Review of **Multi-variable, multi-configuration testing of ORCHIDEE land surface model water flux and storage estimates across semi-arid sites in the southwestern US** by MacBean et al.

The manuscript by MacBean et al. deals with two different soil schematizations of the ORCHIDEE land surface model. One model set-up consists of a 2-layer soil schematization, whereas the other set-up makes use of an 11-layer soil scheme. In addition, resistance for soil evaporation was varied and bare soil fractions were reduced. The model set-ups were evaluated for several sites in the southwestern US. The authors show that adding a more detailed soil schematization improves the model results, especially regarding total evaporation and high frequency moisture dynamics.

The manuscript is generally well-written, and the figures are clear and of high quality. Most of the statements are supported by the data, and I think the article is interesting, because I agree that the hydrology in LSMs deserves attention. Nevertheless, after reading the article, I have several questions that remain.

One of the first things the authors observe is that the forested sites show differences in transpiration and soil moisture. The soil schemes are different between the model runs, but rooting depths and rooting profiles are hardly mentioned by the authors. However, different rooting depths for both set-ups will have a strong influence on the findings of the authors. So how are these parameterized and are these different for the different model set-ups?

Similarly, the authors also often refer to the low and high elevation sites, but these also come with different vegetation types (forested vs grass/shrubland). I think the different vegetation types are much more the reason for the differences between the different sites, so I suggest that the authors distinguish more between the different vegetation types instead of the elevation, especially in the figures.

The authors also decided to model the soils with a thickness of 2 m, and mention that for the 11LAY-model drainage occurs as free gravitational flow at the bottom of the soil. This thickness, which is rather arbitrary, will also have a strong influence on the results as presented. The groundwater tables may influence the soil moisture profiles, and I wonder therefore if the authors have some idea on the groundwater tables at these sites. I do not object to this model choice of a 2 meter soil thickness, as you probably have to make an assumption here, but I believe it would be good to reflect on it, especially as the goal of the authors is to get the hydrology right, from which the groundwater is an important aspect and that is now basically assumed to be negligible.

There are also two methods used to derive ratios of transpiration/evaporation (Figure 6), but also here I have several questions. First, I wonder what the difference is between the two methods and if it is a fair comparison. There is also no data in the first months, and no data for US-Vcp, why is that? In addition, at US-Fuf, the data-derived estimates show that almost half of the total evaporation is transpiration, even during winter. At the same time, the site is described as having snow, at a high elevation, and one would therefore expect hardly any transpiration in winter here. This is also what the model actually does, it shows a strong reduction during winter. So how reliable are the estimated observations here?

The authors often argue that snow is not correctly modelled, and I think the statement of the authors on page 14, lines 442-444 is important here. Snow usually falls within a temperature range around 0 degrees Celsius, and the authors mention that the results improved by changing the temperature threshold, but these results are not shown, so please add these results.
In addition, the reasoning of the authors regarding the snow modelling relates to the overestimation of ET at US-Fuf for 11LAY, but this does not happen for 2LAY. At the same time, US-Vcp also shows an

underestimation and has snow, so it does not seem to be a consistent problem here. Do the two model set-ups use the same snow module and are the parameterizations the same for the different sites?
As suggestion, it could also help the authors to look at remotely sensed snow cover products such as MODIS10A. These products are relatively easy and could provide already a quick check if the snow temporal dynamics are captured in the model.

