Reviewer 1

Thank you for inviting me to review the paper: "Historical and Projected Future Runoff over the Mekong River Basic" by Wang et al.

This is a useful and important paper. It is useful because it builds links between the detailed hydrological modelling community and those developing GCMs. Too often, climate researchers consider a relatively basic land surface model in a GCM as sufficient – but in reality, something much better is needed to help understand future flooding impacts. The paper is important because, as the authors state, 65 million rely on the Mekong River for access to water.

The analysis is well-considered and thorough. My only concern with the paper is that there needs to be better wording and removing ambiguity in places. All of this can be easily rectified in the generation of the new paper version (and I am happy to re-review any revisions).

Response: We appreciate the reviewer's positive feedback and helpful comments, which are highly helpful for us to improve the manuscript. Please kindly find below our detailed responses to each of your comments. Texts in blue are our responses to the comments, while those in red are revisions of the manuscript.

Below are illustrative examples, but please check through the entire manuscript.

P2, Line 48. "However, these studies do not systematically analyse….". Is this suggesting that substantial errors could occur with good GHMs, should there be major biases in GCMs (so a GCM+GHM combination fails, even if the GHM is good).

Response: Yes, as we mentioned in Line 26 of the manuscript, "*Different GCMs use distinct representations of the climate system, leading to "climate model structural uncertainty" (Gosling and Arnell, 2011).* ". We have added this information following the sentence in Line 48 in the revised manuscript:

"*However, these studies do not systematically analyze the runoff simulation results of long-term historical periods (including the historical period of historical scenarios and the real-time period of representative concentration pathways (RCP) scenarios, i.e. from the start simulation year of the RCPs to now pre-2023, for which observed runoff data are available.) under different GCM-GHM combinations. Such an analysis is meaningful and urgent to potentially assess and reduce the uncertainty/bias of runoff simulations introduced by both GCMs and GHMs (Kingston et al., 2011; Hoang et al., 2016) .*"

P2, Line 49. The wording here is clumsy. Maybe something like: "and the simulated years at the beginning of the RCP scenarios, which are now pre-2023 and for which runoff data exists."

Response: Thanks for the helpful suggestion. We have changed this sentence in the revised manuscript:

"*However, these studies do not systematically analyse the runoff simulation results of long-term historical periods (including the historical period of historical scenarios and the real-time period of representative concentration pathways (RCP) scenarios, i.e. from the start simulation year of the RCPs to now pre-2023, for which observed runoff data are available.) under different GCM-GHM combinations.*"

P3, Line 58. Please state what ISI-MIP is. Are the GCM-GHM combinations already calculated in ISI-MIP, or is that database just GCM output? Are any of the outputs from ISI-MIP already bias corrected?

Response: The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) is a community-driven modelling effort and offers a framework for comparing climate impact projections in different sectors and at different scales (Warszawski et al., 2014). Specifically, the ISIMIP 2b scenarios are designed to elicit the contribution of climate change to impacts arising from low-emissions climate-change scenarios (Frieler et al., 2017). You are right that all the GCM-GHM combinations already calculated in ISIMIP 2b, and the results of runoff simulations of five GHMs forced by four GCMs are all derived from the experimental data of the global water sector in ISI-MIP2b. Here, all GCM output meteorological forcing have been bias adjusted. These adjusted meteorological outputs have been collected in the EWEMBI dataset and used as meteorological forcing inputs for all the GHMs in the ISI-MIP2b (Frieler et al., 2017). We have added above information on the ISIMIP project in Section 2.3 in the revised manuscript:

"*2.3 Climate projections and hydrological models*

ISI-MIP is a community-driven modelling effort and offers a framework for comparing climate impact projections in different sectors and at different scales(Warszawski et al., 2014). In the ISI-MIP, the ISIMIP 2b scenarios are designed to elicit the contribution of climate change to impacts arising from low-emissions climate-change scenarios (Frieler et al., 2017). The global climate models (GCMs) selected for this study are derived from the ~~Inter-Sectoral Impact Model Intercomparison Project (~~ISIMIP~~)~~ 2b protocol, which provides four GCMs from CMIP5 and three emission scenarios (i.e., RCP2.6, RCP6.0 and RCP8.5). … All GHMs operate under the meteorological drive of the four GCMs, and the ensemble-averaged results of the GCMs are also evaluated due to the variability of the GCMs and the uncertainty of climate change. The above runoff *simulation results of five GHMs forced by four GCMs are all derived from the experimental data of the global water sector in ISIMIP2b.*"

P3, Line 72. Please clarify why you would use the MK test, and not the standard statistical test of whether a regression line is statistically significant.

