General comments

The authors present a new hydrologic modeling platform over continental China, based on the ParFlow model, aimed at providing information for surface water and groundwater resources management. Their setup is adapted from the CONUS 2.0 modeling platform over the US. The authors discuss the parameters and input data used and they provide a comparison with modeled and observed data for groundwater table depth and river discharge.

The paper is well written and pleasant to read. The results are interesting and well presented via clear figures. My main concern is about the comparison with other datasets, which may be a little too simple, as detailed below.

We are tankful to the reviewer for all the constructive comments which substantially improve our manuscript. We have addressed them point by point below and revised the manuscript at the corresponding locations.

In particular:

- l. 278-293: Do I understand correctly that for the spinup, the authors used the 1981-2010 P-ET average as constant atmospheric forcing until a quasi-steady state was reached? Is this resulting state used for the evaluation in the next section or have the authors simulated a transient run over 1981-2010, starting from this quasi-steady state? This is not clearly stated, while it is very important for the evaluation and its analysis. For the following comments, I will assume that the authors evaluate the resulting quasi-steady state against other data sets.

  Yes, it is a quasi-steady state model forced by the average P-ET of 1981-2010. We clarified it, please refer to lines 293-294 in the revised manuscript.

- section 4.1. and 4.2.: While the main motivations for this modeling platform are (1) the impacts on water resources of the increased frequency, intensity, and duration of extreme weather events and (2) the management of these water resources, e.g., to prevent water scarcity, the authors limit their evaluation to a comparison of the steady state, which represents an idealized situation that never happens in the real world. In particular, the ability of the modeling platform to represent the dynamics (temporal evolution) on a yearly or better monthly or even daily time scale is not considered in this study, while this would be essential to assess whether the modeling platform is able to meet its primary aim (i.e., the aforementioned motivations).

Reviewer well summarized some of the main motivations to develop this modeling platform. This work is the very first and very important step of this modeling platform. The current model is not the whole thing. This first step aims to build the foundation of the modeling platform, focusing on the model structure, parameterizations, data selection and processing, model assembling and spinup, observation data collection and cleaning, comparison of model formulation and simulation results with other models or datasets, and identify the challenges and requirements to move forward (beyond the motivations summarized above). This step costs a team of more than 10 people more than 2 years (all authors and others not listed).

A steady state model representing a long-term average state is important to demonstrate the general reliability of the current modeling formulation in the target area and unravel the remaining deficiencies in the modeling community. This work is an important reference and/or inspiration for the large-scale hydrologic modeling community. This steady state model will be the starting point for the transient run. Starting from a multi-year averaged state will generally reduce the spinup time of the transient run for a specific year. It is computational expensive to run the model, so we first build the steady state model and then run selected years when needed, according to the requirement of focused objectives. One more reason is we will run the transient state model using ParFlow coupled with the latest Common Land Model. The latest Common Land Model has increased functionalities and could be helpful to better understand the hydrologic cycle in China. The workflow and the necessary data based on the new coupling model are under preparation, which is again a huge amount of work.

- section 4.1.
    - esp. l. 347-348: Do the authors use the longest available period for each gauge or the longest overlapping period (i.e., max 9 years between 2002 and 2010)? In any case, this relies on the hypothesis that an observed average over a few years (sometimes even only two years) as well as an observed average over two to several years covering another period (2002-2021) is representative for a steady state based on 1981-2010. I am not convinced that this hypothesis is true. I could agree that, the longer the observation period is, the closer the average gets to a steady state over the same period, even if this should still be verified. But in my opinion, there is no guarantee that the average over 2002-2021 is representative for the 1981-2010 steady state as this ignores potential shifts in the terrestrial water regime, e.g., due to climate change. One could think of the impact on

streamflow of earlier snow melt, less snow accumulation in winter, more frequent extreme events, multi-year droughts, etc. Assuming that an average over a few years is representative for a steady state might be even more questionable, since these few years could be characterized by extreme events (droughts, floods). The resulting average would certainly not correspond to a steady-state, preventing a robust comparison. The authors already mention the potential impact of hydraulic engineering (e.g., dam operations) on the average streamflow (see l. 354), especially over short time periods.

