Response to Reviewer #2

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.

We thank the reviewer very much for their useful and comprehensive review and constructive comments. We have attempted to provide detailed responses to all general and specific comments below. Please note that responses to the reviewer are in blue and additions to the manuscript are in red. Small changes to existing sentences are given in italics within the original sentence.

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?

We have explained the rooting density profiles used for each PFT in Section 2.2.4 in the original text: “Whichever the soil hydrology model, beta depends on soil moisture and on the root density profile R(z)=exp(-c_j z), where z is the soil depth and cj (in m-1) is the thethe root density decay factor for PFT j. For a 2m soil profile, cj is set to 4.0 for grasses, 1.0 for temperate needleleaved trees and 0.8 for temperate broadleaved trees.”

We have added “In both model versions” before “For a 2m soil profile”.

We have kept the same rooting depths for both model versions, so this is not influencing the differences between the versions. We agree that changes in rooting depth can change the hydrological fluxes; however, this was not the aim of our paper so we do not want to test that further here. But we agree our discussion on roots was limited. We have also added in one extra point in the explanation of saturated hydraulic conductivity (in Section 2.2.2) that is related to roots following a comment by Reviewer #1 about infiltration differences between the tree and grass PFTs:

“Ks increases exponentially with depth near the surface to account for increased soil porosity due to bioturbation by roots”

Further to the changes made for Reviewer #1’s infiltration comment (mentioned above), we have added some text in the results discussion to highlight that the root density decay factor may need to be adapted for semi-arid ecosystem PFTs:

“ it is possible that the model description of a vertical root density profile, which is used to calculate changes in Ks with depth, is too simplistic for semi-arid vegetation that typically have extensive shallow root systems that are better adapted for water-limited environments. It is also possible that assigning semi-arid tree and shrub types to temperate PFTs, as we have done in this study in the absence of semi-arid specific PFTs, has resulted in a root density decay factor that is too shallow.”

Finally, in various discussion section where we have talked either about the need for parameter calibration or issues with lateral redistribution of moisture, we have added the need to calibration root density profile or root zone plant water uptake parameters, and we’ve added that LSMs do not currently simulate extensive shallow root systems that are typical of semi-arid vegetation that is more adapted to water limited conditions. We hope these additions significantly improve the discussion related to rooting depths.

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.

This is a fair point by the reviewer. We mainly followed this distinction because of the differences in precipitation regime and sources of available moisture throughout the year. The higher elevation sites are partly driven by snowmelt (as we discuss) as well as monsoon rains, whereas the lower elevation sites’ moisture availability predominantly comes from monsoon rains. This is fundamental to explaining why forests exist at these higher elevation locations but not at all in the lower elevations. But we agree that for most of the text adding in high or low elevation is not needed, so we have removed a good chunk of those references. And to be clearer as to why we talk about differences in high and low elevation in addition to the type of vegetation we have added the following text into the site description in Section 2.1:

(for the low elevation sites): “The four grass- and shrub-dominated sites (US-SRG, US-SRM, US-Whs and US-Wkg) are located at low-elevation (<1600m) in southern Arizona with mean annual temperatures between 16 and 18°C (Biederman et al., 2017).” and “Moisture availability at these low elevation sites is predominantly driven by summer monsoon precipitation; however, winter and spring rains also contribute to the bi-modal growing seasons at these sites (Scott et al., 2015; Biederman et al., 2017).”

(for the high elevation sites): “Both high elevation sites experience cooler mean annual temperatures of 7.1 and 5.7°C respectively and are dominated by ponderosa pine (Anderson-Teixiera et al., 2010; Dore et al., 2012). The high elevation forested sites have two annual growing seasons with available moisture coming both from heavy winter snowfall (and subsequent spring snow melt) and summer monsoon storms. ”

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.

We appreciate the reviewer’s comment here but we are inclined to disagree that the ultimate goal is to make the hydrology *exactly* “right” because, as we discuss in the introduction and discussion, there are many aspects of the hydrology models that we already know are not implemented - groundwater being a good example. We have not set out to test every single parameter that contributes to the soil hydrology schemes (including soil depth and rooting density decay factor etc). The resulting manuscript would be too large; although we agree that both these parameters (and many more of the assumptions that go into the model, as well as known missing processes) could affect the model results. We have done our best to caveat and discuss these decisions and limitations in the discussion.