Response: The Mann-Kendall (MK) test (Mann, 1945; Kendall, 1948) is a rank-based non-parametric method. Compared to parametric tests (e.g., regression coefficient test), non-parametric tests (e.g., the MK test) have no requirements of homoscedasticity or prior assumptions on the distribution of the data sample (Bihrat and Bayazit, 2003) and are less sensitive to outliers (Hamed, 2007). As the MK test statistic is determined by the ranks and sequences of time series rather than the original values, it is robust when dealing with non-normally distributed data, which are commonly encountered in hydrometeorological time series (Wang et al., 2020). We will provide a brief explanation of the choice of the MK test in the revised manuscript.

P5, Line 91. Please state where the GHMs come from. Is it a database such as ISI-MIP. State here that information, and also in the caption of Table 2 (Also, please check the current Caption of Table 2 – it looks wrong, referring to eight hydrological stations).

Response: The runoff simulations results of five GHMs forced by four GCMs were all derived from the experimental data of the global water sector in ISI-MIP2b, which is openly available on ISIMIP protocol (https://data.isimip.org/search/product/). Meanwhile, thanks for your reminder, we have checked the caption of Table 2 and added the information of GHMs sources in the revised manuscript:

"*Table 2:* ~~*Basic statistical information of eight hydrological stations.*~~ *Basic information of the GHMs in the ISIMIP2b Global Water program. The runoff simulation results of the GHMs forced by different GCMs are all derived from the ISIMIP protocol (https://data.isimip.org/search/product/).*"

There is one technical issue. Could the authors describe if there is any bias correction undertaken e.g. of the ESMs? To my knowledge, some ESMs in the "MIPs" are corrected. If so, are these used – because they should reduce climate uncertainty/errors in any GCM+GHM projections of the contemporary period?

Response: As mentioned earlier, the meteorological forcing from the GCMs in ISI-MIP2b have performed bias adjustment to reduce climate uncertainty/error in future projections. For the detailed description of the bias adjustment, please refer to Frieler et al. (2017).

P6, Table 2 – as noted elsewhere, captions appear overly succinct. I often see people give talks where diagrams and tables are extracted from papers, so if they can be more complete (i.e. with essential details in captions), then this is very helpful.

Response: Thanks for your reminder and suggestion. We have changed the caption of Table 2 to make it more complete in the revised manuscript:

"*Table 2: ~~Basic statistical information of eight hydrological stations.~~ Basic information of the GHMs in the ISIMIP2b Global Water program. The runoff simulation results of the GHMs forced by different GCMs were all derived from the ISIMIP protocol (https://data.isimip.org/search/product/).*"

P7, Figure 2 – One possibility to avoid the repeated words "insignificant change" is to give the p-value for the regression. Then where it is significant (e.g. p < 0.05), mark with a star symbol. Something like that… Also, again, expand the caption slightly. For instance, state, "Eight hydrological stations are numbered N1-N8, with their locations presented in the map of Figure 1".