We used all data available during 2002-2021. We fully understand the reviewer's concern. This is exactly one of the biggest challenges we encountered in the modeling in China. Collecting, cleaning, and processing observations are the most time-consuming part in our modeling. In China, we don't have a fully-open, public access to observations of streamflow and water table depth. We contacted many people or institutions and gained little. Also, the monitoring networks in China started very late, e.g., the groundwater monitoring network started with a small number of wells (~900) from 2005. Although the streamflow is slightly earlier, it is not as earlier as that of USGS which could date back to 1900s. The only way we can get observations is to digitize them from the yearbook which again is very time consuming. The final scheme used in our paper are the best we can do at current stage considering the normal/acceptable duration for academic outcomes. We highlighted in the manuscript that this challenge largely hampers our modeling and expect the conversations and collaborations with the data monitoring community. This is also one of motivations of this modeling work. Data sharing or public access is urgent to break this bottleneck and may need policy support. More importantly, this is a modeling platform of dynamic efforts, and the current work is the first step. We have collected more data after the submission and will clean and process the data and incorporate them into the evaluation in future work. Please refer to lines 331-338, 390-400, 486-505, and 516-532 in the revised manuscript for the relevant discussion here.

- section 4.2.
  - the structure of this section might be improved to make it easier for the reader to follow. For example, after the first paragraph (l. 364-371) describing Figure 6, one would expect a first analysis of these results. But this analysis only starts on l. 432.

Revised. We plot Figures 6c and 6d in a new Figure 9.

We understand the reviewer's concern. The original logic here is that we first introduce all the materials we used to evaluate the model, i.e., the global datasets, the observations, and the GRACE data. Then we first analyzed the scatterplots of simulations vs. observations, then the residuals vs. GRACE data, and finally the uncertainties remaining in the groundwater models in the community (i.e., Figures 6c and 6d). We may plot Figures 6c and 6d in a new figure after figure 8, but it may prevent audience to compare 6c and 6d with 6a and 6b. We also tried to analyze some after Figure 6, but it is hard to get general conclusions before the analysis of scatterplots and GRACE data.

- l. 364 and Figure 6: Are the steady states of the two global datasets over the same period as for CONCN (i.e., 1981-2010)? If not, is the hypothesis valid that these steady states, which might have been reached under different climatic conditions, are comparable? For example, if one region experiences less (or more) precipitation and/or higher evapotranspiration due to climate change, the resulting steady state will very likely be different.

  Clarified. Please refer to lines 382 and 385.

  Additional thoughts please refer to the response to the comment '*l. 259: Why did the authors use the period 1981-2010 and not, e.g., 1991-2020? This could have made the evaluation easier, as the authors state further below that more observations are available for the last years (esp. since the 2000s).*' in this letter.

- l. 397-416: In the same way as my comments above for the evaluation of streamflow, I do not see any reason why one could assume that the observed average over 2018 could be considered as representative or close to a steady state generated with data from 1981-2010. Especially for water table depth with a potentially huge impact of inherited conditions from previous years (memory effect), not only 2018, but also the previous years would need to be close to the 1981-2010 average hydrologic regime to – maybe – approach a steady-state-like state. I understand that this is the reason why the authors try to strengthen their evaluation with the analysis of the residuals in the context of the long-term trend from GRACE, thereby trying to make a link between the steady state based on 1981-2010 and 2018, but, in my opinion, the uncertainty of this whole evaluation remains high. The only way to provide a robust and representative evaluation is to do the comparison over a common period.

Please refer to the response to the evaluation of streamflow above for the first part of this comment. Additionally, we used 2018 as it is the first year of the expanded national groundwater monitoring network (> 8000 wells). A much fewer wells (~900) are included in the earlier monitoring network and are mainly distributed in the east China. It is hard to balance the quantity and duration of the observations. Considering the slow variations of groundwater (i.e., the long correlation/memory), we finally used 2018 of more wells. The comparison of residuals with GRACE is an additional approach to evaluate the model and highlights the uncertainties in existing groundwater models in the community. Yes, again, we recognize the mismatch of the durations between simulations and observations is a concern, yet this is the best we can do at current stage.

- l. 401-404: This might be easier to understand if the authors could briefly explain why this analysis integrating GRACE data is needed.

  We aim to evaluate the model use multi-source of data generated by different approaches, especially in such a data poor region. Multi-source data can provide cross-evaluation to improve the reliability of the modeling.

- l. 411: if the global models are calibrated, do they not implicitly account for human interaction, via the observational data used for calibration? This would then be contradictory with the statement in l. 406.

  The global models we cited were calibrated based on observations without explicitly considering human activities, e.g., groundwater pumping. Their calibrations were done based on observations mainly in America and Europe instead of China. This is also due to the data scarcity in China. Then the calibrated models generated the simulation results we used in our study which were not constrained by the observations in China. This is the first time to evaluate their results with observations and GRACE in China area.

