Comment on gmd-2022-173
Anonymous Referee #3

Referee comment on "Continental-scale evaluation of a fully distributed coupled land surface and groundwater model ParFlow-CLM (v3.6.0) over Europe" by Bibi S. Naz et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-173-RC3, 2022

In their manuscript Naz et al. evaluate a pan-European, high-resolution (0.0275°) simulation with the coupled land surface groundwater model ParFlow-CLM, using observations and re-analysis data for streamflow, near-surface soil moisture, evapotranspiration, water table depth and snow water equivalent. In general, the manuscript is well written, the metrics for evaluation seem to be appropriate and the authors go into great detail discussing the potential sources for some of the biases – with respect to possible shortcomings of the model but also of the observational data. Having said that, there was one aspect of the evaluation that did not fully convince me, namely the evaluation of the simulated water table depths, where only the anomalies were being investigated. I understand that it may not be easy to define the reference elevation, but with the sophisticated ground water fluxes being the key component of the model that sets ParFlow-CLM apart from most LSMs, the authors should really think about discussing a comparison of the absolute values – maybe indicating the uncertainty due to the reference surface elevation. Also I did not understand, why the authors limited their comparison to those points with simulated WTD < 10m?

However, my main concern is that I found it somewhat difficult to connect the results to the motivation outlined in the (very well written) introduction of the paper. A large part of the latter is focused on the shortcomings of LSMs and GHMs and their -- admittedly extremely simple – representation of (subsurface) processes. So I would have welcomed a comparison between ParFlow-CLM and a CLM version without ParFlow – possibly the one that is part of ParFlow-CLM -- or with a LSM that includes some simple parametrization of ground water flow (e.g. CLM5 [Felfelani et al., 2021]). Furthermore, the authors indicate that the resolution of the model is important, which I am very willing to believe. Yet they do not show how this affects the simulations in case of their model. Here, a convincing case may have been made by comparing their simulation to the 12km runs in Shastra et al. (2021). If the authors do not want to include an inter-model/-resolution comparison maybe they could think about a different approach to the paper: E.g. as an alternative, the authors could have referred to the study of O’Neil et al. (2021) from the beginning and then set up the paper as a comparison of ParFlow-CLM simulations of Europe and of the CONUS region?

Finally, it is not always easy to estimate whether the model really captures a given variable well or not. E.g the authors state that the model “appropriately captures the seasonal cycles” of the WTD (l. 419). However, with only 20% of the investigated cells
exhibiting an $R > 0.5$, it is debatable whether or not this is appropriate. Again, it would have been much more straightforward if the simulation had been compared to a different model / resolution and the question would have simply been about better or worse than XYZ. Without such a comparison, I am not sure that all of the claims made by the authors – e.g. "the added value of capturing heterogeneities for improved water and energy flux simulations in physically-based fully distributed hydrologic models over very large model domains" (l. 16 ff) – are substantiated by their results.

Additional comments:

I. 144) (Annoying detail, but) I think that here CLM refers to Community Land Model, while CLM was defined in l. 121 for the predecessor Common Land Model.
I. 171) Why do you loop a single year to force the model? Doesn't that include the risk of running the model to a non-representative equilibrium state? Also, how did you decide that a 9-year spin-up is enough and how were the states initialized, that a 9-year spin-up is sufficient?
I. 218) What specific data was assimilated?
I. 226) I think it could also be really interesting to compare SM profiles at the stations in addition to the top layer SM.
I. 269 ff & Fig1) As you indicate a strong dependency on topography, could you maybe include a plot of the topography in Fig. 1. Also, why is the SWC so low and the WTD so high right next to the river?
I. 289 f.) In case of the Rhine (gauges 2-5) the model appears to underestimate the discharge quite a bit, would this still be explainable by human impacts? Or could it not point to an underestimation of P-ET?
I. 290) I am not sure that everyone is so familiar with the KGE as to immediately know what the range of values indicates. Could you maybe add a very brief explanation here? Fig2.) I found it a bit hard to identify the gauges in subplot a, do you think it would be possible to zoom in over the center of the first subplot?
I. 298) I think something went wrong referencing the figure.
Fig3) Could you clarify that the color-code in panel c is the same as in b?
I. 339 ff) How can you be sure that the differences are a result of the different treatment of the lateral groundwater flow? I thought that between CLM3.0 and 3.5 there were also major changes in the terrestrial hydrology – e.g. a TOPMODEL approach to runoff generation and changes to the evaporation calculation?
I. 352) Not Fig. 4c?
I. 353) The R values in subplot 4c go beyond this range.
Fig 4.) When comparing ESACCI and ESSMRA in subplot b, these seem to agree much better than ParFlow-CLM agrees with any of the two datasets. As ESSMRA is the closest to a second model that is shown in the study, one could come to the conclusion that the added complexity of the explicit treatment of groundwater fluxes in PArFlow-CLM does very little to improve the near surface soil moisture. Thus, it would be very helpful if the authors could describe in more detail what was assimilated in ESSMRA, because if it was soil moisture directly then the good agreement between ESACCI and ESSMRA is not very surprising. Otherwise it would be very interesting to understand why the ESSMRA appears to be so much closer to ESACCI.
I. 387) Could this overestimation of ET also be a reason for the underestimation streamflow in the Rhine?
I. 417) I think something went wrong referencing the figure.
I. 419) Here I was a bit surprised at the comparatively low R values. Given that precipitation is prescribed based on observations and that both streamflow and ET show a much better correlation with the observations, does this indicate that the model is missing
something important in the representation of the groundwater dynamics?