Response: Thanks for your helpful suggestions. We have checked the Figure 2 and added p-value where significance tests are involved. We have also changed the caption of Figure 2 in the revised manuscript:

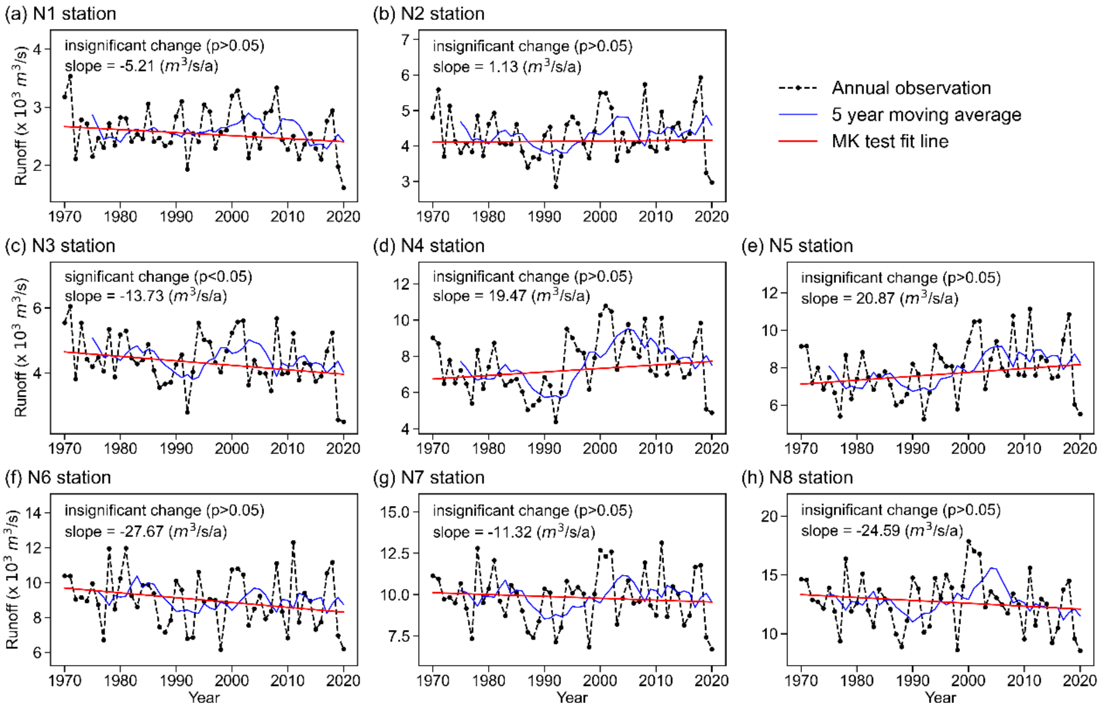


Figure 2: The results of the MK trend test in historical (1971-2020) runoff over the

eight hydrological stations. Eight hydrological stations are numbered N1-N8, with their locations presented in the map of Figure 1.

P7, Figure 2. Although very obvious, please put the word "Year" under panels (f), (g) and (h).