- l. 421-422: In l. 377, I understand that these grid cells are excluded from this analysis (precisely for the reason explained here). Please clarify this.

  Clarified. Please refer to line 391 in the revised manuscript.

Specific comments

- l. 54-55: While I agree that it is pressing to develop a modeling platform accounting for it, this statement suggests that CONCN accounts for water quality control, which is not the case.

  We also have the particle tracking system which can simulate water ages and have implications for water quality. This is also a component that will be added into the modeling platform. To avoid confusing the audience at current stage, we removed the water quality in the revision.

- l. 99: About the "unique dramatic topographic relief". On one side, each part of the world has a unique relief, thus I could agree with this formulation. On the other side, many other regions (e.g., the US, South America, Africa, Europe, New Zealand, Japan, etc.) have transitions from mountains to coastal plains, thus facing similar challenges for hydrologic modeling.

  Revised.

- Figure 1: What is the meaning of the white coloring within the model domain in Fig. 1f? Here, I would interpret it as "no data", is that correct? If it is zero, it should be colored according to the color bar (i.e., dark blue). If it is "no data", how do the authors deal with it as source-sink term for ParFlow? This should be clarified in the text.

  Clarified. These areas have P-ET of 0 in the model. We cannot show them in log plot.

- l. 167-169: The procedure is not clear to me. Did the authors generate D8 connectivity slopes in addition to the aforementioned D4 slopes? If yes, why was it needed? What do they mean by "vector networks"?

  We generated D8 networks as the input of priorityflow to generate the final D4 networks we need. We compared the D8 networks we generated with the vector network generated from the higher resolution MERIT Hydro to avoid obvious errors in the inputs of priorityflow.

- l. 180: How are the sinks handled? Is the inflowing ponding water removed before/after each time step?

  Yes. A specific key in ParFlow did this automatically.

- l. 200: Why do the authors derive the soil texture from this global dataset instead of using directly the soil hydraulic properties?

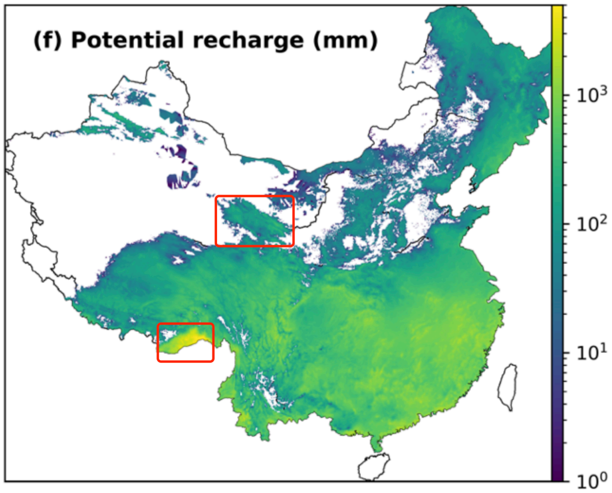
  Could be an option. Two more reasons: 1) wanted to keep consistent with the CONUS 2.0 workflow and 2) our further processing will generate soil textures fewer than the original data which can help the convergence of the model. This has been tested in our North China Plain model in previous studies.

- l. 208: What is meant by "flow barriers"? Is the permeability in these grid cells further reduced by a factor or set to a very small value?

  Yes. Clarified. Please refer to line 218.

- l. 250: "several locations" It might be useful to add some more details here: How many locations? Are they distributed over the country and/or hydroclimatic regions to ensure a representative analysis and selection of datasets?

  This is a qualitative filtering. We mainly focused on the following two locations together with other judgements without a specific location. If the P-ET is negative in the top rectangle or the precipitation is not obviously higher than areas around in the bottom rectangle, we filtered out that combination of P-ET. This is easy to do as it is well known that the P and ET data products, especially ET, are of high uncertainties. This is a common challenge in the community.



- l. 259: Why did the authors use the period 1981-2010 and not, e.g., 1991-2020? This could have made the evaluation easier, as the authors state further below that more observations are available for the last years (esp. since the 2000s).

  We tried to develop a steady-state model that can represent the steady state. This requires the long enough P and ET products in a period without intensified human activities (i.e., better before 1950). This is easy to realize in data-rich US but a challenge in China. If we develop a model forced by P-ET of 1991-2020, the P and ET have been affected by human activities. However, the human activities and the uncertainties are even harder to quantify and represented in the model, which will make the evaluation of the model harder.

- l. 275: It might be useful to add a source or some additional explanation on how Manning's coefficient is "set to vary by land cover type".

  Clarified. Please see lines 285-287.

