1) How did the authors handle uncertainty in datasets for potential recharge and soil properties in regions with sparse observational data, particularly in arid and semi-arid zones? Could more details on uncertainty quantification be provided?

Uncertainties of the large-scale hydrologic modeling are largely determined by the uncertainties in the data products used. The generation of the input datasets is always a huge amount of work and separated from the modeling, i.e., the dataset generation and the large-scale hydrologic modeling are the focuses of two different communities and this division will be clearer moving forward in the big-data era. We mentioned that if it is five years earlier, such a modeling is impossible as many global data products were not available. As one of the most important efforts in our modeling, we tried to choose the best available datasets at current stage to reduce potential uncertainties. However, quantifying the intrinsic uncertainties in these data products are out of the scope of our work. Future work incorporating local available data is necessary to further improve the quality of the input datasets or decrease the uncertainties in them. One of our goals in this regard is to keep an eye on the advances in relevant data products in the community and dynamically replace some of the inputs with those of higher qualities at a feasible frequency.

Regarding the selection of datasets in our modeling, we have lengthy discussions for both potential recharge and soil properties. Please refer to lines 204 to 233 and lines 235 to 270, respectively. We also briefly summarized them here as below.

As we mentioned in the manuscript, our objective is to continuously improve the workflow of large-scale surface water-groundwater modeling using ParFlow for community use globally. Therefore, we started from the workflow of CONUS 2.0. Then we found replacements of some datasets, e.g., those existing in US but are absent in China, or those having better ones in China area. For soil texture and deep geology, we used the same datasets GSDE and GLHYMPS 1.0. For flow barriers, there is a better data product for China area, so we replaced the global one by the new one. Also, all of them are the datasets well recognized by the community, i.e., the best choice we can use in China area not only because of CONUS 2.0 using them. Additionally, the combination of these datasets showed outstanding performance when they were tested in three large basins (the Upper Colorado River basin, the Little Washita basin, and the Delaware-Susquehanna Basin) based on ParFlow simulations evaluated by observed streamflow and water table depth.

For potential recharge (P-ET), we compared those generated by different precipitation and ET products and further constrained them with prior knowledge. We collected the latest P and ET products with higher spatial resolutions and long enough durations. Then we further filtered out those contrasting to prior knowledge. This is easy to do as

it is well-known that P and ET products are of high uncertainties. For example, we know there is recharge in the upstream of Heihe River Basin, so the combinations of P and ET generating zero or negative potential recharge in this area will not be considered any more. In manuscript, we also highlighted such significant uncertainties in the products challenging both the data and modeling communities. We also provided a possible solution in future work to generate P and ET products under a unified modeling framework constrained by the water balance.

2) The CONCN 1.0 model covers a vast area at high resolution, which demands substantial computational resources. Could the authors discuss any measures taken to optimize computational efficiency and how the model's scalability could be extended to similar hydrologic regions?

Yes. We used seepage face as the top boundary condition in the first phase of the spinup and then turned on the overland flow in the second phase. This avoids the meaningless surface water-groundwater exchange in the early stage which mainly stabilizes the groundwater. For the scalability, we also have some experience. The CONCN model and the CONUS 2.0 model have very similar dimensions. Therefore, they take approximately the same wall clock time for spinup. Yet due to the larger area of arid and semi-arid regions in China, where the on and off of overland flow (integrated or groundwater only) may take more time to converge. Thus, the spinup of CONCN model takes slightly longer time. Additionally, ParFlow has excellent parallel scalability for different domain sizes and heterogeneities, which has been carefully tested and discussed in Ashby and Falgout (1996).

3) Would the authors consider using coarser resolution or data assimilation techniques to make the model more computationally accessible, particularly for policy-making applications?

Might be a choice but it is really hard to say that this is what we expect. Coarse resolutions will miss a lot subgrid variations which will cause the deviations of the simulation results. This is a well-known issue in the community of earth system modeling. A model with a higher resolution generally shows better performance if the parameterization is reliable enough (or similar). Thus, we are trying to build a high-resolution model to ensure the model performance instead of moving backward.

