Yang et al. present results for a coupled surface-groundwater model for continental China. The work is appropriate for HESS, the research is very interesting and of high quality, and the manuscript is well written. Nonetheless, I have several recommendations regarding the evaluation of the model(s). I believe that a more process-oriented evaluation would be more meaningful for both the authors and the readers. My main other concern right now is a lack of discussion of the results. Both aspects are straightforward to rectify though.

Sincere thanks to Professor Wagener's kind words, constructive comments, and insightful thoughts on our work. We carefully read the suggested papers, rethought our work relevant to the concerns, and revised our manuscript to improve it. Here, we simply respond how we address each comment while details of revision can be seen in the final revised manuscript (which cannot be attached with the response).

Larger Comments:

[1] The use of a scaled statistical error metric: The authors state that "Note that all performance evaluations in this paper are based on the RSR value which is the ratio of the root mean squared error to the standard deviation of observations. An RSR value of 1.0 suggests good performance while 0.5 suggests excellent performance (O'neill et al., 2021)." These qualitative statements go back to the paper by Moriasi et al. (2007, doi.org/10.13031/2013.23153) who suggested some subjective qualification for normalized statistical metrics. The use of this subjective language persists even though it has been shown multiple times that the ease with which such values can be achieved varies with system properties (e.g. DOI: 10.1002/hyp.6825; doi.org/10.5194/hess-23-4323-2019). Therefore, these statements of good or poor performance with fixed thresholds are very unhelpful because – depending on the system modelled – it will be easy or hard to achieve these values. Personally (the authors do not have to share this view), I find it much more helpful to assess which system properties allow for high or low model performances (e.g. DOI 10.1088/1748-9326/abfac4 Figure 3 or DOI 10.1088/1748-9326/ad52b0). Such analyses are particularly valuable when done across multiple models, which often show that many models work well under specific conditions (often high wetness levels).

Thanks for this insightful suggestion. We reviewed these mentioned papers and agreed with the concern here. Audience should be cautious to treat these values as absolute performances. The RSR values shown in the manuscript are not comparable between different variables (e.g., drainage area, streamflow, water table depth). They are also not comparable with other case studies evaluating other systems or even the same system but in different periods. Yet, insights of relative performance could be gained from Figure 7 as the same benchmark (observations) is used for evaluating the same behavior (long-term average performance) in generally the same simulation period.

We first added an overall clarification, following the definition of RSR, about the limitation of using RSR as discussed in these listed papers. Then we added the variations of residuals of water table depth with critical factors into the paper, which is, essentially, also the response to comment [2].

[2] Possibility for understanding process controls: The focus on statistical metrics and maps for the comparison of the model with observations or other models provides limited insights into how and (potentially) why the models differ. A simple but effective way to provide more insight is to plot the water table depth (WTD, or other output variables) against (potentially) controlling

variables as functional relationships. For example, when plotting WTD against topographic slope for two of the models used by the authors – GLOBGM and Fan, the recent study by Reinecke et al. (doi.org/10.1088/1748-9326/ad8587) showed that GLOBGM is strongly correlated with slope, while the Fan model and global observations do so much less. Also, the Fan model shows distinct WTD differences between water and energy limited regions, while GLOBGM hardly does so. Similarly to my point 1, what controls the variability of model outputs and the output differences? These plots would include data, which the authors should have readily available – hence there is not much additional effort needed to try this.

Thanks for this constructive comment. We do have a substantial discussion about the shallowed simulated water table depth and the uncertainties caused by human activities. It's unfortunate that they were buried in the original manuscript probably due to the limitation of the manuscript structure as mentioned by Reviewer in comment [5]. In the revision, we added the variations of residuals with key factors (e.g., elevation, slope) into the manuscript (following figure 3 in DOI 10.1088/1748-9326/abfac4 or figure 9 in doi.org/10.5194/gmd-12-2401-2019) and reorganized the paper structure of relevant sections to better deliver our points. Yes, this is also a response to comment [1].

[3] Model omissions: Over 0.5 million km2 of Southern China has Karst geology (doi.org/10.1007/s10980-019-00912-w), which shows significantly different recharge patterns than many other geologies (doi.org/10.1073/pnas.1614941114). How is this reflected in the model set-up? Do these regions show distinctly different patterns than other areas regarding recharge or other variables?

Good point. Previous studies using ParFlow in Karst regions, such as the entire continental US (Yang et al., 2023, doi.org/10.1016/j.jhydrol.2023.130294) and the individual watershed in Florida (Srivastava et al., 2014; doi.org/10.1016/j.jhydrol.2014.10.020) show satisfied performances. Therefore, we didn't take specific actions in such regions. But we fully understand the recharge patterns in Karst regions might be highly different from other regions. The basic idea behind our work is that, at large enough scale, the Karst geology can be assumed as porous media while we recognize that the limitation of this idea must exist. Nevertheless, high hydraulic conductivities were setup in Karst regions in our model. We rechecked the residuals of water table depth shown in Figure 7 in the original manuscript, we do see something special, i.e., deeper simulated water table in all three models in the Karst regions and that GLOBGM v1.0 is the most significant one. We inferred that this might be caused by a larger P-ET in 2018 than long-term average P-ET but we cannot reject that this might be also attributed to the Karst geology. For example, wells are always drilled in places without significant Karst signatures and thus hold normal water table depths. Yet the higher average hydraulic conductivity might cause deeper water table in the simulation. Thanks for this good point motivating us to rethink this important question and we added this additional discussion into the revised manuscript.

[4] Comparison with global models: Global models are rather crude approximations of local hydrology – shown regularly. Comparison to these models is a good starting point, but also limited in what one can learn. Do any national scale modelling efforts exist for China that would also provide a comparison for the model introduced here? Clearly the model presented here has tremendous potential – given its coupled nature – but how would it have to be further improved? It would be interesting to discuss more what additional aspects local or regional models might consider relevant.

We had a lot of efforts regarding this. Unfortunately, we didn't get the results of relevant models. We understand and respect the preferences of authors of these models. As a result, we highlighted in the discussion that model comparison is encouraged. Yet it may take time to build a desired environment of the community.

[5] Lack of discussion: As is often the danger when Results and Discussion sections are not separated, there is a lack of actual discussion. The discussion section should place the results in context of existing literature. This has not yet been done. Other evaluations of the models used exist. Other modelling studies have assessed different strategies for China or globally Etc. The authors need to place their results into such context, preferably by separating Results and Discussion into distinct sections.

We added the new discussion mentioned above into the manuscript and reorganized the structure to make the paper more readable.

Minor Comments:

[6] Line 85ff.: The authors state that "Significant progresses or consensus have been achieved in community discussions regarding model parameterization, evaluation, calibration, and intercomparison". Given that at least the cited Gleeson et al. stresses the current lack of adequate evaluation strategies for global models, I would personally not frame it quite this positively. I do think that there is still significant advancement needed to derive at adequate strategies, and I also think that consensus is not yet there.

Corrected and cited new relevant papers, e.g., Heinicke et al. (ERL, 2024) and Reinecke et al. (ERL, 2024).

[7] Figure 6. The lower plots show positive and negative deviations from 0. The maps would be much clearer if the authors were to use a diverging color scheme as they do in Figure 7. Though I can also see that the authors prefer to keep the colors similar to the actual values.

Revised.