My most important point relates however to the fact that the article misses sometimes a bit focus regarding the goal of the authors, which is comparing a simple two-layer scheme with a more complex scheme in order to improve the hydrology. A couple of times the authors only look at the 11LAY-results, or do not use observations to assess if there are any improvements. For example, the authors only compare 11LAY with the soil moisture measurements (Fig. 4,5, paragraph 3.2). I do understand why, as the authors explain this in paragraph 2.3.2, but I am not sure if there is any point in evaluating 11LAY-results with soil moisture data, if you can not do the same for 2LAY. After reading paragraph 2.2.2 I still think the authors could at least compare also the temporal dynamics in the 2LAY-model, as this is what the authors do anyway with equation 5.
Similarly, a large part of paragraph 3.1 gives a description on the differences between the two model set-ups, and discusses Figure 1. Nevertheless, without any idea on how reality looks like, it is hard to really get an understanding on what is actually better. So I am not sure if this part of the paragraph really adds something, unless the authors add some observations. The authors do have soil moisture data and flux tower data, so I suggest to add these to Figure 1.
One of the main conclusions is also that the high frequency soil moisture dynamics are more realistic for the 11LAY-model. This conclusion is however not supported by the data as shown, there is no figure in the manuscript and supplementary material that actually compares both 11LAY and 2LAY soil moisture values with observations, so you can unfortunately not state that 11LAY is clearly better here. The conclusion that surface runoff is more realistic (P21.L669) came even as a bigger surprise to me, I believe there is no data on surface runoff in the manuscript, or I must have completely missed this.

Concluding, the manuscript is interesting, but the authors should make sure they build a systematic case why one hydrological schematization should be preferred over another. I have sometimes the feeling the authors have a preference for the 11LAY-scheme, but I think it is important to objectively assess the performance of both set-ups. I hope my comments are useful for the authors and look forward to an improved manuscript.

Minor comments
P1.L36. Results better → results in a better?
P2.L62. A evaporation → an evaporation
P3.L79 have been rarely been → have rarely been
P4.L115. Define PFT
P6. L187. What do you mean with soil tile? The spatial distribution of different soils within a grid cell?
P6.L189. “all three PFT’s” –> It is mentioned before that there are 12, so why three now?
P6.L191. Related parameters) → remove “)”
P7.L210. At al → et al
P7.L217. At al → et al
P8.L227. Seems a bit arbitrary to me, why these numbers?
P8.L229. Has been test → have been tested
P8.L256. The the root density → the root density
P8.L256-257. Why these values? What are they based on?
Eq3. Please define and describe also h_t and d
P8.L267. Is T here transpiration? Please define.

Eq5. Please define your variables
P12.L351. Higher compared to the other sites? It is not higher than the 11LAY-scheme.
P12.L380. I do not see any values going to 0 in Figure S1 for VWC in the upper 2m. Basically 2LAY seems to drain the upper layer faster.
P12.L383-384. I do not think you can conclude 11LAY is better based on the data as shown, there are no observations shown of soil moisture in Fig. 2.
P14.L421. Fig 4 → Fig. 4
P14.L422. So which sites in fig4 do you mean? It's easier to add the names, then the reader knows where to look.
P14.L445-448. Where can I see this? Please make sure you back up your conclusions by showing the evidence.
P15.L460-480. I was a bit confused by the term evaporation E, whereas you also discuss evapotranspiration ET (which are often used interchangeably), but you mean here interception evaporation, correct? For clarity it might be good to add a subscript Ei and talk about interception evaporation.
P15.L467. You mention before that US-Vcp underestimated ET, instead of overestimated.
P16.L480. Are be responsible –> are responsible?
P17.L517. You do not show that T/ET fractions are better with the reduced bare soil fraction.
P17.L523. TeNE-forest?
P17.L529. Spring → spring
P19.L592. ORCHIEE → ORCHIDEE
P21.L669. I am not sure how you can conclude this without runoff data and never evaluating it.

Table3. Please note that RMSE also has a unit

Figure 3. The unit is mmm-1, I believe you mean mm/month, but please make this clearer.
Figure 6. Why not include also the 2LAY-estimates? There are two methods used to estimate the ratios for the high and low elevation sites, is this a fair comparison then? Why is there no data for the first months? Why no data for US-Vcp?

Figure 7. Why would you average over all the sites? This is just removing information, please show all sites individually, there is no point in lumping this together.

Figures S5 and S6. Please add units and a legend. And as these are regressions, why are there no data points shown? I only see a regression line, so I am not sure how to interpret these figures.

Data availability:
Where are the model results shared?