarities and differences among participated GHMs

Although "global hydrological models" (GHMs) has been used as a universal terminology to represent the participating models in this study, there are two groups of models with fundamental differences:

- Hydrological models (HMs): this group includes H08, MPI-HM, PCR-GLOBWB, and WaterGAP2, which focused on quantitatively simulate the water balance components such as streamflow.
- Dynamic global vegetation models (DGVMs): this group includes LPJmL and ORCHIDEE, which focused on the shifts in vegetation cycle under natural and anthropogenic factors.

Table S2 summarizes the differences in schematization across the models while the following paragraphs highlight the key differences. The information contained in this table was synthesized by the ISIMIP community. We noted that adapted versions of Table S2 have been used as Supplementary of two other manuscripts.

Generally, different models can potentially simulate the timing and magnitude of the streamflow differently due to their different structure, and the features that are included/excluded from the model schematization. Nevertheless, it is difficult to attribute discrepancy of simulated changes in streamflow indices for differences

in model schematization, as there is no study that has explored the influence of specific component on changes in streamflow indices. Below are some key differences across models.

Interception

H08 and MPI-HM models do not use an interception scheme. PCR-GLOBWB simulates canopy interception as a function of vegetation type, which is annually prescribed by HYDE, MIRCA, and GLOBCOVER datasets (ESA Globcover 2005 Project, led by MEDIAS-France/POSTEL). LPJmL, ORCHIDEE, and WaterGAP2 models take into consideration the leaf area index. Furthermore, DGVMs also take into consideration the CO_2 fertilization effect and the dynamic vegetation effects.

Snow

Four models (LPJmL, MPI-HM, PCR-GLOBWB, and WaterGAP2) use the degree-day method to simulate snow accumulation and melt, while H08 and ORCHIDEE models use the physically based energy balance method. The energy balance method generally determines lower snow water equivalent values than the degree-day method (Haddeland et al., 2011). However, H08 only consider a single snow layer, which means the model tend to produce abundant snowmelt sooner (relative to ORCHIDEE) when enough energy has been accumulated.

Soil profile and groundwater

Generally, HMs (H08, MPI-HM, PCR-GLOBWB, and WaterGAP2) have one or two soil layers, because they focused on matching observed and simulated streamflow rather than dynamic vegetation growth like DGVMs (which have six and eleven soil layers). Most HMs (except MPI-HM) have a single groundwater layer. LPJmL doesn't have a groundwater layer, but its seepage is considered to have the role of groundwater recharge and groundwater runoff.

Components determining surface runoff and river discharge

Predominantly, the surface runoff is modelled as saturation excess overland flow and subsurface runoff as a function of soil. The streamflow routing is made wherever possible but the approach varies across models such as linear reservoir cascade (MPI-HM and WaterGAP2), continuity equation derived from linear reservoir model (LPJmL), drainage direction map 30 minutes (H08), STN-30p river network (ORCHIDEE), travel time routing (characteristic distance) linked with dynamic reservoir management (PCR-GLOBWB). Other varying features that could lead to differences in simulated runoff as well as discharge are:

- Only MPI-HM and ORCHIDEE do not include the reservoir management.
- H08, LPJmL, and ORCHIDEE do not include lakes in their structure.
- H08 and LPJmL do not include wetland scheme in their structure.

Human water management

ORCHIDEE model does not simulate the water use sectors. Others models simulate, mainly, irrigation, but H08, PCR-GLOBWB, and WaterGAP2 also simulate differently water used in the domestic, industry and livestock sectors.

Table S2 Hydrological processes represented in the Global Hydrological Models included in the present study.

Models	Model version ISIMIP2a / ISIMIP2b	Interception scheme	CO_2 fertilization effect	Snow scheme	Soil Layer / Total Soil Layer Depth (m)	Groundwater scheme	Surface runoff / subsurface runoff	Routing scheme	Reservoir operation	Lakes scheme	Wetlands scheme	Water use sectors scheme	References
H08	Hanasaki et al., (2008a&b) / Hanasaki et al., (2018)	no	no	energy balance method	1 / 1	1 renewable and 1 nonrenewable groundwater layer	saturation excess / *f*(soil) runoff properties varies with climate zones	based on 30' drainage direction map (DDM30)	yes	no	no	Irrigation, industry, domestic	Hanasaki et al., (2008a&b); Hanasaki et al., (2018).
LPJmL	Version 3.5 but with update of irrigation scheme in ISIMIP2b	*f*(LAI)	yes	degree-day method with precipitation factor	6 / 13	seepage reported as groundwater recharge and groundwater runoff	saturation excess / *f*(soil)	continuity equation derived from linear reservoir model	yes	no	no	irrigation	Bondeau et al., (2007) Schaphoff et al. (2013)
MPI-HM	R44 / v1.2	no	no	degree-day method	1 / 0[1]	not included	saturation excess / *f*(soil)	linear reservoir cascade	no	dynamical wetland extent scheme	dynamical wetland extent scheme	irrigation	Stacke and Hagemann, (2012)
ORCHIDEE	ORCHIDEE-Trunk Rev3013 / ORCHIDEE-MICT v8.4.1	*f*(LAI)	yes	physically based snow module + energy balance method	11 / 2	1 groundwater layer	infiltration excess / *f*(soil)	STN-30p river network	no	no	wetlands act as floodplains	no	Guimberteau et al., (2014) and Guimberteau et al., (2018)
PCR-GLOBWB	same version 2	*f*(veg)	no	degree-day method	2 / 1.5	1 groundwater layer	saturation excess / *f*(soil and gw)	travel time routing (characteristic distance) linked with dynamic reservoir operation	yes	yes	columns of water (no soil)	irrigation, domestic, industry, livestock	Wada et al. (2014) Sutanudjaja et al., (2018)
WaterGAP2	2.2 / 2.2c	*f*(LAI)	no	degree-day method	1 / Depending on land cover type between 0.1 and 4	1 groundwater layer	saturation excess, Beta function	linear reservoir cascade	yes	local and global lakes	local and global wetlands	irrigation, domestic, electricity, manufacturing, livestock	Müller Schmied et al., (2016)

Notes:
[1]: MPI-HM defines the soil storage in terms of maximum water column, varying between 0 m and 5 m;
f(gw) = subsurface flow or interflow modelled as a function of groundwater; f(veg) = function of vegetation type;
f(soil) = subsurface flow or interflow modelled as a function of soil moisture (soil).

1.3 Model versions used in ISIMIP2a and ISIMIP2b

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" capacity of the model used to simulate trends in floods during the future period (2006-2099). Specifically, PCR-GLOBWB is the only GHM that used the same version for both ISIMIP phases (noted in Figure S2). Below are the main modifications that have undertaken:

(1) LPJmL

The LPJmL model version used for ISIMIP2b was updated compared to the one used for ISIMIP2a in particular regarding the implementation of a new scheme to model irrigation systems, after Jägermeyr et al. (2015). In simulations with irrigation, this leads to various effects (differing in space and time) on most water balance components including discharge. Also the albedo of bare soil is made dependent on the soil moisture status.

(2) ORCHIDEE

ORCHIDEE-MICT v8.4.1 is a branch developed from ORCHIDEE-Trunk. ORCHIDEE-MICT improved the representation of the interactions between soil carbon, soil temperature and hydrology, and their resulting feedbacks on water and CO2 fluxes at high latitude, in addition to a recently developed fire module (Guimberteau et al., 2018).

ORCHIDEE-MICT focusing on high-latitude phenomena include the following non-exhaustive series of pivotal hydrological and biogeochemical interactions.

- A representation of permafrost physics and seasonal freeze–thaw cycles, which determine the soil hydrologic and thermal budgets and the volume and timing of lateral water flows to rivers.
- The impact of winter snow acting as an insulating "barrier" between soils and overlying air from fall to early spring. These have subsequent effects on soil temperature and water content, feeding back onto snow thickness itself.
- The seasonal mediation of plant water availability via snowmelt water, transpiration losses and the depth of the permafrost table (active layer thickness), which in turn determine the availability of the lateral water flows that feed rivers in the warmer months.
- The limitations on plant productivity and biomass due to acute climatic conditions in high-latitude regions. These primarily involve biotically prohibitive cold temperatures from fall to late spring, low soil moisture in dry-summer regions, and fire events caused by hot and dry conditions.
- The buildup of large soil carbon stocks under cold conditions through the slow burial of organic matter in the permafrost via cryoturbation and sedimentary soil formation processes.
- Feedbacks between high soil carbon concentrations and profiles of soil temperature, water and permafrost carbon content.

(3) WaterGAP2

Modifications of water use models compared to model version in ISIMIP2a

Deficit irrigation with 70% of optimal irrigation was applied in grid cells, which were selected based on Döll et al. (2014) and have 1) groundwater depletion of > 5 mm yr^{-1} over 1989–2009 and 2) a >5% fraction of mean annual irrigation water withdrawals in total water withdrawals over 1989–2009.

