General comments

Thank you for the well written and organized manuscript.

Calibration of parameters for hydrological and hydraulic processes is very important for river discharge, and I believe that calibration should be conducted carefully when applying to the global scale. It is difficult to apply the current results to the global scale. Further analysis of the current results is necessary.

Response: Thank you for reviewing our manuscript and providing the constructive comments for us to improve the manuscript. We agree calibration should be conducted carefully at global scales. Indeed, this work was inspired by current global calibration studies. As we argued in the manuscript, current calibration studies only focused on one process to improve the simulated streamflow but resulted in poor performance in other relevant processes. We aim to address this gap by proposing a two-step calibration method. In the current manuscript, we already tested our two-step calibration method in another watershed with different watershed characteristic and climatology (e.g., Susquehanna River basin locates in northeastern US). Without modification in the calibration procedure, the calibrated model shows improved performance in capturing baseflow index, surface water dynamics, streamflow variation, and annual streamflow trend. We note that the proposed two-step computational method is expensive to apply at global scales since it would take 2,000 global coupled ELM-MOSART simulations, which is beyond the scope of this study. In addition, another challenge for the calibration at global scales with the two-step method is that many basins are not gauged for monitoring streamflow. We will add a paragraph to discuss the challenges of applying our method to global scales in the revised manuscript. We think we need a follow up study in the future for the calibration, analysis, and evaluations at global scales.

Please find our point-to-point response to your comments in the following.

Although it may not be the purpose of this study, specifically, an analysis of the relationship between the parameters and the characteristics of the target river basin (precipitation, soil, topography/geology, etc.) would increase its applicability to the global scale.

Response: Many parameters in ELM and MOSART are derived from surface and subsurface conditions. For example, in ELM, saturated hydraulic conductivity and specific yield are estimated based on soil types, maximum drainage rate is determined by topographic slope, etc. In MOSART, river length and slope are derived from high resolution DEM. However, some other parameters should be determined based on sensitivity analysis and calibration, such as the parameters selected in this study. This is because ESM is typically too coarse to represent some physical processes, and empirical functions are used to parameterize those processes with simplifications. Specifically, f_{drain} and f_{over} are decay factors of the exponential function in the subsurface and surface runoff generation processes, respectively. According to the development of the runoff scheme in our model (e.g., simple TOPMODEL-based runoff parameterization), f_{drain} and f_{over} should be determined through sensitivity analysis or calibration against hydrograph recession curve (Niu et al., 2005; Niu et al., 2007). f_c is a threshold below which no single connected inundated area spans the grid cell. It is used to quantify the fraction of the

inundated portion of the grid cell that is interconnected according to percolation theory. In other words, f_c determines the maximum inundation extent, above which the water will outflow from the inundated area. Although the maximum inundated area should be controlled by topographic variation, high resolution data that captures the topographic variation under the inundated areas is not available.

There exist observations of river width and depth at very high spatial resolution. But it is challenging to upscale the observed river width and depth to ESM resolution (e.g., around 1deg) (Liao et al., 2022). The relationship between discharge (or drainage area) and river channel geometry is commonly used to determine the river width and depth. However, such relationship varies from watershed to watershed. It is not possible to use single factor to derive the river geometry as it is affected by multiple factors such as discharge magnitude, seasonality, lithology, channel slope, etc. Overall, the coarse spatial resolution of ESM and its simplifications in physical process make it hard to identify the relationship between selected parameters and watershed characteristic. In addition, according to our calibration experience in ESM, the most effective method to improve model performance and representation of the selected parameters is calibration.

Furthermore, we have difficulty in determining whether this two-step calibration is a good idea.The reason is that we do not know what kind of changes the first and second steps brought about in the river discharge, respectively.

Since this is a two-step calibration, there should be an answer for the first step, which can be shown in Fig. 12 to indicate how the river discharge changed from the default. It would be necessary to show how the river discharge changed and at what locations.

Response: Please find in the Figure R1 for the simulated streamflow after calibration of step 1 and step 2. Compared to the default simulation (red dashed line), the step 1 (blue dashed line). calibration shifts the streamflow by about 1 month. The Step 2 calibration (blue solid line) further delays the streamflow seasonality, especially for the downstream subbasins (subbasin #5 and #6). We will update Figure 12 in the main text in the revision.

