Referee comments are highlighted in bold with our response in normal font

General comments:

This paper examines eight global hydrologic models (GHMs) across the Amazon basin for several basic performance metrics. The model configurations allow for somewhat limited comparisons of coarse granularity methodological choices (e.g. routing model, input forcing data) across the performance metrics for a subset of the GHMs. The submission is easy to read and logically organized with clearly stated objectives. The overall conclusion that the precipitation dataset may be the most important for essentially uncalibrated GHMs and that calibration of the routing model using daily KGE as the objective function has limited effectiveness, particularly for floods, intuitively make sense and agree with the presented results. While not particularly novel, they are useful conclusions to reiterate to the GHM model development and user community.

We thank Dr. Newman for his useful evaluation of the manuscript and suggestions for improvement. We will address these in the revised manuscript, as according to our responses to each comment below.

Specific comments:

1) This paper suffers from the same issues all macroscale intercomparison studies do. In general the analysis is somewhat superficial and the conclusions regarding which modeling component is best (e.g. routing or hydrologic model) are limited to that very coarse scale. Also, while I do appreciate that the authors describe their model selection efforts, it remains that only a subset of the model decision matrix is populated in an adhoc manner. This limits comparisons and conclusions to specific subsets of models depending on the question asked.

It may be beneficial for the authors to add some statements describing the limitations of these types of studies. For example, question 3.3 examines routing models, but the comparison is for one hydrologic model and one precipitation dataset only. The conclusions in this section are reasonable for that specific comparison; the results may change given another modeling chain upstream of the routing model. When discussing future work/paths forward, it may be helpful to talk about increasing the granularity at which modeling decisions are tested and filling out the model decision matrix in a more quantitative way so that more generalized conclusions can be made.

Overall, we agree with the limitation of comparing only macroscale features between each GHM (e.g. routing model/precipitation forcing) and the suggestion to explicitly highlight some of these limitations. As Dr. Newman mentions, we do highlight the reasoning behind the model and comparison selections in the manuscript with acknowledgment to alternative methodologies.

We plan to change the sub-heading for Sect. 3.6 to "Limitations and future work" by discussing the limitations of using a macroscale intercomparison approach. We will add to the discussion the possibility to expand the study by increasing the granularity at which modelling decisions are tested; thereby allowing more generalised conclusions to be made.

2) There are references relevant to the routing model calibration discussion the authors should consider. Their conclusion that performance is improved for metrics more closely related to the objective function is correct, however the statements on page 17, lines 13-15, and again on page 18, lines 4-5 need further elaboration. Gupta et al. (2009) and Mizukami et al. (2019) discuss in detail how squared error metrics relate to high flow performance. Mizukami et al. (2019) also show how an application specific metric, annual peak flow bias (APFB) can improve model

performance for that specific metric, but at the expense of decreased performance in related metrics. This is directly relevant to the final sentence of the conclusions section, and it would be good to note that some application specific metrics could degrade model performance for other parts of the hydrograph, so thoughtful consideration to the full use of the modeling system should be given when performing parameter optimization.

We thank Dr. Newman for this very useful suggestion. We agree that the original conclusion would benefit from further elaboration and evidence from previous works.

We will incorporate both the suggested references to support our discussions and conclusions. Specifically:

i) By commenting on the suitability of square error related metrics for model calibration when the application requires robust performance for high flows (in Sect. 3.6)
ii) By providing the example of increased model performance when using application specific metrics (i.e. Annual Peak Flow Bias), as in Mizukami et al. (in review)
iii) By discussing the need to carefully consider the metric used in calibration, with the possibility of a loss in skill for other related metrics upon evaluation when using application specific metrics (Sect 3.6 and Sect. 4)