Modifications of WaterGAP Global Hydrology Model compared to model version in ISIMIP2a

- Groundwater recharge below surface water bodies is enabled in semi-arid and arid regions.
- Dynamic land area fractions as consequence of dynamic surface water extents.
- Precipitation input on surface water bodies is now also multiplied with the evaporation reduction factor (as evaporation) to keep water balance consistent.
- Modified routing approach where water is routed through the storages dependent upon the fraction of surface water bodies; otherwise water is routed directly into the river.
- New total water capacity input based on Batjes (2012).

- For global lakes and reservoirs (where the water balance is calculated in the outflow cell), water demand of all riparian cells is included in the water balance of the outflow cell and thus can be satisfied by global lake or reservoir storage.
- All water storage equations in horizontal water balance are solved analytically (except for local lakes). Those equations now include net abstractions from surface water or groundwater. As a consequence, sequence of net abstractions has been changed to 1) global lakes or reservoirs, 2) rivers, 3) local lakes.
- Net cell runoff is strictly the difference between the outflow of a cell and inflow from upstream cells at the end of a time step.
- Area correction factor (CFA) is included in water balance of lakes and wetlands.
- In 2.2 (ISIMIP2a), local and global lake storage could vary between the maximum storage S_{max} and zero. In 2.2c (as in versions before 2.2), local and global lake storage can drop to $-S_{max}$ as described in Hunger and Döll (2008). The area reduction factor (corresponding to the evaporation reduction factor in Hunger and Döll (2008), their eq. 1) has been changed accordingly (denominator: $2 \times S_{max}$). If lake storage S equals $-S_{max}$, the rediction factor is 1; if S equals S_{max}, the reduction factor is 0.
- Modified calibration routine: an uncertainty of 10% of long-term average river discharge is allowed (following Coxon et al., 2015), meaning that calibration runs in four steps: 1) test if γ alone is enough to calibrate to ±1% of observed value; 2) test if γ alone is enough to calibrate when 10% uncertainty of observed values are allowed; 3) adapt observed value by 10%, and test if γ plus CFA are sufficient for calibration; 4) add station correction factor (CFS) if all other steps were not successful, and set CFS values to 1 if between 0.98 and 1.02.
- All model parameters which are potentially used for the calibration/data assimilation integration (including also multiplicators) are now read from a text file in Javascript Object Notation (JSON) format.
- Regional changes based on Döll et al. (2014): 1) for Mississippi Embayment Regional Aquifer, groundwater recharge was overestimated, and thus the fraction of runoff from land recharging groundwater was reduced from 80–90% to 10% in these cells; 2) groundwater depletion in the North China Plain was overestimated by a factor of 4, and thus runoff coefficient γ was reduced from 3– 5 to 0.1 in this area; 3) all wetlands in Bangladesh were removed since diffuse groundwater recharge was unrealistically low.
- Due to different bug fixes reducing water balance error to a global sum of $<1*10^{-4}$ km³ yr^{-1}.
- In semi-arid/arid grid cells: In case of less precipitation then 12.5 mm day^{-1}, groundwater recharge is remaining in soil column (and not handled as runoff as in the version before).

(4) MPI-HM

The MPI-HM versions R44 and v1.2, as used in ISIMIP 2a and 2b respectively, differ only slightly in I/O infrastructure together with some modifications which concern only human impact simulations (which are not considered in ISIMIP 2a). More specifically changes are:

- Dynamic field allocation to allow for different model resolutions.
- Consistent reading and writing of parameter and restart files.
- Improvements for setup script.
- Limit irrigation gift to a maximum of 5% of the river flow storage per time step [1 day].
- Fix for parameter input (not affecting ISIMIP simulations).
- Fix for wetland water balance diagnostic (not affecting simulation results).

(5) H08

The newer version of the H08 model was used in ISIMIP2b (Hanasaki et al., 2018) while the older version used in ISIMIP2a (Hanasaki et al., 2008b). The main modifications of the updated version are:

- Revised irrigation/industrial and municipal water allocation.
- Inclusion of water transfer using aqueduct.
- Inclusion of seawater desalination scheme.
- Local reservoir implementation.

- Revised groundwater scheme.

Although the comparison between trends introduced by two versions of WaterGAP2 shows minor effects on changes in the key results of our investigation (see Section 3), simulations for other models are not readily available. As a result, the effect of modifications in GHMs cannot be checked in the context of this study.

2 Simulated streamflow extraction

For very large catchments, where excess rainfall takes a significant amount of time to reach the outlet, the routing scheme plays an important role in model performance related to high flow events (Zhao et al., 2017) and thus routed discharge is the more appropriate measure of simulated streamflow. The same simulation product, however, potentially does not perform well for small catchments, partially due to the coarse resolution of GHMs (Hunger and Döll, 2008). To address this concern, we adopted a common threshold of 9,~~000km~~^{2}000 km^2 (approximate the size of 1°×1° grid cell) to separate the selected catchments into two groups and applied different procedures to extract simulated streamflow.

2.1 Weighted-area average for stations with catchment from 0 to 9000 km^2

2.1.1 Producing weighted-area tables

For stations with catchment area less than or equal to 9000 km^2, the catchment boundary was superimposed to the ISIMIP grid to identify intersecting cells, and a weighted-area table was calculated for each case. Simulated runoff was extracted by averaging the un-routed surface runoff from all intersect cells (considering weight). Runoff was then converted into discharge data.

Figure S1 provides an illustration of the weighted-area table for station US_0002282 (red dot; Merrill catchment of Pascagoula River, Mississippi, US) which has the total number of 15 upstream cells (dark-grey cells). Two components of the weighted-area table were used to label intersect cells: (1) cell number (dark red) and (2) normalised fraction of each cell (weights) that is covered by the catchment boundary (dark blue). The normalisation was performed such that the weights add up to one for each catchment, and these weights are used to extract simulated runoff for this catchment.

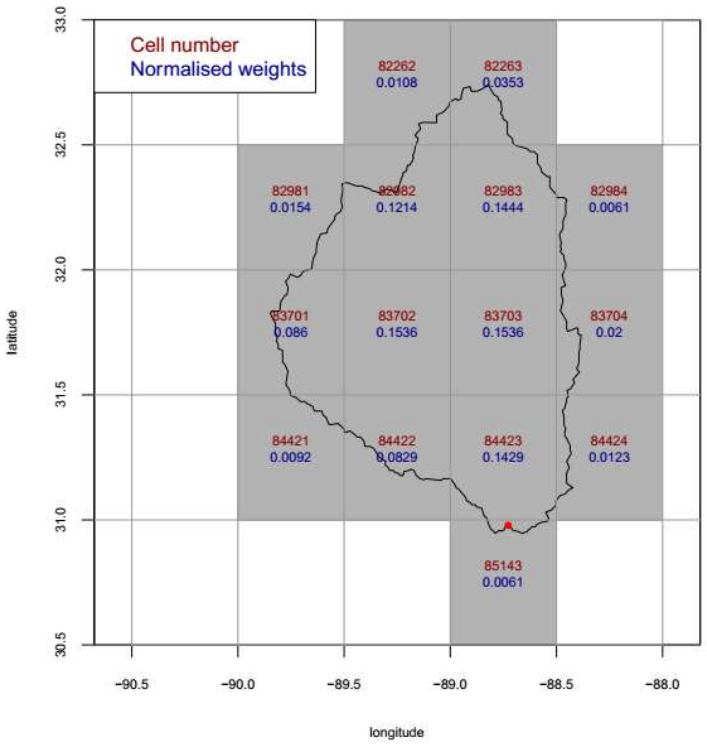
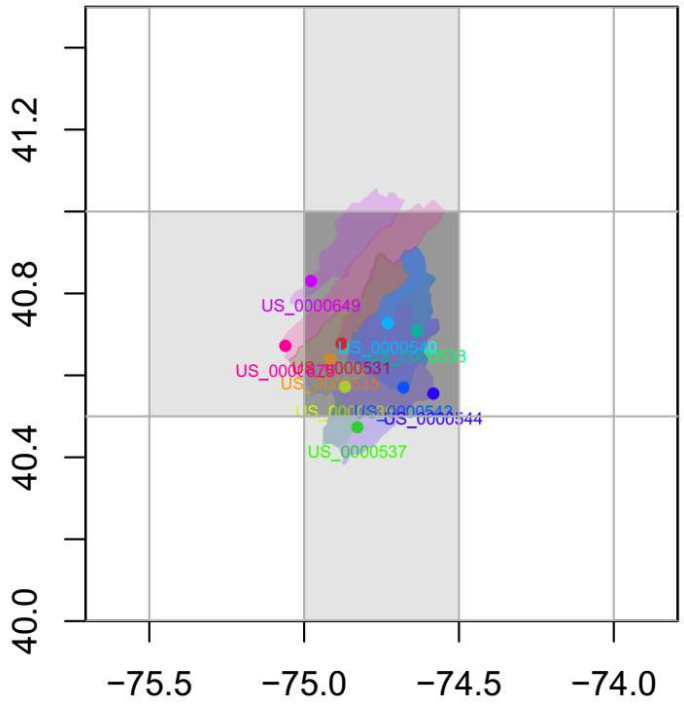


Figure S1. Illustration of the table of weights.

