Thank you for inviting me to review paper: “The Dilemma of Including Hot Models in Climate Impact Studies: A Hydrological Study” by Asenjan et al. At the outset, I should mention that I am a climate researcher (rather than a hydrologist).

The authors address an issue of maximum concern to climate change research, that at present, the CMIP6 ensembles contain a few models with an especially high climate sensitivity (“hot models”). The concern is that such simulations may project impacts that are especially difficult to adapt to – causing alarm - yet, we may find in the decades ahead that these hot-running ESMs are unrealistic.

The authors focus, in particular, on testing the effect of especially warm ESMs on streamflow. I like this approach because the emphasis is on a major impact and of much concern i.e. flood risk.

We appreciate the reviewer for taking the time to review our paper and provide insightful recommendations for improvement. We are pleased to have the insights of a climate scientist in evaluating our work. In the following section, we have provided a detailed response addressing each of the reviewer's comments.

Based on the Abstract alone as a start, in some ways, I like the catchy title, but using the word Dilemma implies something unresolved. In fact, this paper provides definite findings. Maybe something a bit more factual such as e.g. : “Including Hot Climate Models in estimating Streamflow does not alter the assessment their future variability”.

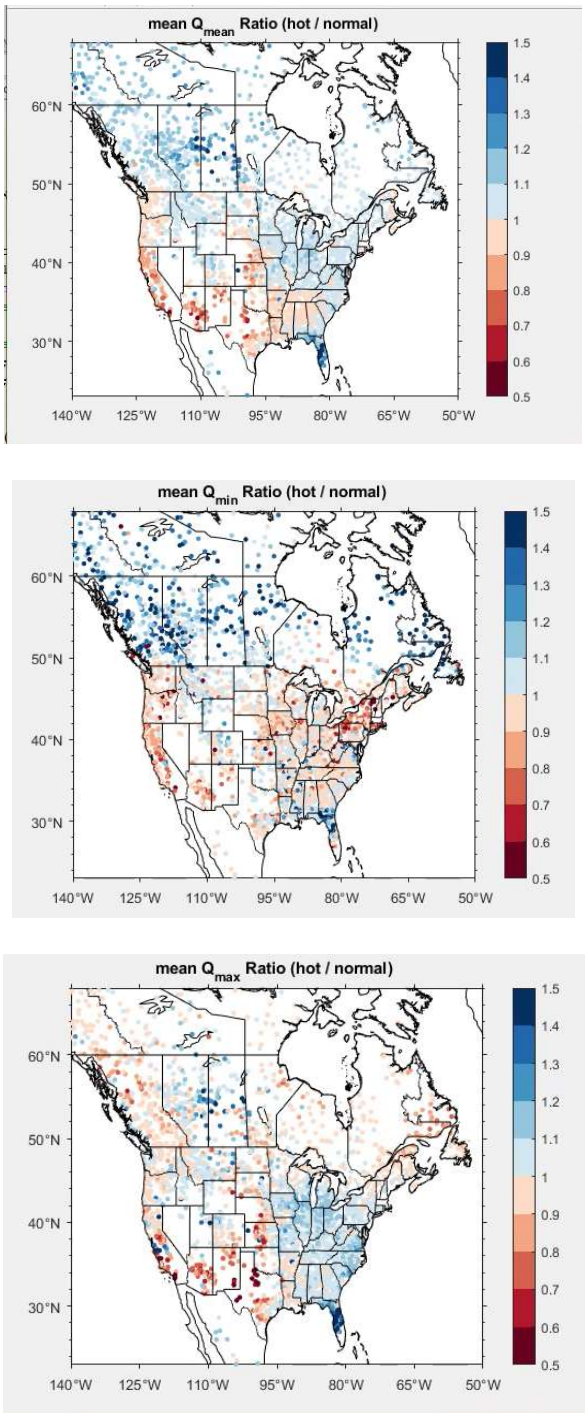
'Dilemma' may indeed be a bit too dramatic. We did not settle on your title suggestion, however, because including 'hot' models does affect variability in some regions, while in most others, it indeed has minimal effects. We have yet to settle on a final revised title, but here are the top candidates so far:

- Evaluating the Inclusion of 'Hot' Models in Climate Impact Studies: Insights from a Hydrological Study
- Assessing the Impact of Incorporating 'Hot' Models in Climate Impact Studies: A Hydrological Viewpoint
- Understanding the Influence of 'Hot' Models in Climate Impact Studies: A Hydrological Perspective

Second, in general and as the world warms, there will be a hydrological intensification for the planet. Hence, it might be expected that hot models generate higher river flows. The authors need to give a good reason for concentrating on variability, rather than mean trends. However, reading further into the manuscript, then mean changes e.g. “Qmean” are considered – maybe reword the Abstract to “mean changes and changes to variability”?