4) I recommend that the authors consider including a comparison with data assimilation approaches to enhance model accuracy and reduce uncertainties, especially in data-scarce regions. Data assimilation has been effectively applied in hydrologic modeling to integrate observed data with model predictions, often improving the alignment with real-world conditions. Techniques like Kalman filters or

variational data assimilation could complement the current workflow, particularly for improving estimates of potential recharge and water table depth in arid and semi-arid regions where observational data is limited. A comparison with data assimilation methods may also highlight the strengths of the CONCN model and provide a pathway for future enhancements in large-scale hydrologic modeling.

Data assimilation is an efficient approach for incorporating observations and doing parameter inversion. This is in our future plan of our modeling platform. The foundation of a sustainable modeling platform is to build a model of reliable/acceptable performances, i.e., the very first thing. Then the strength of the data assimilation can be fully leveraged. As a result, our first step focuses on the model structure, data selection, spinup, evaluation etc. We also identified the challenges in the modeling. It has been a huge step from scratch, and costs more than two years involving all authors and other collaborators. We are not to build a perfect modeling platform with everything in one paper which is impossible in a short time. We are doing step by step to gradually improve the modeling platform and timely share the results of each step with the community.

Once there is a model with acceptable performance, not only data assimilation but also many other approaches, e.g., emulators, could be incorporated into this modeling framework. The data assimilation improves some of the parameters relying on the observations of some others. This means it still has high requirements of observations. As mentioned by reviewer, it is data scarce in arid and semi-arid areas. Collecting long-term observations of enough spatial density, e.g., water table depth, which has been confirmed useful in data assimilation of groundwater modeling, is a big challenge. In some regions, it is even impossible as the observation network has not been built. Therefore, we also discussed in the manuscript that, moving forward, this modeling platform needs collaborative efforts from different communities.

5) To strengthen the contextual foundation of this study, I recommend the authors cite established integrated hydrologic models like SWAT-MODFLOW in the introduction. SWAT-MODFLOW, widely used for its integration of surface and subsurface processes, has significantly advanced our understanding of coupled surface-groundwater systems across various scales. Citing SWAT-MODFLOW alongside ParFlow and other large-scale models would provide readers with a broader perspective on the tools available for integrated hydrologic modeling. This comparison may also underscore the unique challenges and innovations of applying ParFlow within China's hydrologic and geologic context, while highlighting the importance of diverse model approaches for managing complex water resources.

I strongly recommend to cite below paper:

"Assessing regional-scale spatio-temporal patterns of groundwater–surface water interactions using a coupled SWAT-MODFLOW"

"Assimilation of sentinel-based leaf area index for modeling surface-ground water interactions in irrigation districts"

"Development and application of the integrated SWAT–MODFLOW model."

Our objective is in the framework of large-scale hydrologic modeling. We have done substantial literature review and listed the latest large-scale hydrologic models either in China or at global/national scale, including those using MODFLOW. Though SWAT-MODFLOW is relevant to integrated hydrologic modeling, these three papers are neither relevant to China nor to global/national scale. We fully respect reviewer's strong desire to cite the new published WRR paper, so we cite all three papers in the discussion. Please refer to line 508 in the revised manuscript.

Overall, we appreciate reviewer's interesting thoughts and are pleasant to exchange our ideas on these thoughts. However, these thoughts are more or less deviated from the objective of this work or beyond the scope of this very first and important step. Open questions remain in the large-scale high resolution groundwater modeling. As we mentioned in the discussion, all three groundwater models show different water table depths implying large uncertainties. This is more challenging in an integrated framework and in a data-poor region. Therefore, we are trying to use what we can use in such a region to build a model with acceptable performances (actually, it is unexpected excellent performances) as a reference for the community. The data/datasets collection, selection, processing, assembling the model, fetching computational resources, running the model, and analyzing the simulation results and comparing them with previous models, etc., have been substantial work. We don't aim to finish everything in one step, but to gradually improve it with time.