2.1.2 Averaging approach for cases where there were more than one catchment sharing similar weighted-area tables

Among catchments that have area less than 9000km^2, there are many instances where two or more catchments have (almost) identical simulated runoff as they have similar weighted-area tables. All ISIMIP models have a common assumption of uniform parameterisation for runoff generation in the 0.5×0.5 grid area, which in concept should represent an average value of runoff at finer resolution. Note that ORCHIDEE in ISIMIP2b (GCMs driven) was run at 1°×1° resolution, and the outputs were disaggregated evenly 0.5×0.5 resolution. Here we also treat catchments that intersect an identical set of dominant contributing grid-cells (total weights of at least 70%) as samples of an identical simulation domain. As a result, the area-weighted mean discharge of these catchments was calculated and used for model-observation comparison.

A search was conducted across all weighted-area tables to identify cases that have an identical set of intersecting cells contributing at least 70% to the total weighting. Figure S2 provides an example of these cases. In the top panel, boundaries of ten catchments were superimposed on top of the ISIMIP gridline (0.5×0.5 degree), demonstrating that they share a common cell (number 70051) which contributes at least 70% to the total weight (showed in the bottom panel).



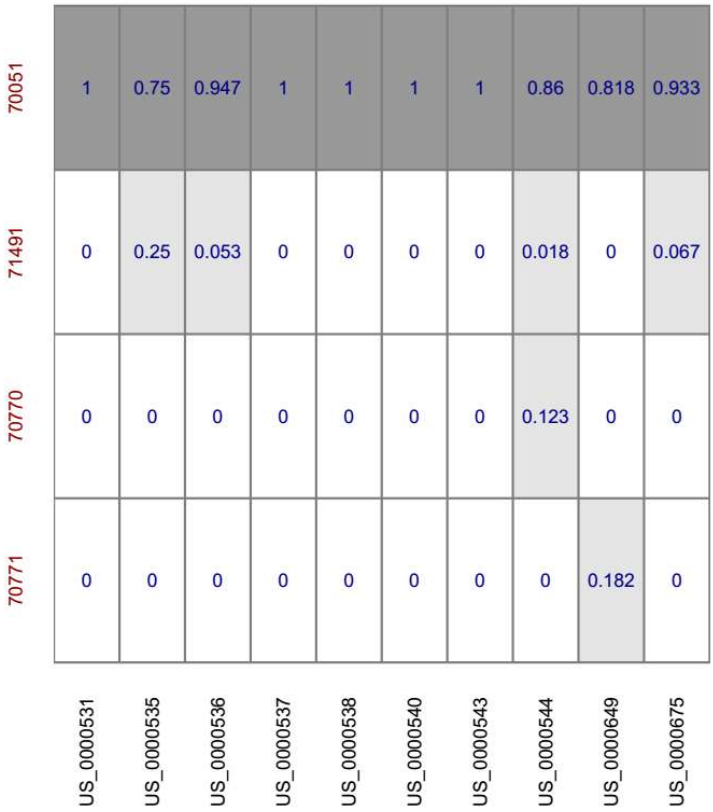
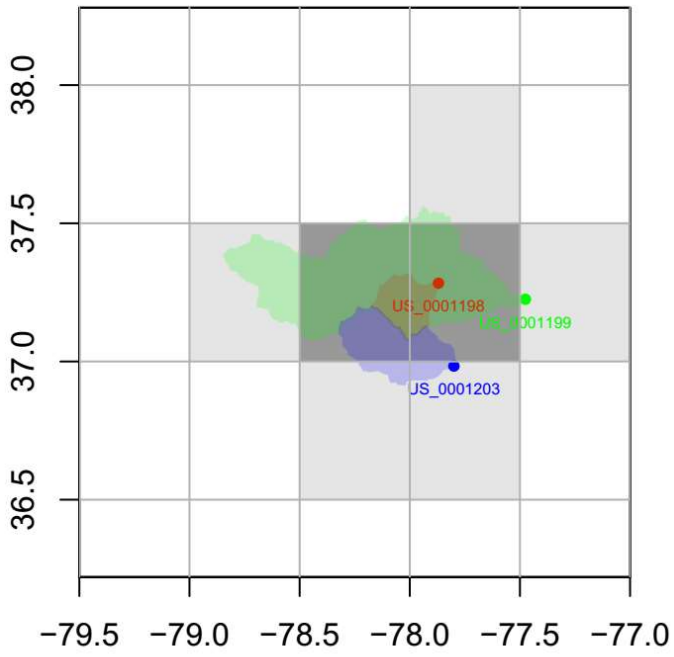


Figure S2. Example of instances where there is a significant overlap in contributing cells. Top panel: locations of 10 catchments that share a common contributing grid-cell (cell number 70051 (in dark-grey colour) contributes at least 70% to the total weight of each catchment) although specific catchments have different contributing cells. Bottom panel: weighted-area table of these 10 catchments.

Figure S3 illustrates another case where three different catchments share two common cells (no. 76524 and 76525). These cells contribute 100%, 79.1%, and 76.4% to the weighted-area tables of catchment US_0001198, US_0001199, and US_0001203 respectively. In both examples, the identified catchments were considered samples of the same modeling domain.



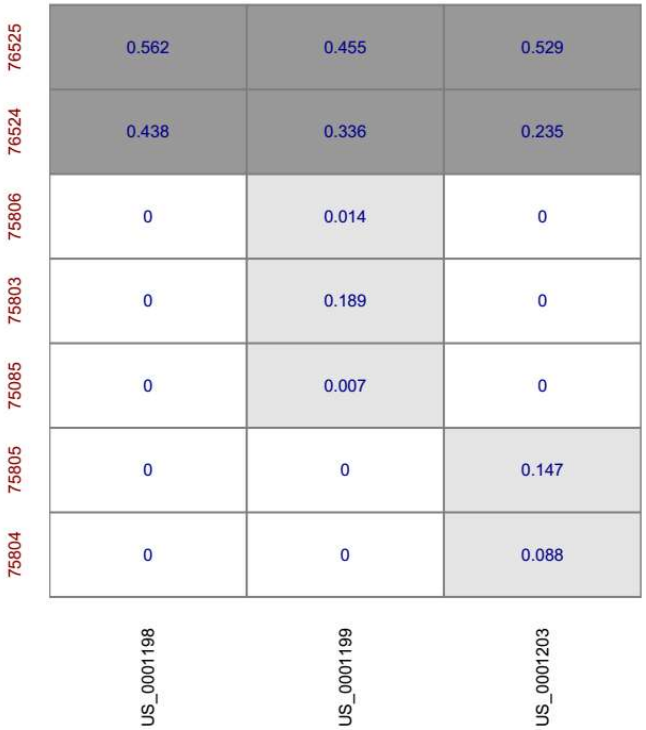


Figure S3. Similar to Figure S2, but here we have two contributing cells. The total weight of these common cells (number 76524 and 76525, highlighted in dark-grey colour) is higher than 0.7 in all cases and thus these three catchments were considered samples of the same modelling domain.

For each set of *n* catchments with similar weighted-area tables, a single average discharge $\bar{Q}$(m^3/s) was calculated to represent these individual time series in the model-observation comparison following below procedures:

For observed discharge:

1. Convert discharge Q (units: m^3/s) to runoff rate R (units: m/day) using catchment area A (units: m^2) for each catchment *i*.
$$R_i = Q_i \times 24 \times 3600/A_i \quad \text{(m/day)}$$
Average catchment size was also recorded:
$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i \quad \text{(m}^2\text{)}$$
2. Average runoff rate across all catchments (considering area-weights)
$$\bar{R} = \frac{\sum_{i=1}^{n} R_i A_i}{\sum_{i=1}^{n} A_i} \quad \text{(m/day)}$$
3. Back-calculate average discharge (m^3/s):
$$\bar{Q} = \frac{\bar{R}\bar{A}}{24\times3600} \quad \text{(m}^3\text{/s)}$$

For simulated discharge:

1. Extract runoff rate using weighted-area tables as described in Section 2.1 for all catchments.
2. Follow Step 2 and Step 3 of the observation procedure.