Indeed, our paper focuses on the potential to reduce uncertainty in future changes, as this is a significant issue for decision-makers. A high level of uncertainty can be a deterrent to implementing sound adaptation measures. If this uncertainty could be reduced in a scientifically sound manner, it would be extremely beneficial. However, the intensification of the hydrological cycle does not necessarily mean that higher flows will result, as evapotranspiration could also be significantly affected by the increase in mean temperature. The combined impact of both effects depends on geographical location and primary climate zones, and is not straightforward since, as shown in the paper, 'hot' models tend to also be 'wet.' In light of your comment, we propose to add the following figure, which shows the

actual changes as the ratio of the mean projected changes (‘hot’ models / normal models). Mean flows are indeed increasing over most of the study domain for 'hot' models, except in the south-west regions, where increased evapotranspiration nullifies potential increases in precipitation.



New Figure XX1: Ratio of mean projected changes: ‘hot’ divided by normal (not ‘hot’) models. Upper graph: Qmean; Middle graph (Qmin), Lower graph (Qmax). A red color indicates that hot models, on average, have lower flows (mean, minimum, maximum).

It is often asked: “Given the problem of climate change is now emerging strongly in the measurement record, why is it so difficult to constrain future warming estimates?”. The main reason for this is the historical record also includes a strong cooling aerosol effect. There are a few papers out there that make this point, and could be worth citing around line 45? In other words, we do not know from present-day measurements if we are in a high sensitivity fast warming world with strong contemporary aerosol cooling, or the opposite. That is the usual reason why we cannot reject outlier ESMs.

The reviewer has raised a very good point. We will incorporate this point, along with appropriate citations, into the revised version of the paper.

The description of the methods feels a bit too short in places, which is a shame because the authors have worked especially hard to entrain data, bias-correct a hydrological model against ERA5, bias-correct the ESMS. The test data is especially comprehensive (~14K sites). Would a schematic work that can capture some of this activity?

We appreciate the reviewer's comments. It’s always a fine line between having too little vs too much methodological information. In light of the comment, we will include a schematic to better describe the modeling procedure of this study, and we will add extra details to the methods section for a more comprehensive representation of our methods.

I had not heard of the KGE metric / statistic before, and where it is introduced, there is no citation. Can a reference be provided, and maybe a short sentence that explains its basic features? Presumably, the statistic allows an assessment of which hydrological models are performing best (based on aspects of variability in timeseries?).

Considering your background, it's probably not surprising that the KGE metric, which is rooted in hydrological science, is not familiar to you. Given the multidisciplinary nature of HESS, we will expand the text to provide additional details on this metric, along with references.

When calibrating and assessing hydrological models, performance criteria are generally used to quantify the degree of similarity between observed and simulated discharges. The Nash-Sutcliffe Efficiency (NSE, Nash & Sutcliffe, 1970), a normalized RMSE metric, has long been the standard for assessing model efficiency. However, it has a number of drawbacks, including those affecting bias and correlation, that have been addressed by the Kling-Gupta Efficiency metric (KGE) (Gupta et al., 2009). KGE has gained popularity over the past 10 years and is gradually replacing the NSE in literature dealing with the calibration of hydrological models. The KGE metric directly combines the bias, ratio of variance, and correlation into a single metric.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

where r is the linear correlation between observations and simulations, α a measure of the flow variability error, and β a bias term

$$KGE = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_{sim}}{\sigma_{obs}}-1)^2 + (\frac{\mu_{sim}}{\mu_{obs}}-1)^2}$$

where σ_{obs} is the standard deviation in observations, σ_{sim} the standard deviation in simulations, μ_{sim} the simulation mean, and μ_{obs} the observation mean.

Please also check the order of presentation. Figure 1 is cited in the text (line 84) before the modelling is described in full (around line 124).

We will take this point into account in the revised version of the manuscript and correct the order of the presentation.

There could be improvement of the manuscript around page 7 (lines 135-158, and Figure 2). Can it be confirmed that Figure 2 is a generic plot, referring to any metric of interest? Line 143 "Figure 2 present the three dispersion metrics…. streamflow". This is too vague – does the "y" axis of Figure 2 related to Qmean, Qmax or Qmin? Apologies if I am missing something obvious.

Figure 2 is indeed a generic plot referring to all the metrics considered in the manuscript. The "y" axis of Figure 2 represents the Streamflow metrics, e.g. Qmean, Qmax or Qmin, while the X axis represents the time. we will improve the explanation in the revised version to avoid any confusion. In addition, we will make minor modifications to the Figure.

In results, line 186, there is reference to ECS. However, neither here, or in the caption to Figure 3 (or caption to the Table) is a reference to the source of the ECS values. For Figure 3, you could potentially use some sort of marker to differentiate the five warmest models on the left. For instance, a horizontal arrow under the first five models, with the words "warmest models".

Correct, this point was also mentioned by the first reviewer, and we will include the references to the source of ECS values in the revised version.

We will also edit figure 3 to distinguish the hot models from the rest.

In places I thought the captions to diagrams and tables were very short. Obviously, captions cannot repeat everything that is in the paper text, but a little more information in places might help the reader (this is important if, upon publication, people extract diagrams and captions to place in powerpoints for talks).