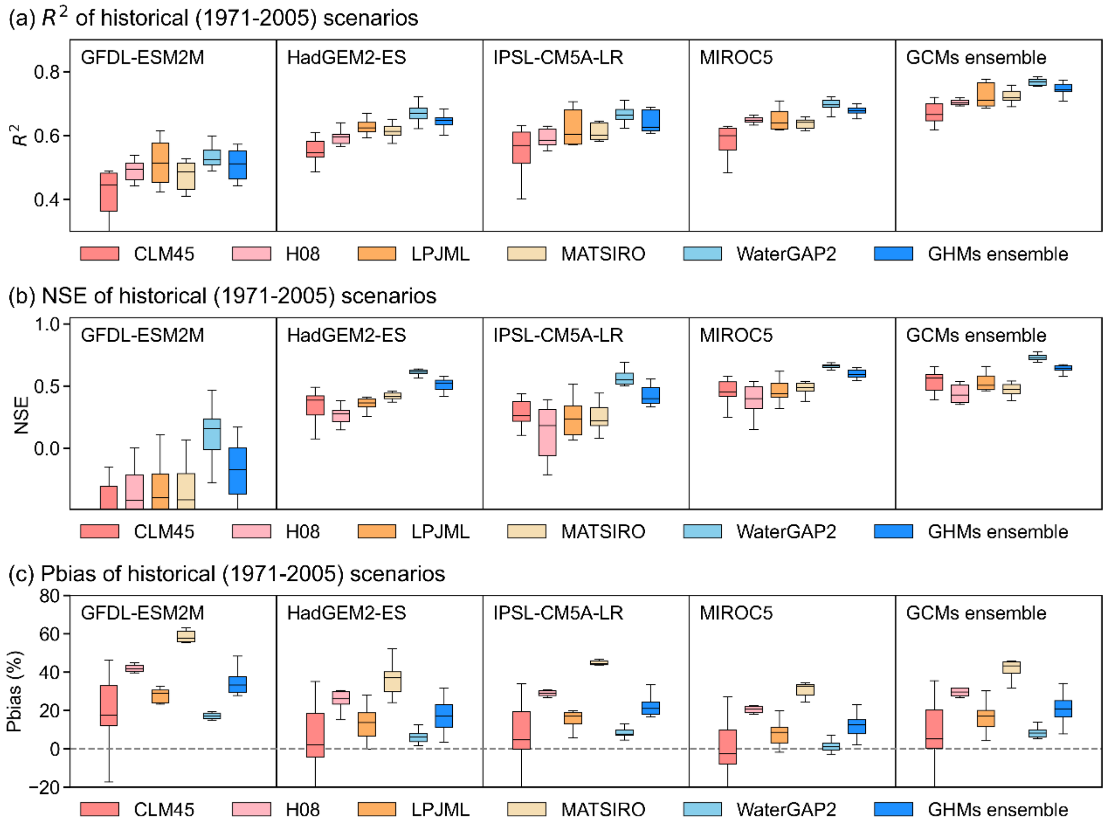
Response: Thanks for your helpful suggestions. We have revised Figure 2 accordingly in the revised manuscript, which can be found in the response to the previous comment.

P8. Table 3. The table is nice, but isn't all the information in Figures 1 and 2?

Response: Thanks for your helpful suggestions. The purpose of Table 3 is to provide the information of significant test of the changing trend. Considering the redundancy of the information, we have added significance test information to Figure 2 and then deleted Table 3. The revised Figure 2 can be found in the response to the previous comment.

Figure 3. Again, please make the caption much more informative. Something like "Figure 3. Performance of all combinations of GCMs and GHMs. The three rows correspond to three performance matrices……. In each row, each panel is for a different GCM, as annotated. Then in each panel, the different colors are for each GHM, as marked under each row……"

Response: Thanks for your helpful suggestions. We have changed the caption of Figure 3 in the revised manuscript according to the above suggestions:



"*Figure 3. Performance of all combinations of GCMs and GHMs during historical (1971-2005) periods. The three rows correspond to three performance matrices (R^2, NSE and Pbias) of all GCM-GHM combinations at the eight hydrological stations. In each row, each panel is for a different GCM, as annotated. Then in each panel, the different colors are for each GHM, as marked under each row.* "

Furthermore, we have also changed the captions of the remaining figures to make them more informative in the revised manuscript.

Figure 3. Although it is important to make captions informative, placing a result there is not usual practice. So please reconsider the words "WaterGap2 has the best performance compared to other models", and should they be in a caption?

Response: Thanks for your helpful suggestions. We have removed this sentence from the captions of Figure 3. The revised the captions of Figure 3 can be found in the response to the previous comment.

Figure 3. And on the same point above, statements such as "…best performance compared to other models" requires very careful quantification. Is it the best performance when WaterGap2 is driven by a specific GCM? Is it the best for one statistical metric or many?

Response: Thanks for your helpful suggestions. Specifically, we have quantitatively compared three performance metrices of WaterGap2 and other hydrological models under forcing from each of the four GCMs. It was found that WaterGap2 has the highest R^2 and NSE and the lowest Pbias than other GHMs under any *same* GCM forcing. This shows that despite the uncertainty between different GCMs, the simulation of the MRB runoff by WaterGap2 has better performance and reliability than other models. We have clarified this point in Line 141 of the revised manuscript:

Line 141: "*Among these GHMs, ~~the performance of the WaterGAP2 model is the best~~ WaterGap2 has the highest R^2 and NSE and the lowest Pbias than others under the same GCM forcing.*"