2.3 Discharge output identification for catchment with area greater than 9000 km^2

For catchments with area greater than 9000km^2, the 'discharge output' approach was adopted to find GHM cells corresponding to the catchments following Zhao et al. (2017). For a specific catchment, the grid cell corresponding to the catchment outlet was identified by matching catchment area available in a 0.5° drainage direction map (DDM30 dataset, freely available at http://www.uni-

frankfurt.de/45218101/DDM30) and the reported area. The identified grid cell was then used to extract simulated discharge available in the ISIMIP data repository. Stations were removed if the procedure could not identify any DDM30 grid cell surrounding the reported geographical location with a drainage area discrepancy less than 30% (see supplementary of Zhao et al. (2017) for detail).

3 Supplementary Figures

3.1 Capacity of GHMs to reproduce observed trends at continental scale

As stream gauges are not evenly distributed across the world, Figure S4 provides a zoomed-in map for four regions with relatively high number of stations (North America, Europe, South America, and Oceania). The most notable feature is a significantly lower strength of trends exhibited through GSWP3/GCMHIND ensemble average compared to GSIM observed trends. This pattern is likely the result of averaging technique (smoothed out variability of ensemble members) as the feature is more pronounced in GCMHIND (21 simulations) compared to GSWP3 (6 simulations). Visual inspection of these results suggests that the overall spatial pattern of observed trends seems to be preserved in GSWP3 while GCMHIND simulations tend to incorrectly simulate some spatial pattern of trends (e.g. over Oceania).

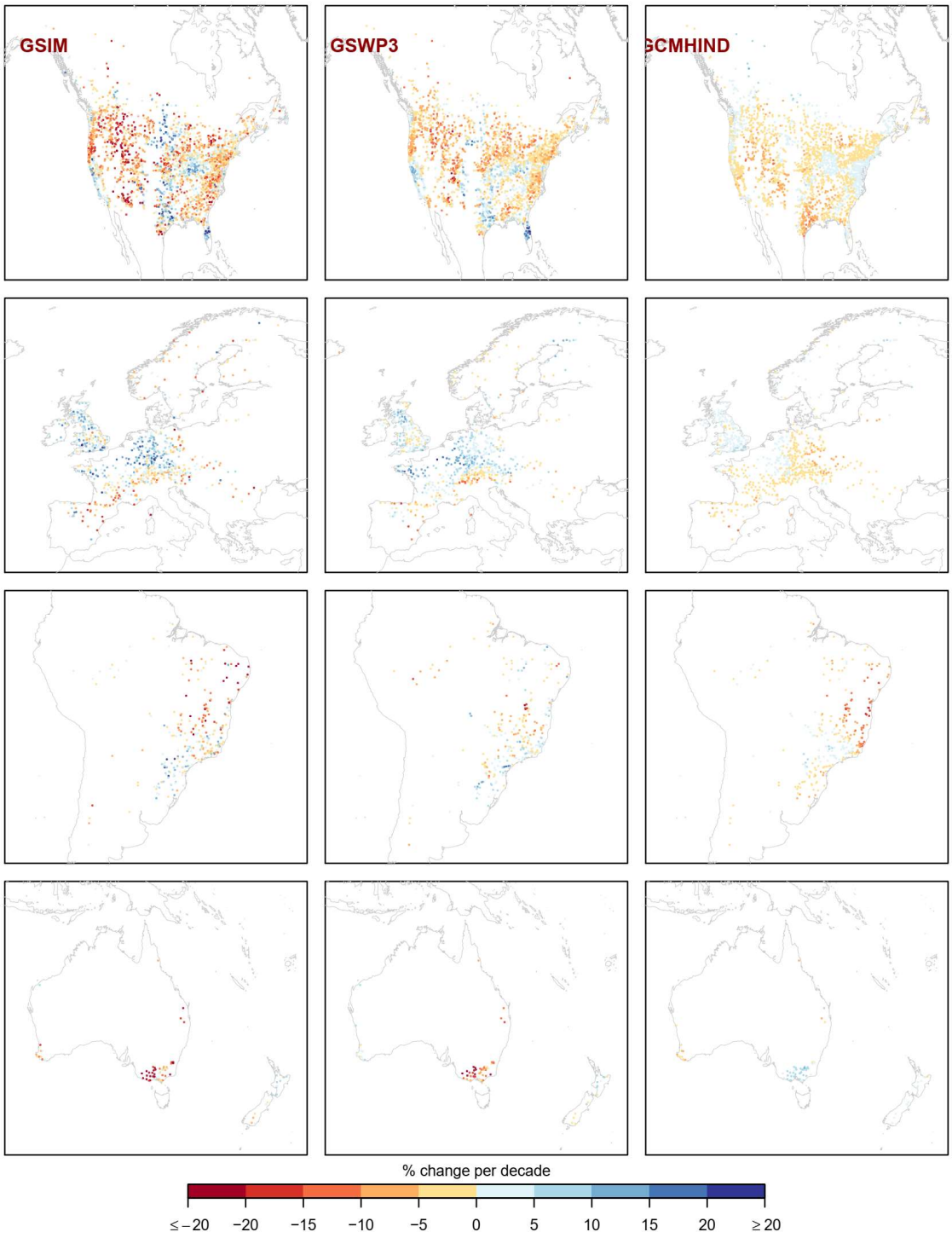
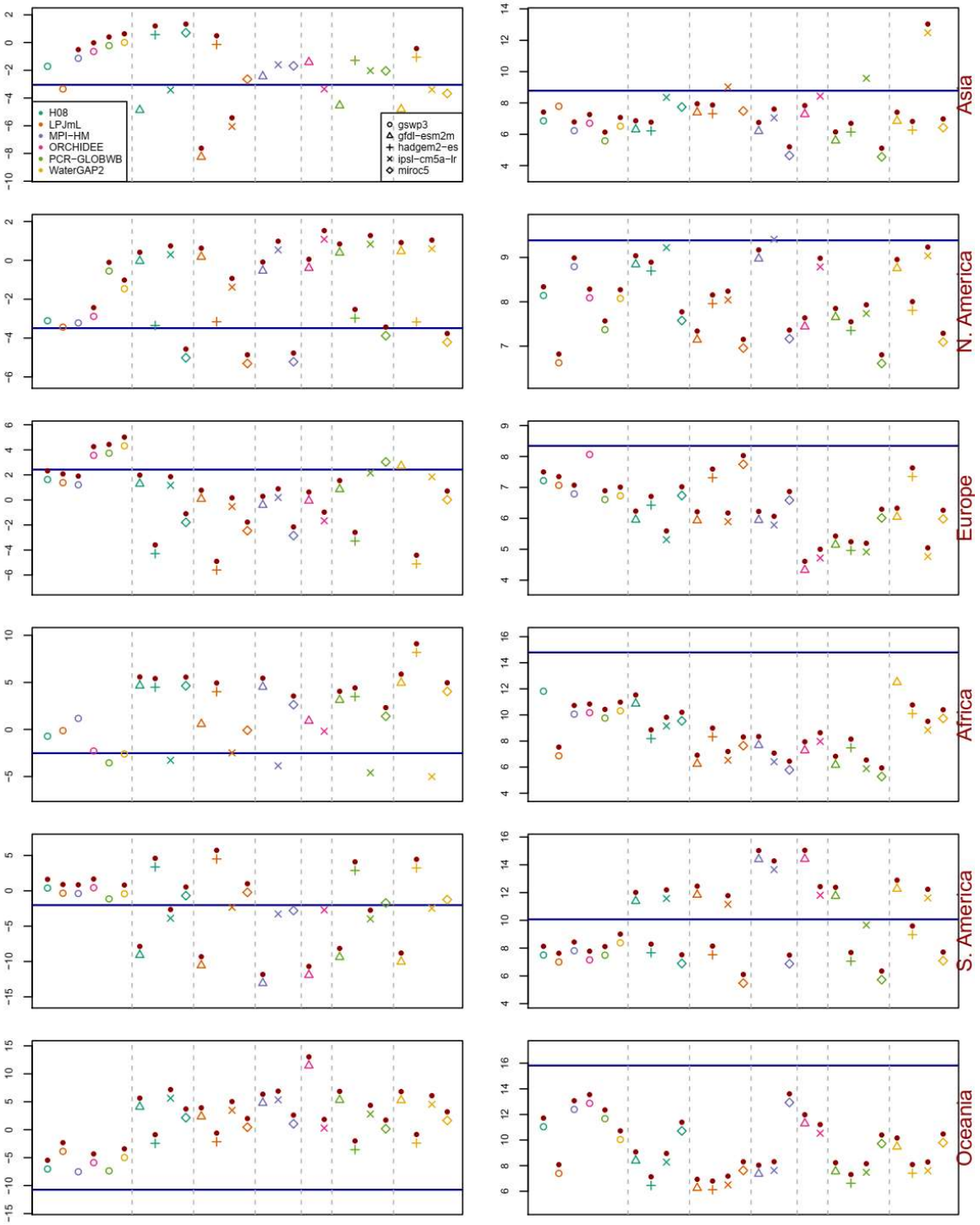


Figure S4. Normalised Theil-Sen slope for historical trends in flood magnitude (MAX7 index) over South America, Europe, South America and Oceania (left panels: GSIM; middle panels: GSWP3; right panels: GCMHIND). Multi-model average is shown for simulated trends. Trend is expressed in % change per decade.