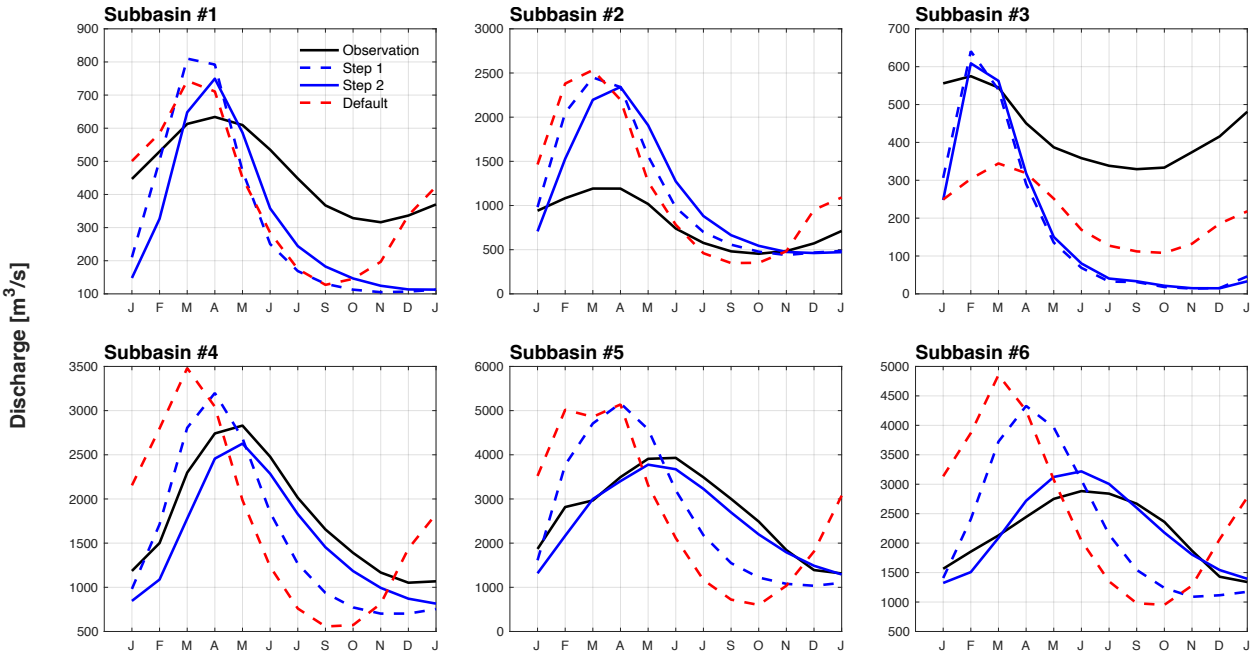

Figure R1. Simulated streamflow seasonality during 1981-2010 for the selected subbasins with default parameter values (red dashed line), calibrated ELM parameter values and default MOSART parameter values (blue dashed line), and calibrated ELM and MOSART parameter values (blue solid line). The black solid line denotes the observed streamflow seasonality. The blue dashed and solid lines represent the calibrated streamflow after Step 1 and Step 2 calibration, respectively.

Specific comments

How about describing the characteristics of the sub basin observation points in 2.2? Currently, the entire river basin is described with a focus on outlet points, but it would help to reinforce the last part of the discussion. In particular, SB#3 has a high river discharge for the area of the upstream river basin. It would be nice to have a comparison between precipitation and runoff height, even if it is shown on the supplement.

Response: Thanks for your suggestion. We will add some descriptions of the selected sub-basins in section 2.2. We note the reviewer is correct that SB#3 has a relatively high discharge for the area of the upstream river basin. As shown in Figure R2, the runoff coefficient in SB#3 is much larger than other sub-basins. This is because SB#3 locates in elevated area with steeper slope than other sub-basins.