In the Conclusions, critical is consistency between direct drivers of runoff change and changes imposed by raised GHGs and related altered rainfall patterns. Here, the impression is that the former is small (i.e. "However, the impact of the reservoir on the annual runoff after the completion of water storage is small"). Elsewhere in the paper, there is the suggestion that humans have impacted – directly – runoff strongly. I still think it would be useful to have a summary statistic that is some sort of ratio between historical direct change and the impact of raised GHGs on runoff. This should be easy to do, as all the numerical values to build such a single comparison statistic are calculated at different points within the manuscript.

Response: Thanks for your helpful suggestions. The ratio you mention, comparing historical direct changes and future climate impacts on runoff, is undoubtedly meaningful and attractive. In our current manuscript, the historical direct changes are calculated from observed runoff and are also influenced by both historical climate change and historical human activities. Direct comparisons between the combined effects of historical climate change and human activities on runoff and the separate effects of future climate change on runoff may lead to misleading conclusions because the interactions of climate change and human activities on runoff over historical periods are usually complex and unclear. With sufficient reservoir/dam-related data, we can hopefully separate the effects of historical climate change from the effects of historical human activities and compare them to the effects of future climate change. In the absence of reservoir/dam data, our current work only calculates direct historical changes and does not compare them to future climate change impacts. In the revised manuscript, we will include a more thorough discussion on this point, and- highlight the importance of future efforts to record or collect local reservoir/dam data to further explore the relationship between human activities impacts and climate change impacts.

Small things – here, for Abstract but may be representative elsewhere

Abstract: These need to avoid ambiguity, as often read in isolation by a reader in a hurry. Hence please:

(1) tighten line 7 and explain the difference between how "four GCMs" and "five GHMs" are used. Are they operated independently e.g. raw runoff output from the GCMs are used – while the GHMs are forced with known near-surface meteorological drivers? The next sentence, however, talks about "best simulation combination", so state this is 4 x 5 simulations – all combinations of GCMs and GHMs (I realise the main body of the paper makes this clearer, but such very basic information should be in an Abstract).

Response: Thanks for your helpful suggestions. We apologize for any confusion caused by the wording here. Here we use the runoff simulation output of five GHMs driven by four GCMs, which is the combination of GCM and GHM described in the manuscript. To avoid confusion, we have clarified in Line 7 in the revised manuscript:

"*With these runoff data, we then evaluate the runoff simulation performance of ~~four global climate models (GCMs) and five global hydrological models (GHMs)~~ five global hydrological models (GHMs) forced by four global climate models (GCMs) under the ISI-MIP project.*"

(2) Line 11. State what "WaterGap2" is (i.e. GHM).

Response: Thanks for your helpful suggestions. We have clarified in Line 11 in the revised manuscript:

"*~~WaterGap2 forced by GCMs ensemble-averaged climates~~ The ensemble-averaged result of WaterGap2 (i.e., GHM) forced by four GCMs has the best runoff simulation performance.*"

(3) The Abstract presents two lines of investigation but does not bring them together in a coherent way. One direction is that for the contemporary period, it is dams and reservoirs that have had the biggest effect on runoff. However, when describing the future based on RCPs, runoff is described as "projected to increase significantly". The question is then whether future changes caused by climate change are bigger than current changes caused by dams/reservoirs? (See similar comment above)

Response: Thanks for your helpful suggestions. Quantitative comparisons of future changes caused by climate change and current changes caused by dams/reservoirs are meaningful and attractive. Unfortunately, we currently lack sufficient dams/reservoirs data to quantify the current impact of dams/reservoirs on runoff (see previous response above). Our current manuscript focuses on qualitative analysis of the current impact of dams/reservoirs on runoff and quantitative analysis of the future impact of climate change on runoff. In future work, we expect to be able to quantitatively analyze the impact of human activities and climate change on runoff by acquiring or collecting reservoir data.

(4) Line 13. Is "increase significantly" a formal statistical statement, and should there be a p-value?