Figure S5 illustrates the mean and standard deviation of simulated trends across all locations (% change per decade) for each individual ensemble member (multi-model average was showed in the manuscript). The mean and standard deviation of all trends (referred to as trend mean and trend standard deviation here-after) obtained from GSIM archive were also showed as dark blue line. GSWP3 simulations generally produced a higher trend mean and a lower trend standard deviation across all continents compared to the observed trends. The discrepancy varies substantially across different regions. For instance, Oceania

exhibited a discrepancy up to 7% per decade for the trend mean and 8% per decade for the trend standard deviation. This feature indicates a substantial inconsistency between simulated trends and observed trends. Among the six GHMs, ORCHIDEE, PCR-GLOBWB and ~~WaterGAP~~WaterGAP2 tend to have a higher trend mean with the exception of Africa. This pattern potentially indicates the influence of either (i) parameterisation, (ii) model capacity in reproducing observed trend characteristics, or (iiI) a bias of the GSWP3 forcing trends.

Figure S5 also shows relatively lower capacity of GCMHIND simulation in terms of reproducing observed trend mean and trend standard deviation in streamflow maxima. There is no clear ranking pattern in terms of the modelled atmospheric forcing being used, suggesting that uncertainty in GCM model was inherited differently across GHMs, likely due to the variation of parameterisation strategies.



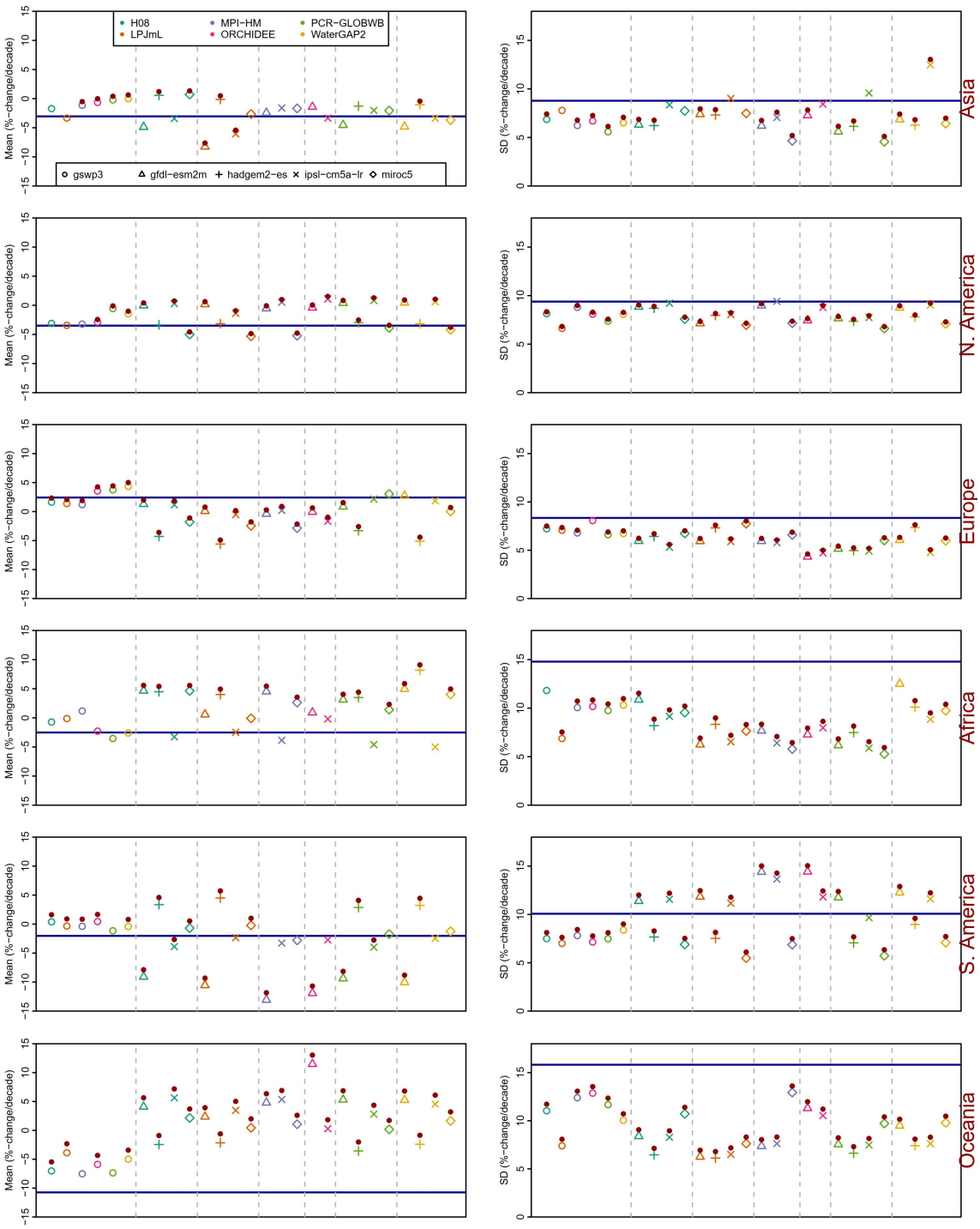


Figure S5. Mean (left panels) and standard deviation (right panels) of trends (% change per decade) exhibited from GSIM (horizontal blue line) observed trends and GSWP3/GCMHIND (hollow dots) simulated trends at the continental scale. The x-axis indicates different models. Note that y-axis range varies across panels. A null-hypothesis test was conducted to assess whether the mean/standard deviation of simulated trends is statistically different to that obtained from observed GSIM trends (horizontal blue line). Dark-red filled dots indicate simulations rejecting the null-hypothesis (i.e. which is that simulated trend mean/trend standard deviation is not statistically different to that obtained from GSIM).

3.2 Spatial uncertainty across simulated trends forced with different modelled atmospheric forcing

The assessment in section 3.3 of the main text suggests the combined GCM-GHM uncertainty has led to the presence of high uncertainty in terms of regions with significant projected trends in streamflow extremes. That is, a region could be projected by an overall increasing trend by one member and a decreasing trend by another member. This feature is illustrated in Figure S6, which shows a notable mismatch in the spatial structure of projected trends in MAX7 index between two ensemble members. Under the RCP2.6 greenhouse gas emission scenario, H08 forced with GFDL-ESM2M (top panels) projects an increasing trend for the majority of Australia and Siberia, while ORCHIDEE forced with IPSL-CM5A-LR (bottom panels) projects an overall decreasing trend for the same regions. This spatial uncertainty could come from either the climate trends introduced by GCMs (differentiate across GCMs), different RCPs, and model characteristics.

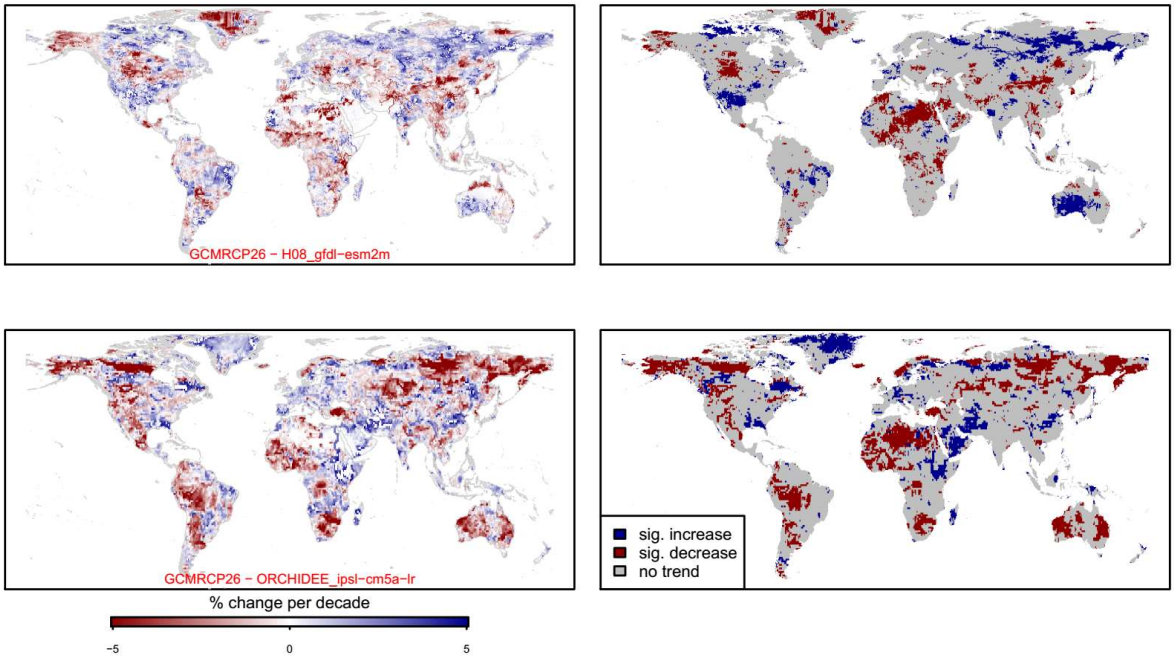


Figure S6. The magnitude (left panels) and significance (right panels) of trends in simulated MAX7 time series across all grid cells under RCP26 greenhouse gas emission scenario (2006-2099). Top panels: H08 forced with gfdl-esm2m climate data; bottom panels: ORCHIDEE forced with ipsl-cm5a-lr climate data. These two models had the lowest value of pattern similarity (correlation of -0.17).