- l. 284: What is the role of the seepage face boundary condition?

  This is to speed up the spinup. In the early stage of the spinup, the state of groundwater is far from the final quasi-steady state, so the interactions between groundwater and surface water are meaningless. Hence, we used seepage face instead of overland flow at the beginning to reduce computational load. When the groundwater is almost steady (the river channels are generated), we turn on the overland flow. Please refer to lines 301-302.

- l. 284-285 and 287: On which time scale does the total storage change have to be less than 1% (resp. 3%)? Is it e.g., between two consecutive time steps or on an inter-annual basis?

  Either is fine. I did the latter one.

- l. 292: I understand that it is important to reach an equilibrium for groundwater and for river discharge, but does a quasi-steady state for discharge in arid and semi-arid regions really make sense? Is the resulting discharge not too far away from reality? I would guess that in reality, the discharge is highly variable in these regions, with very low flow, or even no flow at all, most of the time alternating with high discharge after precipitation events or snow melt.

  Good point. Actually, the current modeling has bigger challenges than reviewer's concern as the large intrinsic uncertainties in P and ET datasets, especially in ET. The simulation results are unsatisfied in these areas such as the Endorheic and the Hai River Basins, which we highlighted in the discussion of Figure 5.

- Figure 3: How can the authors explain that they have streamflow values everywhere and not just in the streambeds in the south-east and north-east of the model domain? Or is it just an impression due to the visualization of a dense hydrographic network?

  Partly, yes. Additionally, the surface water and groundwater shared the same head in the top layer. Therefore, the pressure used to calculate the streamflow is actually 0.05 m below the land surface. Therefore, in areas with water table depths smaller than 0.05 m, there are also 'streamflow'.

- Figure 3: It would be useful to add in the caption which period is shown. Or is it the end of the spinup (i.e., resulting quasi-steady state)?

  Added

- Figure 5: It might be useful to add in the caption that the gauges are grouped per basin as shown on Fig. 1b.

  Added

- Figure 7 and in the text: Do the "residuals" correspond to the difference between CONCN and the observed values at the wells?

Clarified.

- l. 453: All regions in the world experience increasing extreme weather events such as droughts and floods. What may make China “one of the most significant ecohydrologic hotspots in the world” could be the intense water use in the highly populated areas of the country. However, this is not accounted for in the model platform presented in this paper.

  Yes, water use is important. This work is the very first step of this modeling platform and we will consider water use in the future work.

Technical corrections

- l. 29: Meaning of RSR?

  We clarified it in lines 124-127. As it is too long to explain it in the abstract.

- l. 56: Correct “ with a 10 km resolution”.

  Corrected.

- l. 67: Meaning of USGS?

  Clarified.

- l. 112: Correct “key components of the ParFlow model”?

  Corrected.

- Figure 1: The north-eastern edge of the domain is hidden behind the color bars.

  Revised.

- Figure 1: In the caption, what do “f.g.”, “sil.”, and “c.g.” stand for?

  Clarified.

- l. 154-156: For clarity, it might be good to specify that this concerns each grid cell individually, e.g., something like “D4 connectivity means that, within each grid cell, streamflow is allowed…”.

  Corrected

- l. 182: Correct “with those in IHU”?

  Corrected

- l. 199: In l. 125, the thickness of the second layer (from the top) is 0.3 m. Here, it is indicated to be 0.4 m.

  Corrected

- l. 260: Correct “Tarim River Basin”?

  Corrected

- l. 267: It is important to expand the acronym of CLM to avoid any confusion, as nowadays CLM usually means “Community Land Model”, while the CLM integrated in ParFlow is the “Common Land Model”.

  Clarified

- l. 340-343: There is a mismatch in the number of gauges: 95 (total) – 6 (no location) – 1 (close to another) – 1 (outside of domain) = 87, not 88.

  Corrected. It should be 95-5-1-1=88.

- Figure 7: Correct “The background shows the average decrease of groundwater storage”.

  Corrected

- l. 397: Correct “by the three models”?

  Corrected

- Figure 8: Indicate in the caption that you compare the steady state over 1981-2010 with observations from 2018.

  Indicated

- l. 433: Correct “and the two global models”?

  Corrected

- l. 435: Correct “across the three models”?

Corrected

- l. 445: Correct “below – these require” or maybe “below. These require”?

  Corrected

- l. 515: Correct “have been cited”?

  Corrected

- l. 524: Correct “reported in this paper”?

  Corrected

- l. 525: Correct “which is a consortium”?

  Corrected

- l. 526: Correct “and the Office”?

  Corrected