Response: Yes, here is a formal statistical statement where a p-value is required. We have clarified this in Line 13 in the revised manuscript:

"*Under representative concentration pathways (RCPs, i.e., RCP2.6, RCP6.0 and RCP8.5), runoff of the MR is projected to increase significantly (*$p<0.05$*, from 3.81* m^3 s^{-1} a^{-1} *to 16.36* m^3 s^{-1} a^{-1}*).*"

To be more rigorous, we have added p-values to all formal statistical statements and replaced "significant" with synonyms for informal statistical statements throughout the current manuscript. All these revisions are included in the revised manuscript. The significance level used in this study is clarified in Section 2.2 in the revised manuscript:

"*The null hypothesis in this test is that there is no significant trend in the time series at the significance level of p. If* $|U|\geq U_{p/2}$*, where* $U_{p/2}$ *is the critical value of the standard normal distribution with a probability exceeding p/2, then the null hypothesis is rejected, namely the trend is significant (Guan et al., 2021). This study adopts the significance level of 0.05, which means that there is a significant trend of change when the p-value is less than 0.05.*"

(5) Line 13. Actual values are given here (units of m^3 s^-1 a^-1). Similar to the comments above, how large are the 3.81 – 16.36 numbers compared to the effects of dams/reservoirs? And how large are these numbers compared to background contemporary flows. Would a simple statistical value help?

Response: Thanks for your helpful suggestions. Considering the quantitative effects of dams/reservoirs are not available, we have only added the ratio of the actual values to their background contemporary. We have clarified in Line 13 in the revised manuscript:

"*Under representative concentration pathways (RCPs, i.e., RCP2.6, RCP6.0 and RCP8.5), runoff of the MR is projected to increase significantly (*$p<0.05$*), e.g., 3.81* m^3 s^{-1} a^{-1} *(9% increase in 100 years) at the upstream station under RCP2.6 and 16.36* m^3 s^{-1} a^{-1} *(13% increase in 100 years) at the downstream station under RCP6.0.*"

References:

Bihrat, Ö. and Bayazit, M.: The power of statistical tests for trend detection, 27, 247-251, 2003.

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), Geoscientific Model Development, 10, 4321-4345, 10.5194/gmd-10-4321-2017, 2017.

Gosling, S. N. and Arnell, N. W.: Simulating current global river runoff with a global hydrological model: model revisions, validation, and sensitivity analysis, Hydrological Processes, 25, 1129-1145, 10.1002/hyp.7727, 2011.

Guan, X., Zhang, J., Bao, Z., Liu, C., Jin, J., and Wang, G.: Past variations and future projection of runoff in typical basins in 10 water zones, China, Sci Total Environ, 798, 149277, 10.1016/j.scitotenv.2021.149277, 2021.

Hamed, K. H.: Improved finite-sample Hurst exponent estimates using rescaled range analysis, 43, 2007.

Hoang, L. P., Lauri, H., Kummu, M., Koponen, J., van Vliet, M. T. H., Supit, I., Leemans, R., Kabat, P., and Ludwig, F.: Mekong River flow and hydrological extremes under climate change, Hydrology and Earth System Sciences, 20, 3027-3041, 10.5194/hess-20-3027-2016, 2016.

Kendall, M. G.: Rank correlation methods, 1948.

Kingston, D. G., Thompson, J. R., and Kite, G.: Uncertainty in climate change projections of discharge for the Mekong River Basin, Hydrology and Earth System Sciences, 15, 1459-1471, 10.5194/hess-15-1459-2011, 2011.

Mann, H. B.: Nonparametric tests against trend, 245-259, 1945.

Wang, F., Shao, W., Yu, H., Kan, G., He, X., Zhang, D., Ren, M., and Wang, G.: Re-evaluation of the Power of the Mann-Kendall Test for Detecting Monotonic Trends in Hydrometeorological Time Series, Frontiers in Earth Science, 8, 10.3389/feart.2020.00014, 2020.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework, Proc Natl Acad Sci U S A, 111, 3228-3232, 10.1073/pnas.1312330110, 2014.