3.3 Potential influence of model versions on detected trends

As mentioned in section 1.3 of this supplementary, there are changes in model versions that were used in two phases of ISIMIP. 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).

While some models undergone minor changes (e.g., changes and bug-fixes done in MPI-HM affect only the human impact simulations – and the influence is insignificant), the different versions of the other models might lead to substantial differences of simulated trends. Within the context of this study, we managed to compare trends simulated by two versions of WaterGAP2 (Figure S7), and the influence of model versions to trends seem minor. However, not all simulations for the other models are readily available, thus the influence of model versions to the results cannot be explicitly identified in this study.

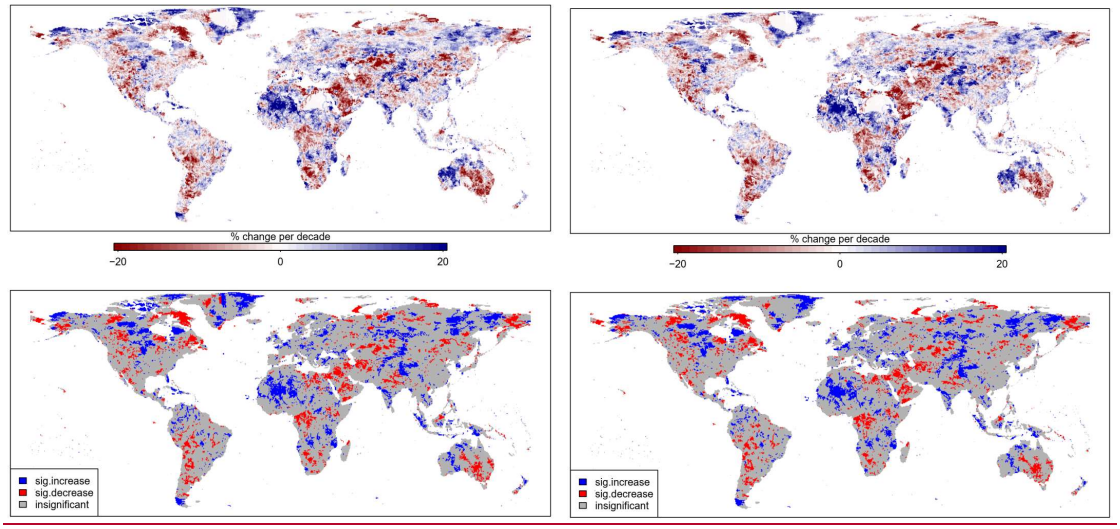


Figure S7. The magnitude (top panels) and significance (lower panels) of historical trends (1971-2005) in simulated MAX7 time series across all grid cells using two versions of WaterGAP. Left panels: WaterGAP2.2 (ISIMIP2a) which was used in ISIMIP2a; right panels: WaterGAP2.2c which was used in ISIMIP2b. Both simulations were forced with GSWP3 observed climate data.

4 Supplementary Tables

Considering a large number of simulations available (73 in total), the main text mostly used multi-model min/max/average to illustrate the results for cases where there is more than one simulation available for an identical GHM/spatial-domain. Table S2 provides a list of all 73 available models reported in this section together, with their simulation settings. Note that:

(i) GSWP3_VARSOC simulations (listed in Table S2 as H08_GSWVAR, LPJ_GSWVAR, PCR_GSWVAR, and WAT_GSWVAR) were not reported in the main text as (1) there were only four simulations available (comparing to six simulations of GSWP3_NOSOC) and (2) the results obtained from GSWP3_NOSOC and GSWP3_VARSOC are similar (Table S3).

(ii) In the main text, OBSHIS_NOSOC simulations were denoted as GSWP3.

Table ~~S2~~S3. Available ISIMIP streamflow simulations and associated setting.

Seq	Streamflow simulations	GHM	Climate	Human	Period
1.	H08_GSWVAR	H08	Observation (GSWPv3)	varsoc	1971-2005
2.	H08_GSWNO		Observation (GSWPv3)	nosoc	
3.	H08_HIN_G		HINDCAST (GFDL-ESM2M)	2005soc	
4.	H08_HIN_H		HINDCAST (HadGEM2-ES)		
5.	H08_HIN_I		HINDCAST (IPSL-CM5A-LR)		
6.	H08_HIN_M		HINDCAST (MIROC5)		
7.	H08_RCP2.6_G		RCP2.6 (GFDL-ESM2M)		2006-2099
8.	H08_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
9.	H08_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
10.	H08_RCP2.6_M		RCP2.6 (MIROC5)		
11.	H08_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
12.	H08_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
13.	H08_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
14.	H08_RCP6.0_M		RCP6.0 (MIROC5)		

15.	LPJ_GSWVAR	LPJmL	Observation (GSWPv3)	varsoc	1971-2005
16.	LPJ_GSWNO		Observation (GSWPv3)	nosoc	
17.	LPJ_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
18.	LPJ_HIN_H		HINDCAST (HadGEM2-ES)		
19.	LPJ_HIN_I		HINDCAST (IPSL-CM5A-LR)		
20.	LPJ_HIN_M		HINDCAST (MIROC5)		
21.	LPJ_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
22.	LPJ_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
23.	LPJ_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
24.	LPJ_RCP2.6_M		RCP2.6 (MIROC5)		
25.	LPJ_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
26.	LPJ_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
27.	LPJ _RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
28.	LPJ_RCP6.0_M		RCP6.0 (MIROC5)		
29.	MPI_GSWNO	MPI-HM	Observation (GSWPv3)	nosoc	1971-2005
30.	MPI_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
31.	MPI_HIN_I		HINDCAST (IPSL-CM5A-LR)		
32.	MPI_HIN_M		HINDCAST (MIROC5)		
33.	MPI_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
34.	MPI_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
35.	MPI_RCP2.6_M		RCP2.6 (MIROC5)		
36.	MPI_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
37.	MPI_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
38.	MPI_RCP6.0_M		RCP6.0 (MIROC5)		
39.	ORC_GSWNO	ORCHIDEE	Observation (GSWPv3)	nosoc	1971-2005
40.	ORC_HIN_G		HINDCAST (GFDL-ESM2M)		
41.	ORC_HIN_I		HINDCAST (IPSL-CM5A-LR)		
42.	ORC_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	nosoc (land use changes was considered)	2006-2099
43.	ORC_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
44.	ORC_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
45.	ORC_RCP6.0_G		RCP6.0 (IPSL-CM5A-LR)		
46.	PCR_GSWVAR	PCR-GLOBWB	Observation (GSWPv3)	varsoc	1971-2005
47.	PCR_GSWNO		Observation (GSWPv3)	nosoc	
48.	PCR_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
49.	PCR_HIN_H		HINDCAST (HadGEM2-ES)		
50.	PCR_HIN_I		HINDCAST (IPSL-CM5A-LR)		
51.	PCR_HIN_M		HINDCAST (MIROC5)		
52.	PCR_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
53.	PCR_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
54.	PCR_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
55.	PCR_RCP2.6_M		RCP2.6 (MIROC5)		
56.	PCR_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
57.	PCR_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
58.	PCR_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
59.	PCR_RCP6.0_M		RCP6.0 (MIROC5)		
60.	WAT_GSWVAR	WaterGAP2	Observation (GSWPv3)	varsoc	1971-2005
61.	WAT_GSWNO		Observation (GSWPv3)	nosoc	
62.	WAT_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
63.	WAT_HIN_H		HINDCAST (HadGEM2-ES)		
64.	WAT_HIN_I		HINDCAST (IPSL-CM5A-LR)		
65.	WAT_HIN_M		HINDCAST (MIROC5)		
66.	WAT_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
67.	WAT_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
68.	WAT_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
69.	WAT_RCP2.6_M		RCP2.6 (MIROC5)		
70.	WAT_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		

71.	WAT_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
72.	WAT_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
73.	WAT_RCP6.0_M		RCP6.0 (MIROC5)		

Most results of the main text only showed the multi-model average for GCMHIND simulations of each GHM (up to four simulations per GHM) (e.g. Table 3 of the main text, which presents the characteristics of trends in the MAX7 index over 1971-2005 period across 3666 locations globally). The following tables, therefore, provide the results of each experiment at the global scale for individual models to complement the key findings, in which:

- Table S3 (adapted from Table 2 in the main text) describe the hypothesis tests.
- Table S4 and S5 report trend mean/standard deviation, percentage of locations exhibiting significant trends and the correlation of simulated trends against observed trends (historical period from 1971 to 2005). The results of hypothesis test (described in Table S3) are also highlighted in Table S4 and Table S5.
- Tables S6 and S7 report the value of simulated trend mean/trend standard deviation and the percentage of cells exhibiting significant trends for future period (2006-2099). Note that the statistical test described in Table S3 was not adopted for these results.