In an earlier version of the manuscript, we did have a small discussion section on the impact of soil depth (and texture). We removed it because many co-authors thought the paper was already long enough and we had to prioritize the points we discussed. However, if the reviewer would like we can add it back in. It read:

“**Soil texture and depth**
Total water content is unquestionably dependent on both the texture and depth of the soil, which is fixed at 2m in the 11-layer discretized hydrology model. However, semi-arid region soils are likely to be shallower with a higher concentration of rock and gravel (Grippa et al., 2017) – both of which are not represented in the ORCHIDEE soil texture classes. These two issues could introduce a bias in the soil moisture magnitude that is not easy to assess with the current observations. The inclusion of a mechanistic surface hydraulic conductivity parameter in the 11LAY version has allowed more water to be partitioned as runoff. However, it is nevertheless possible that too much water is still being held in the soil as a result of an incorrect soil depth and texture; in reality, more water might be partitioned to runoff or drainage. Ultimately, more different types of observations (such as runoff) are needed to test multiple different model versions.”

In terms of groundwater at these sites, the depths to the groundwater are much deeper at these sites (10-100s m depth) and therefore groundwater access is not thought to be a large

contributor. We have added in a sentence into Section 2.1 describing the sites to add that groundwater depths are typically 10s to 100s metres deep.

Several previous studies have shown that in ORCHIDEE, only rather shallow WTDs are likely to significantly increase ET. As an example, Fig 4g in Wang et al. 2018 indicates that a forced WT at a depth of 1m from the soil surface can increase ET by more than 2.4 mm/d in SW USA. In this area, complementary work, but not yet published, shows that ET increases of 1% or more can be achieved with WTDs down to 5m or less (Ducharne et al., submitted).

Campoy A, Ducharne A, Cheruy F, Hourdin F, Polcher J, Dupont JC (2013). Response of land surface fluxes and precipitation to different soil bottom hydrological conditions in a general circulation model. JGR-Atmospheres, 118, 10,725–10,739, doi:10.1002/jgrd.50627.

Wang F, Ducharne A, Cheruy F, Lo MH, Grandpeix JL (2018). Impact of a shallow groundwater table on the global water cycle in the IPSL land-atmosphere coupled model, Climate Dynamics, 50, 3505-3522, doi:10.1007/s00382-017-3820-9

Ducharne A, Lo MH, Decharme B, Chien RY, Ghattas J, Colin J, Tyteca S, Cheruy F, Wu WY, Lan CW. Compared sensitivity of land surface fluxes to water table depth in three climate models. Submitted to Journal of Hydrometeorology.

However, in relation to one of Reviewer #1's comments we have proposed adding a few new sentence in the discussion about the need for groundwater to be included to the model:

"When discussing woody plant responses to drought, it is also worth noting that many LSMs to date are also missing any representation of groundwater (Clark et al., 2015). As described in Section 2.1, the water table is typically very deep (10s to 100s metres) at these sites. Previous modeling studies have shown that only rather shallow water tables (~1m) are likely to significantly increase ET in the SW US (e.g. by 2.4mmd-1 in Fig. 4g of Wang et al., 2018). However, the fact LSMs typically do not include adequate descriptions of groundwater access could impact their ability to simulate savanna ecosystem dry season water uptake given that drought deciduous shrubs in Mediterranean and semi-arid ecosystems are more resilient to droughts due to their ability to tap groundwater reserves (e.g. Miller et al., 2010). A new groundwater module is being developed for ORCHIDEE and will be tested in future studies."

The answer to the question of why we decided to use a thickness of 2m is similar to our answer to Reviewer #1's question of why 11 layers (and why not 100). The discretization of the soil column and the depth have been tested in previous studies testing the implementation of the finite difference integration needed to solve the Richards' equations in De Rosnay et al. (2000) and are now set as default parameters in the model. Aside from comparing these two schemes, and some additional tests related to the bare soil fraction and bare soil evaporation resistance term (which we decided to test based on the most obvious model-data discrepancies we found), we do not attempt to test any of the other options that may contribute to differences in water stores and fluxes for the reasons given above - it is too much for one paper. Rather, in the absence of insights that these

parameters may be the main cause of model-data discrepancies, we prefer to leave most of the parameters (such as soil depth) set to the default values that have been set based on these previous studies. This has the additional benefit of providing a reference as to how the default model (used in ongoing CMIP6 simulations) compares to observations for this region. We have added the following on to the end of the sentence that originally detailed that 2m soil depth was used for both versions:

“In this study, the depth of the soil for both schemes is set to 2m based on previous studies that tested the implementation of the soil hydrology schemes (de Rosnay and Polcher 1998; de Rosnay et al., 2000; de Rosnay et al., 2002). ”

We have also added the following sentence to the hydrology model description in Section 2.2.2.