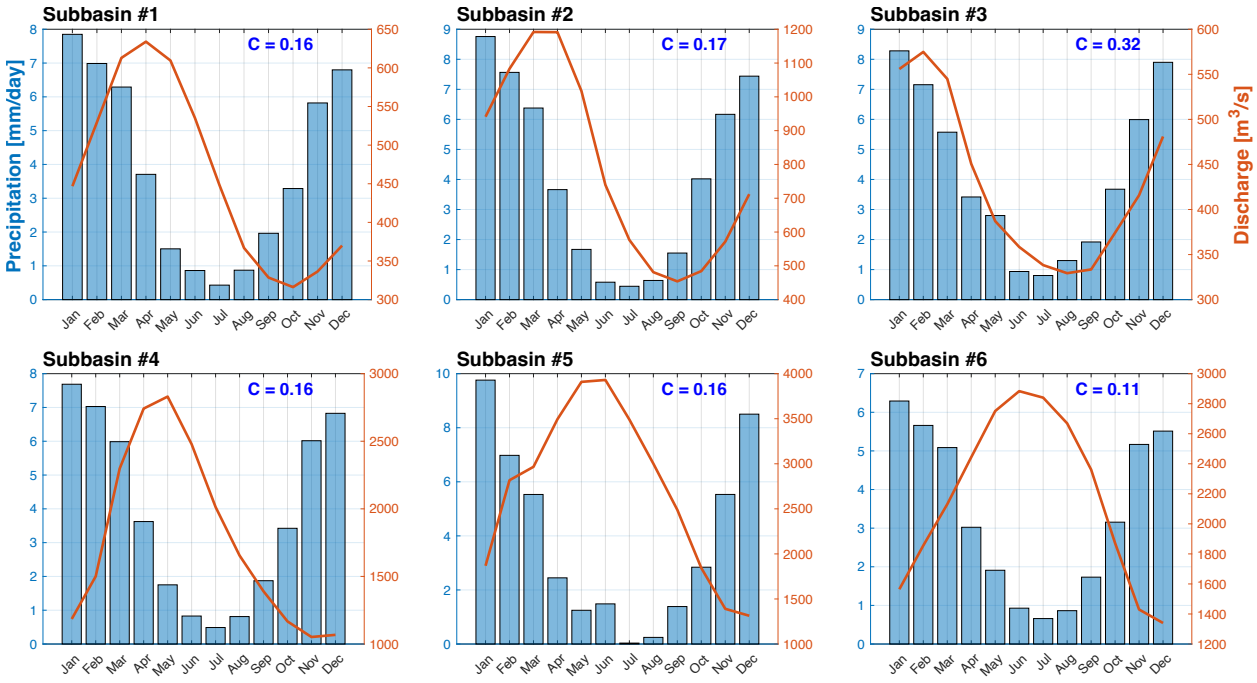

Figure R2. Monthly averaged precipitation and discharge for the selected subbasins. C denotes runoff coefficient, which is the ratio of discharge to precipitation.

What is the reason why you chose these three components for multi object funcition in 2.5, you wrote that you didn't include streamflow variability in 3.5 because we didn't know how much it would be affected by runoff process only, but why did you include SWF? I think the SWF is also affected by river routine model, MOSART.

Response: We picked these three components in the multi-objective function because we think these are very critical metrics relevant to streamflow variability. We acknowledge other objective can be used for model calibration in other models, for example, some ESMs/LSMs may not have the surface water dynamics. We argue that both hydrological and hydraulic processes should be carefully calibrated together in the revised manuscript, however, the modelers could adopt other multi-objective functions based on the available processes in the calibrated models and theirs needs for the calibration.

In simulation, SWF is the sum of pluvial inundation (i.e., simulated by the land component, ELM) and fluvial inundation (i.e., simulated by the river component, MOSART). Therefore, SWF is affected by both runoff generation and river routing processes. In the first step calibration for the multi-objective function, we only calibrate ELM parameters, with MOSART parameters constant. Thus, the SWF in the multi-objective function is not affected by the routing process. We will add a clarification for this in the revised manuscript.

In the comparison between 3.4 and 3.5 (Fig. 9(a)), why is the river discharge lower in the two-step case than in the case where hydrology and hydraulics are calibrated separately?

Response: When hydrology is calibrated separately, only streamflow variability is considered in the objective function. However, in the two-step calibration, multi-objective function is used. The calibrated parameter corresponds to the smallest multi-objective function of Eq (4) is different from the calibrated parameter in calibrating hydrology and hydraulic separately as other variables were considered. We will update the text in the revised manuscript for clarity.

Regarding discussion 3.5, you list three factors for underestimation of river discharge, but I think we need to be sure that these three factors are really contributing to the problem. For example, if we consider average evapotranspiration, increase the area of the river basin by 5%, and increase precipitation, how much runoff will be generated and whether this runoff can represent an underestimation of the river discharge, we can check this at the order level. Alternatively, we could implement the results for other precipitation products.

Response: The simulated streamflow underestimates the observed streamflow by about 11%, which is $-17\ [mm/yr]$ in absolute value. The smaller contributing area can contribute to ~5% in the streamflow underestimation, assuming the runoff from the missing area is similar to the basin averaged runoff. The precipitation ensemble (e.g., based on the selected precipitation datasets) in Table S1) leads to the average annual precipitation during 1979-2009 to be $1{,}107 - 1{,}229\ [mm/yr]$. CRUNCEP forcing that used in our study has the average annual precipitation to be $1{,}156\ [mm/yr]$. Therefore, the long-term uncertainty of the precipitation in the used forcing is about $-73 - 49\ [mm/yr]$. Similarly, considering the two evapotranspiration (ET) references, the simulated ET uncertainty is about $54 - 154\ [mm/yr]$. Assuming the water storage doesn't change in the simulation period, the long-term averaged runoff is the difference between the long-term precipitation and evaporation. Therefore, an approximate estimate of runoff uncertainty from precipitation and evaporation is about $-227 - -5\ [mm/yr]$. Overall, the three factors can explain the underestimation.