As noted in the main text, trends in peak discharge exhibited from 'naturalised runs' (GSWP3_NOSOC) are similar to those obtained from 'human impact runs' (GSWP3_VARSOC). This is specifically illustrated through Table S4, in which the trends characteristic are quite similar between two settings. For instance, PCR_GSWVAR suggests a global trend mean (standard deviation) of 0.0 (7.7) % change per decade, with a spatial correlation against observed trends of 0.5. These results are very similar to that reported for PCR_GSWNO.

Table ~~S3~~S4. Summary of the hypothesis tests conducted to address the first two objectives. The significance of these tests was reported in Table S4 and S5.

Objective	Null-Hypotheses	Streamflow dataset	Statistical tests
Objective 1: Capacity of GHMs to reproduce observed trends in flood hazards	Hypothesis 1: Trend means obtained from two streamflow datasets over observation locations were not statistically different from each other.	(i) Observed discharge across 3,666 observation locations (ii) Simulated discharge across 3,666 observation locations (extraction processes outlined in Section 2)	Two-sample *t*-test at the 10% two-sided significance level
	Hypothesis 2: Trend standard deviations obtained from two streamflow datasets over observation locations were not statistically different from each other.		Two-variance *F*-test at the 10% two-sided significance level
	Hypothesis 3: Percentage of significant trends obtained from all observation locations of a specific streamflow dataset was not produced by random chance.		Field significance test similar to that presented in Do et al. (2017) was adopted. A moving-block-bootstrap (block-length $L = 2$) was used to derive a null-hypothesis distribution of the change that occurred due to random chance. The null hypothesis is rejected at 5% one-sided significance level when the true percentage falls on the right-hand side of the 95th percentile of the resampled distributions.
	Hypothesis 4: The correlation between trends obtained from two streamflow datasets was not significantly higher than '0' (i.e. zero pattern similarity).		'Zero pattern similarity' was compared to the probability distribution function (PDF) of pairwise correlation between simulated and observed trends, drawn from a bootstrap procedure similar to that proposed by Kiktev et al. (2003). The null hypothesis is rejected at 5% one-sided significance level when zero correlation falls on the left-hand side of the 5th percentile of the resampled distributions.
	Hypothesis 5: The correlation between GCMHIND simulated trends and observed trends was not significantly lower than the correlation between GSWP3 simulated trends and observed trends		The actual pairwise correlation between GCMHIND simulated trends and observed trends (denoted by $r_{GCMHIND}$) was compared to the bootstrapped PDF of correlation exhibited from GSWP3 simulated trends (denoted by r^*_{GSWP3}). If $r_{GCMHIND}$ falls on the left-hand side of the 5th percentile r^*_{GSWP3}, there is evidence to reject the null-hypothesis at the 5% one-sided significance level.
Objective 2: The representativeness of observation locations in the GHM simulations	Hypothesis 6: Trend mean obtained from observation locations was not statistically different to that obtained from all grid cells.	(i) Simulated discharge across 3,666 observation locations (extraction processes outlined in Section 2)	Two-sample *t*-test at the 10% two-sided significance level
	Hypothesis 7: Trend standard deviation obtained from observation locations was not statistically different to that obtained from all grid cells.		Two-variance *F*-test at the 10% two-sided significance level

	Hypothesis 8: Percentage of significant trends obtained from all grid cells of a specific streamflow dataset was not produced by random chance.	(ii) Routed discharge across all landmass grid cells (59,033 cells)	Field significance test similar to that presented in Hypothesis 3 but trends obtained from all grid cells were the subject of the assessment.

Table ~~S4~~S5. Characteristics of trends in the MAX7 index (introduced by GHMs) over the 1971-2005 period averaged across the 3666 locations. Trend mean and trend standard deviation have units of %-change per decade. Gauge-based significant trends were identified using a Mann-Kendall test (10% two-sided significance level). The global significance of this result is then calculated using field significance test (5% one-sided significant level; highlighted in boldface text). Trend mean, trend standard deviation and trend spatial structure were compared against that exhibited by GSIM (see Hypothesis 1 to hypothesis 5 of Table S3 for description of hypothesis tests; significant values were represented in boldface text).

Streamflow simulations	Trend mean	Trend standard deviation	Percentages of significant		Correlation against observed trends
			Increasing trend	Decreasing trend	
H08_GSWVAR	-2.0	**8.3**	4.8	6.7	**0.4**
LPJ_GSWVAR	-2.6	**7.5**	4.6	9.2	**0.4**
PCR_GSWVAR	**0.0**	**7.7**	9.4	6.1	**0.5**
WAT_GSWVAR	**-0.7**	**8.5**	8.4	5.8	**0.5**
H08_GSWNO	**-1.9**	**8.3**	4.8	6.7	**0.4**
LPJ_GSWNO	-2.2	**7.1**	4.5	7.3	**0.4**
ORC_GSWNO	**-1.4**	**8.6**	7	8.2	**0.4**
MPI_GSWNO	-2.1	**8.7**	5.6	7.5	**0.5**
PCR_GSWNO	**0.1**	**7.7**	9.6	6.1	**0.5**
WAT_GSWNO	**-0.3**	**8.2**	8.5	4.2	**0.5**
H08_HIN_G	**-0.4**	**8.9**	6.1	7.8	**0.1**
H08_HIN_H	**-2.8**	**8.4**	2.2	10.8	-0.1
H08_HIN_I	**0.1**	**8.9**	7.7	4.4	0.0
H08_HIN_M	**-3.6**	**7.8**	3.4	12.0	**0.1**
LPJ_HIN_G	**-0.8**	**8.0**	6.3	8.3	**0.1**
LPJ_HIN_H	**-2.9**	**8.1**	2.8	**14.6**	0.0
LPJ_HIN_I	**-1.3**	**8.0**	4.1	10.1	**0.1**
LPJ_HIN_M	**-4.1**	**7.3**	3.5	**17.3**	**0.2**
ORC_HIN_G	**-0.9**	**8.6**	5.2	7.6	0.0
ORC_HIN_I	**0.1**	**8.6**	8.6	6.4	**0.1**
MPI_HIN_G	**-1.3**	**9.5**	5.9	7.9	**0.1**
MPI_HIN_I	**0.2**	**9.2**	8.8	5.6	0.0
MPI_HIN_M	**-4.2**	**7.3**	2.3	**16.3**	**0.1**
PCR_HIN_G	**-0.2**	**8.0**	8.3	9.0	**0.1**
PCR_HIN_H	-2.5	**7.1**	2.7	11.0	0.0
PCR_HIN_I	**0.6**	**7.6**	12.2	4.1	0.0
PCR_HIN_M	-2.1	**7.0**	6.9	13.5	**0.1**
WAT_HIN_G	**0.2**	**9.2**	8.2	5.6	**0.1**
WAT_HIN_H	**-2.9**	**8.1**	2.7	10.9	-0.1
WAT_HIN_I	**0.5**	**8.8**	6.2	4.2	-0.1
WAT_HIN_M	**-2.9**	**7.3**	4.3	11.4	**0.1**

Table ~~S5~~S6. Trend mean, trend standard deviation and percentage of significant trends averaged across all simulation grid cells. Trend mean and trend standard deviation have units of %-change per decade. Cell-based significance was identified using the Mann-Kendall test at the 10% significance level. The global significance of this result is then calculated using field significance test at 5% one-sided level (highlighted in boldface text). Trend mean and trend standard deviation across all land mass were compared against that obtained across 3666 observation locations (reported in Table S4) and significant values are highlighted in boldface text (see Hypothesis 6 to hypothesis 8 of Table S3 for description of hypothesis tests).