“De Rosnay et al. (2000) tested a number of different vertical soil discretizations in a 2m soil column and decided 11 layers was a good compromise between computational cost and accuracy in simulating vertical hydraulic gradients.”

de Rosnay, P., Bruen, M. and Polcher, J.: Sensitivity of surface fluxes to the number of layers in the soil model used in GCMs, Geophysical Research Letters, 27(20), 3329–3332, doi:10.1029/2000gl011574, 2000.

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 reviewer is absolutely right that we did not outline the difference between the two methods to derive the T/ET ratios. We also did not explain the S&B17 method well and we did not explain the Zhou et al. (2016) method at all in the methods. We also did not provide Zhou estimates for US-Vcp. These were oversights by the authors. We have corrected all these issues in the revised manuscript but the reasons are explained below.

Initially, we used Scott and Biederman (2017) for the low elevation more water-limited shrub- and grass sites because it was deemed that this method is better at detecting T/ET for water limited sites following reasons given in that paper, namely that "Because we do not force the regression through the origin, our approach is more appropriate for water-limited sites, where it is often found that the ET ≠ 0 (i.e., the intercept) for GEP = 0 [Biederman et al., 2016].". However, the method does not work well at the less water-limited forested sites - there is only a month or two where there are significant linear fits and where those fits yield positive ET axis intercepts. Indeed, Scott and Biederman had no intention of this method

being universally used but just found that it worked particularly well for their sites (low elevation shrub and grassland). Thus, for the Fuf sites we used the Zhou method.

At the forested sites we only keep the Zhou et al. estimates for the reasons given above and at the lower elevation grass and shrub sites we now give estimates from both Zhou et al. (2016) and Scott and Biederman (2017) to show that indeed there is uncertainty in estimating T/ET ratios based on assumptions in different methods. We detail both of the these methods and our reasoning for having only Zhou at the forested sites and both at the grassland sites in Section 2.3.1 ("Site-level meteorological and eddy covariance data and processing") with the following sentence:

"Estimates of T/ET ratios were derived from Zhou et al. (2016) for the forested sites, and both Zhou et al. (2016) and Scott and Biederman (2017) at the more water-limited low elevation grass- and shrub-dominated sites. Zhou et al. (2016) (hereafter Z16) used eddy covariance tower GPP, ET and vapor pressure deficit (VPD) data to estimate T/ET ratios based on the ratio of the actual or apparent underlying water use efficiency (uWUEa) to the potential uWUE (uWUEp). uWUEa is calculated based on a linear regression between ET and GPP.VPD0.5 at observation timescales for a given site, whereas uWUEp was calculated based on a quantile regression between ET and GPP.VPD0.5 using all the half-hourly data for a given site. Scott and Biederman (2017) (hereafter SB17) developed a new method to estimate average monthly T/ET from eddy covariance data that was more specifically designed for the most water-limited sites. The SB17 method is based on a linear regression between monthly GPP and ET across all site years. One of the main differences between the Z16 and SB17 method is that the regression between GPP and ET is not forced through the origin in SB17 because at water-limited sites it is often the case that ET ≠ 0 when GPP = zero (Biederman et al., 2016). The Z16 method also assumes the uWUEp is when T/ET = 1, which rarely occurs in water-limited environments (Scott and Biederman, 2017)."

Based on the fact we now have also have T/ET estimates for US-Vcp and we also have two T/ET estimates for the grass and shrub dominated sites, we have adapted Figure 6 (and its caption) to include both estimates for the grass- and shrub-dominated sites and included the Zhou et al. (2016) method for the US-Vcp site. We have also altered the description of these results in Section 3.3 as described below.