We will add this additional analysis in the revised manuscript.

After that, I think it is necessary to consider the factors that lead to an overestimation of evapotranspiration.

Response: There are several attributions for the uncertainty of simulated evapotranspiration (ET). First, the uncertainty in the forcing can affect ET process. Specifically, the shortwave incoming radiation and air temperature in the used atmosphere forcing (i.e., CRUNCEP) is larger than atmosphere forcing of The Global Soil Wetness Project Phase 3 (GSWP3) (Figure R3). The higher shortwave incoming radiation and air temperature can both lead to higher evapotranspiration. The second factor is uncertainty in the surface dataset in ELM, for example, vegetation types and corresponding lead area index (LAI), which are a very sensitive surface conditions for ET process. The third uncertainty is due to parameterization. The ET parameterization is based on Monin-Obukhov similarity theory. The estimation of aerodynamic resistance and stomatal resistance commonly contain substantial uncertainty. Although the streamflow is significantly impacted by ET processes, it is challenging to calibrate ET process due to the lack of ET observation. We will add additional discussion of potential factors for the ET overestimation in the revised manuscript.

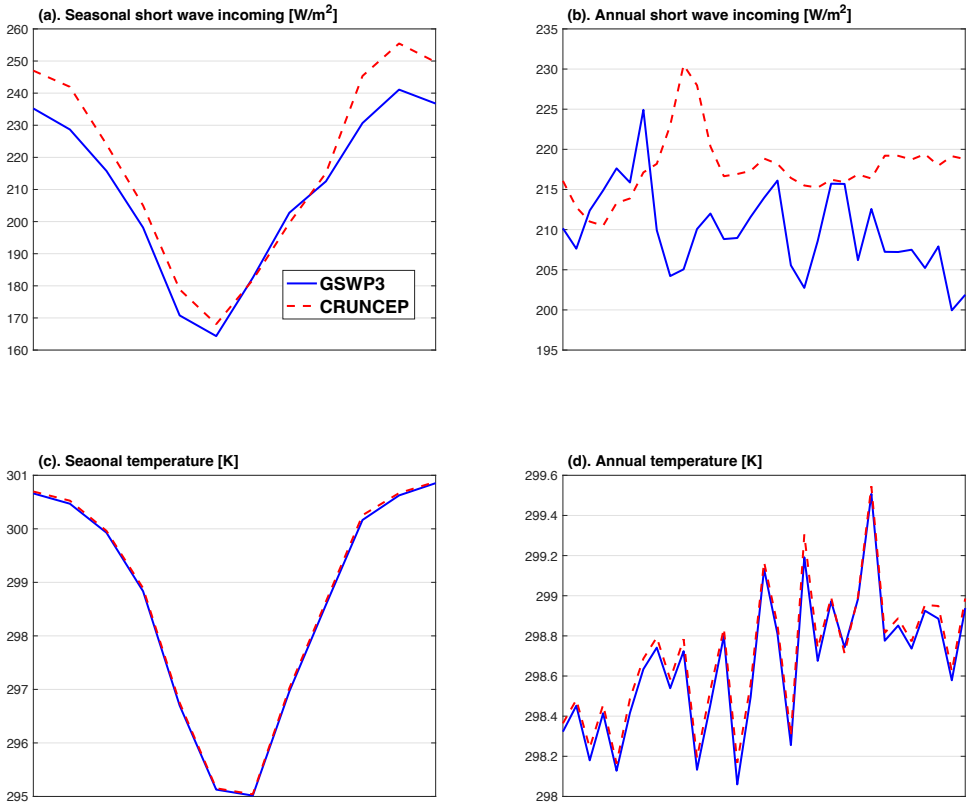

Figure R3. Comparison of (a). seasonal short waving incoming; (b). annual short waving incoming; (c). seasonal temperature; and (d) annual temperature between GSWP3 (blue solid line) and CRUNCEP (red dashed line).

Liao, C., Zhou, T., Xu, D., Barnes, R., Bisht, G., Li, H.-Y., Tan, Z., Tesfa, T., Duan, Z., Engwirda, D. and Leung, L.R. 2022. Advances in hexagon mesh-based flow direction modeling. Adv Water Resour 160, 104099.

Niu, G.-Y., Yang, Z.-L., Dickinson, R.E. and Gulden, L.E. 2005. A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. Journal of Geophysical Research: Atmospheres 110(D21).

Niu, G.-Y., Yang, Z.-L., Dickinson, R.E., Gulden, L.E. and Su, H. 2007. Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. Journal of Geophysical Research: Atmospheres 112(D7).