Streamflow simulations	Trend mean	Trend standard deviation	Percentages of significant	
			Increasing trend	Decreasing trend
H08_GSWVAR	**-0.5**	**10.1**	8.4	**10.7**
LPJ_GSWVAR	**-1.6**	**10.4**	7.2	**14.0**
PCR_GSWVAR	**-1.1**	**11.0**	10.4	**15.0**
WAT_GSWVAR	**-0.3**	**11.4**	10.8	**11.0**
H08_GSWNO	**-0.3**	**9.9**	8.3	9.6
LPJ_GSWNO	**-0.9**	**9.9**	7.4	**11.5**
ORC_GSWNO	**-0.9**	**9.6**	6.1	**7.8**
MPI_GSWNO	**-0.7**	**10.2**	6.4	**7.5**
PCR_GSWNO	**-1.0**	**10.9**	10.7	**14.7**
WAT_GSWNO	**0.0**	**11.1**	**10.9**	10.1
H08_HIN_G	**1.5**	**10.8**	**15.4**	10.4
H08_HIN_H	**0.0**	**8.5**	7.4	9
H08_HIN_I	**-0.7**	**9.3**	7	**10.7**
H08_HIN_M	**0.4**	**8.9**	8.7	8
LPJ_HIN_G	**-0.3**	**9.3**	8.9	9.1
LPJ_HIN_H	**-1.1**	**8.7**	5.1	**9.9**
LPJ_HIN_I	-1.1	**8.7**	6.1	9.2
LPJ_HIN_M	**-0.8**	**9.1**	7.7	9.4
ORC_HIN_G	**0.6**	9.5	**8.4**	6.3
ORC_HIN_I	**-0.9**	**8.2**	3.9	**6.8**
MPI_HIN_G	**-0.1**	7.3	4.5	5
MPI_HIN_I	**-0.2**	**10.3**	**10.9**	**11.2**
MPI_HIN_M	**-1.4**	**9.3**	5.5	**11.1**
PCR_HIN_G	**1.3**	**11.3**	**14.9**	11.1
PCR_HIN_H	**-0.4**	**8.7**	8.1	10.5
PCR_HIN_I	**-1.3**	**10.7**	7.7	**12.2**
PCR_HIN_M	**0.4**	**9**	11.7	9.9
WAT_HIN_G	**1.5**	**10.9**	**15.3**	7.2
WAT_HIN_H	**0.0**	**9.1**	6.3	7.3
WAT_HIN_I	**0.0**	**9.4**	6.9	7.5
WAT_HIN_M	**0.4**	**9.7**	**10.8**	7.2

Table ~~S6~~S7. Characteristics of projected trends (GCMRCP2.6) across 18 members at the global scale. Mean and standard deviation have unit of %-change per decade. Note that no statistical test was conducted.

Streamflow simulations	Trend mean	Trend standard deviation	Percentages of significant	
			Increasing trend	Decreasing trend
H08_RCP2.6_G	0.0	2.1	10.9	9.6
H08_RCP2.6_H	0.4	2.7	18.0	11.0
H08_RCP2.6_I	0.0	2.3	11.5	14.2
H08_RCP2.6_M	0.0	2.8	16.2	11.6
LPJ_RCP2.6_G	-0.1	1.8	7.5	7.4
LPJ_RCP2.6_H	0.0	2.1	10.7	10.6
LPJ_RCP2.6_I	-0.1	2.1	9.1	10.6
LPJ_RCP2.6_M	0.0	2.2	12.6	9.0
ORC_RCP2.6_G	-0.3	2.3	9.0	13.9
ORC_RCP2.6_I	-0.6	2.9	9.2	21.2
PCR_RCP2.6_G	0.1	2.1	11.0	9.0
PCR_RCP2.6_H	0.3	2.3	16.6	11.2
PCR_RCP2.6_I	0.0	2.8	15.5	13.9
PCR_RCP2.6_M	0.1	2.5	17.4	12.4
WAT_RCP2.6_G	0.0	2.1	9.6	7.1
WAT_RCP2.6_H	0.4	2.2	14.1	7.5
WAT_RCP2.6_I	0.2	2.3	12.3	10.0
WAT_RCP2.6_M	0.2	2.4	16.1	7.3

Table ~~S7~~S8. Characteristics of projected trend (GCMRCP6.0) across 18 members at the global scale. Trend mean and trend standard deviation have unit of %-change per decade. Note that no statistical test was conducted.

Streamflow simulations	Trend mean	Trend standard deviation	Percentages of significant	
			Increasing trend	Decreasing trend
H08_RCP6.0_G	0.3	3.0	19.7	17.1
H08_RCP6.0_H	0.7	4.0	27.2	18
H08_RCP6.0_I	-0.4	3.4	15.3	27.1
H08_RCP6.0_M	0.4	3.3	26.2	14.9
LPJ_RCP6.0_G	-0.1	2.6	17.5	15.7
LPJ_RCP6.0_H	-0.2	3.4	22.3	21.9
LPJ_RCP6.0_I	-0.6	3.1	14.0	24.8
LPJ_RCP6.0_M	0.1	3.0	22.6	16.2
ORC_RCP6.0_G	-0.3	3.0	16.4	21.1
ORC_RCP6.0_I	-1.3	4.1	12.3	35.0
PCR_RCP6.0_G	-0.1	3.0	18.9	18.7
PCR_RCP6.0_H	0.1	3.8	26.0	22.2
PCR_RCP6.0_I	-0.5	3.6	18.3	25.6
PCR_RCP6.0_M	0.5	3.0	27.7	14.4
WAT_RCP6.0_G	0.4	2.6	23.5	9.8
WAT_RCP6.0_H	0.7	3.2	29.6	10.7
WAT_RCP6.0_I	0.0	3.2	20.4	16.9

WAT_RCP6.0_M	0.8	3.1	30.1	9.6

Reference

Batjes, N. H.: ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2), ISRIC-World Soil Information, 2012.

Bondeau, A., Smith, P. C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., LOTZE-CAMPEN, H., Müller, C., and Reichstein, M.: Modelling the role of agriculture for the 20th century global terrestrial carbon balance, Global Change Biology, 13, 679-706, 2007.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531-5546, 10.1002/2014wr016532, 2015.

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

Döll, P., Müller Schmied, H., Schuh, C., Portmann, F. T., and Eicker, A.: Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites, Water Resources Research, 50, 5698-5720, 10.1002/2014wr015595, 2014.

Guimberteau, M., Ducharne, A., Ciais, P., Boisier, J.-P., Peng, S., De Weirdt, M., and Verbeeck, H.: Testing conceptual and physically based soil hydrology schemes against observations for the Amazon Basin, 2014.

Guimberteau, M., Zhu, D., Maignan, F., Huang, Y., Yue, C., Dantec-Nédélec, S., Ottlé, C., Jornet-Puig, A., Bastos, A., Laurent, P., Goll, D., Bowring, S., Chang, J., Guenet, B., Tifafi, M., Peng, S., Krinner, G., Ducharne, A., Wang, F., Wang, T., Wang, X., Wang, Y., Yin, Z., Lauerwald, R., Joetzjer, E., Qiu, C., Kim, H., and Ciais, P.: ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: model description and validation, Geosci. Model Dev., 11, 121-163, 10.5194/gmd-11-121-2018, 2018.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 2: Applications and assessments, Hydrol. Earth Syst. Sci., 12, 1027-1037, 10.5194/hess-12-1027-2008, 2008a.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 1: Model description and input meteorological forcing, Hydrol. Earth Syst. Sci., 12, 1007-1025, 10.5194/hess-12-1007-2008, 2008b.

Hanasaki, N., Yoshikawa, S., Pokhrel, Y., and Kanae, S.: A global hydrological simulation to specify the sources of water used by humans, Hydrol. Earth Syst. Sci., 22, 789-817, 10.5194/hess-22-789-2018, 2018.

Hunger, M., and Döll, P.: Value of river discharge data for global-scale hydrological modeling, Hydrology and Earth System Sciences Discussions, 12, 841-861, 2008.

Jägermeyr, J., Gerten, D., Heinke, J., Schaphoff, S., Kummu, M., and Lucht, W.: Water savings potentials of irrigation systems: global simulation of processes and linkages, Hydrol. Earth Syst. Sci., 19, 3073-3091, 10.5194/hess-19-3073-2015, 2015.

Mueller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F. T., Reinecke, R., and Riedel, C.: Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use, Hydrology and Earth System Sciences, 20, 2877-2898, 2016.

Schaphoff, S., Heyder, U., Ostberg, S., Gerten, D., Heinke, J., and Lucht, W.: Contribution of permafrost soils to the global carbon budget, Environmental Research Letters, 8, 014026, 2013.

Stacke, T., and Hagemann, S.: Development and validation of a global dynamical wetlands extent scheme, Hydrology and Earth System Science, 16, 2915-2933, 2012.

Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., Van Der Ent, R. J., De Graaf, I. E., Hoch, J. M., and De Jong, K.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geoscientific Model Development, 11, 2429-2453, 2018.

Wada, Y., Wisser, D., and Bierkens, M. F. P.: Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, Earth Syst. Dynam., 5, 15-40, 10.5194/esd-5-15-2014, 2014.

Zhao, F., Veldkamp, T. I., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., and Portmann, F. T.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environmental Research Letters, 12, 075003, 2017.