For the forested sites, we have edited this paragraph: "Further support for the suggestion that modelled E is overestimated comes from examining the T/ET ratios. Although both E and T increase in the US-Fuf 11LAY simulations (compared to the 2LAY – Fig. S3a) – due to the increase in soil moisture (as previously described in Section 3.1 and Figs. 2 and S2a) – the larger increase in 11LAY E compared to T resulted in lower 11LAY T/ET ratios (Fig. S3a). The seasonal trajectory of T/ET ratios at US-Fuf appear to match data-derived estimates following the Zhou et al. (2016) method: the ratio peaks in the Spring before decreasing in July, with monsoon period T/ET values that are on average lower than the spring (Fig. 6). However, the magnitude of T/ET ratios are too low in all seasons given the 100% tree cover at this site with a LAI ~2.4. Whilst low spring 11LAY T/ET ratios may be due to overestimated E as a result of higher soil moisture and underestimated snow cover, the

generally low bias in T/ET ratios may also be due to the fact there is no bare soil evaporation resistance term included in the default 11LAY version."

to include a broader description of issues at the forested sites now we have T/ET estimates for US-Vcp as well as US-Fuf. The edited text now reads:

"Further support for the suggestion that modelled spring E is overestimated comes from comparing the model to estimated T/ET ratios (Fig. 6). Although both E and T increase in the US-Fuf and US-Vcp 11LAY simulations (compared to the 2LAY – Fig. S3a and b) due to the increase in soil moisture (as previously described in Section 3.1 and Figs. 2 and S2a), the stronger increase in 11LAY E compared to T resulted in lower 11LAY T/ET ratios across all seasons (Fig. S3a and b). While the model captures the bimodal seasonality at the forested sites as seen in the Z16 data-derived estimates (Fig. 6), the magnitude of model T/ET ratios appear to be too low in all seasons given the 100% tree cover at these sites with a maximum LAI of ~2.4. Whilst low spring 11LAY T/ET ratios at may be due to overestimated E as a result of higher soil moisture and underestimated snow cover, the generally low bias in T/ET ratios across all seasons at both US-Fuf and US-Vcp may also point to the issue that no bare soil evaporation resistance term is included in the default 11LAY version. This may also explain why the model T/ET ratios do not increase as rapidly as estimated values at the start of the monsoon (Fig. 6). Discrepancies in the timing of T/ET ratio peak and troughs between the model and data-derived estimates at the forested sites could also be due to the fact evergreen PFTs have no associated phenology modules in ORCHIDEE; instead, changes in LAI are just only subject to leaf turnover as a result of leaf longevity, which may be an oversimplification."

One of the main changes to the results following the inclusion of both methods is in the paragraph relating to US-SRM spring T/ET given that the model now lies in between the two estimates for this time period. Therefore, we have replaced this original text: "We can also glean some information on whether T or E (or both) are be responsible for the 11LAY overestimate of springtime ET at US-SRM by comparing modelled T/ET ratios against data-derived estimates. Observed T/ET ratios at the low-elevation sites were derived from independent eddy covariance data following the method of Scott and Biederman (2017) (Fig. 6). The observed spring T/ET at US-SRM is slightly underestimated by the model (Fig. 6). Given that T/ET ratios are underestimated by the model but ET is overestimated by the model, it is probable that spring E at this site is too high. Spring T could also be overestimated at US-SRM due potentially due to an overestimate in LAI (Fig. S5); however, the positive bias in E must be larger than the bias in T. If model LAI at US-SRM is too high during the spring, it is impossible to determine whether the shrub or grass LAI are inaccurate without independent, accurate estimates of seasonal leaf area for each vegetation type; however, in the field the spring C4 grass LAI is typically half that of its monsoon peak (R.L. Scott – pers. comm.) – a pattern not seen in the model (Fig. S6)."

with

“At US-SRM, the modelled spring T/ET ratio overestimates the Z16 estimate and underestimates the SB17 estimate (Fig. 6). The current state of the art is that different methods for estimating T/ET typically compare well in terms of seasonality but differ in absolute magnitude; therefore, the uncertainty in T/ET magnitude during the spring at US-SRM makes it difficult to glean any information on whether T or E (or both) are be responsible for the 11LAY overestimate of springtime ET (Fig. S3c). If the SB17 method is more accurate, then it is probable that modelled spring E at this site is too high. However, if the Z16 estimate is accurate, then it is likely that spring T is overestimated at US-SRM, potentially due to an overestimate in LAI. The model-data bias in spring mean monthly ET is well correlated (0.XX) with spring mean LAI at US-SRM (Fig. S5). If model LAI at US-SRM is too high during the spring, it is impossible to determine whether the shrub or grass LAI are inaccurate without independent, accurate estimates of seasonal leaf area for each vegetation type, which are not available at present; however, in the field the spring C4 grass LAI is typically half that of its monsoon peak – a pattern not seen in the model (Fig. S6). We will test both of these hypotheses (overestimate in either T or E) in Section 3.4.”