We acknowledge that in some instances, the captions may rely too heavily on the description provided in the main text. In the revised version of the paper, we will ensure that we include more details in the figure and table captions. This will allow each figure or table to be understood independently

This might be something for the HESS editor to advise on, but in my view, the main plot of the paper is Figure 5. This presents very clearly the geographical effect of removal of models with a high ECS. After that, there are many helpful diagrams, but I am not convinced that all are needed. Or some sort of multi-panel plot might help, e.g. merging the box plots. The authors could consider placing some of the additional information and plots after that in an SI – e.g. Figure 12.

We appreciate the reviewer's suggestion. We agree that relocating some figures to the supporting information could help emphasize the main points of our paper. However, we also believe that retaining some figures, such as those beyond figure 5, is necessary to effectively communicate the crux of our research. For instance, figure 5 displays the total spread ratio, a measure of the data range that is heavily influenced by outliers. Conversely, figure 9 illustrates the standard deviation ratio, a measure of data dispersion. We recognize the similarity in the patterns shown in Figures 7 and 9, so much so that we could move either Figure 7 or Figure 9 to the supporting information while retaining the discussion in the revised manuscript.

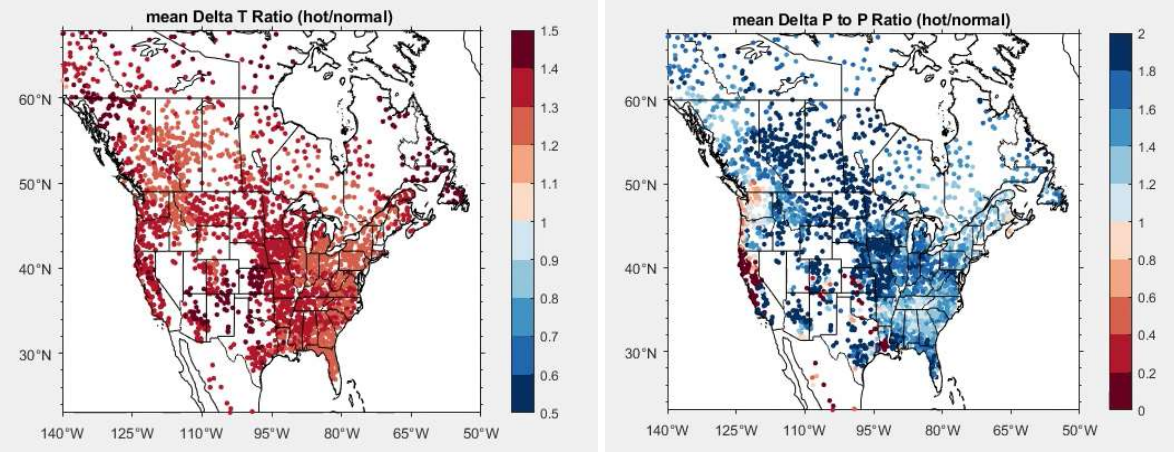
We will also merge the boxplots with their corresponding figures (Figure 5 with 6, Figure 7 with 8, Figure 9 with 10), and move Figures 12 and 13 to the supporting information.

One small frustration with this manuscript is that once variables are declared, and given abbreviated names, often such names are not then used systematically through the paper. This is especially noticeable when presenting keynote diagrams, such as Figure 5. Hence, the caption for Figure 5 should read something like: “Total spread ratio, TS_nd, for Q_mean…..”

Good point. In the revised version of the paper, we will ensure that we use the abbreviated names of each variable systematically.

Would the authors like to comment on the spatial cohesion in plots such as Figure 5. This diagram in particular seems to show that it is the higher latitudes that will see the bigger effects if the high ECS models are correct? Certainly the largest levels of warming will be seen towards to poles, which is a generic result for ESMs. That is the ESMs with the largest ECS which show the most pronounced warming poleward. This might also be true for precipitation?

To address this valid point, we propose adding a figure showing the mean climate change ratio for ΔT and $\Delta P/P$. As discussed earlier, the ratio represents the mean projected changes (‘hot’ models to normal models). These figures will assist in interpreting some of our other figures. This new figure reveals intriguing patterns, particularly for precipitation. The temperature ratio is relatively consistent, with the additional warming projected by the 'hot' models always exceeding 1 and showing less spatial dependency compared to precipitation. For precipitation, strong spatial patterns emerge. The 'hot' models predict considerably higher precipitation over central North America, while they project lower precipitation over the West coast of the USA.



New Figure XX2: Mean ΔT (left-hand side) and ΔP/P (right-hand side) ratios (hot models to normal models). For ΔT, a red color indicates that hot models, on average, are warmer than their normal (non-hot) counterparts. For ΔP/P, a blue color shows that hot models are wetter than their normal (non-hot) counterparts. ΔT represents the difference between future and historical temperatures.

It is noted that the reference list is balanced and comprehensive. I also like the Introduction, and the references cited where attempts have been made (either via emergent constraints or historical data) to constrain the bounds of Equilibrium Climate Sensitivity (ECS) bounds. This paper contains a wealth of important information, addresses an important problem of how to deal with ESM outliers and in the important context of such model impacts on streamflow. I am very happy to see any other revised paper version.

We appreciate the reviewer's positive feedback and insightful comments. We will ensure that the revised version of the paper addresses the identified shortcomings and points raised.