We have also edited the following original text: “Data-derived T/ET ratios also help to diagnose why the 11LAY model underestimates monsoon ET at the low-elevation shrub sites (US-SRM and US-Whs– Figs. S3 c-d). Fig. 6 shows that the 11LAY model also underestimates monthly T/ET ratios, and furthermore, that the model does not capture the correct temporal trajectory (Fig. 6). Although the earlier summer drop in T/ET ratios in the 11LAY compared to the 2LAY simulations at grass and shrubland sites (Figs. S3 c-f) does result in a better match in ET between the model and the observations (Fig. 3), the 11LAY T/ET ratios are slightly out of phase. Observed T/ET ratios decline in June during the hottest, driest month, whereas model values decrease one month later in July (Fig. 6). Furthermore, the ratios do not increase as rapidly as observed during the wet monsoon period (July – September).
The underestimate in modelled monsoon T/ET ratios across all grassland and shrubland sites (and likely at US-Fuf and US-Vcp) suggests either that transpiration is too low or bare soil evaporation is too high. At the shrubland sites (US-SRM and US- 500 Whs), both monsoon ET and T/ET are underestimated; therefore, for these sites it is plausible that the dominant cause is a lack of transpiring leaf area. Certainly, monsoon model-data ET biases are better correlated with LAI at shrubland sites compared to grassland sites (Fig. S7). The underestimate in modelled monsoon period leaf area could either be: i) an underestimate of maximum LAI for either grasses or shrubs; or ii) due to the fact the static vegetation fractions prescribed in the model do not allow for an increase in vegetation cover during the wet season (e.g. the lack grass growth in the model in interstitial bare soil 505 areas). In contrast, at the grassland sites (US-SRG and US-Wkg) monsoon ET is well approximated by the 11LAY model; thus, the underestimate in T/ET ratios suggests that both the transpiration is too low and the bare soil evaporation too high.” to include both T/ET methods, to make the text more understandable, and to provide further explanation of the “out of phase” seasonality in T/ET ratios at the low elevation sites. The new text is:

“At the low elevation grass- and shrub-dominated sites, both data-derived estimates of T/ET agree on their seasonality and sign with respect to the model magnitude during the

monsoon. Given this agreement, both sets of estimated values can help to diagnose why the 11LAY model underestimates monsoon peak ET at the low-elevation shrub sites (US-SRM and US-Whs– Figs. S3 c-d). Fig. 6 shows that the 11LAY model also underestimates both Z16 and SB18 monthly monsoon period T/ET estimates across all low elevation sites. The underestimate in modelled monsoon T/ET ratios across all grassland and shrubland sites suggests either that T is too low or E is too high. At the shrubland sites (US-SRM and US-Whs), both monsoon ET and T/ET are underestimated; therefore, for these sites it is plausible that the dominant cause is a lack of transpiring leaf area. As was the case for spring ET at US-SRM, monsoon model-data ET biases are better correlated with LAI at shrubland sites compared to grassland sites (Fig. S7). In contrast, at the grassland sites (US-SRG and US-Wkg) monsoon ET is well approximated by the 11LAY model; thus, the underestimate in T/ET ratios suggests that both the transpiration is too low and the bare soil evaporation too high.
Furthermore, although the 11LAY does capture the decrease in ET during the hot, dry period of May to June (which is a significant improvement compared to the 2LAY – see Section 3.1), the 11LAY T/ET ratios are slightly out of phase with the estimated values. Both data-derived estimates agree that T/ET ratios at all low elevation sites decline in June during the hottest, driest month (as expected); however, the model T/ET ratios reach a minimum one month later in July (Fig. 6). This one month lag in model T/ET ratios is apparent despite the fact that the ET minimum is accurately captured by the model (Figs. 3b and S3). The modelled T/ET ratios also do not increase as rapidly as both estimates during the wet monsoon period (July – September), which can be explained by the fact that the model E at the start of the monsoon increases much more rapidly than modelled T. Taken together, these results suggest that LAI is not increasing rapidly enough after the start of monsoon rains (see Fig. S6), resulting in low biased T/ET ratios in July. Meanwhile the increase in available moisture from monsoon rains is causing a biased high model E that compensates for the lower T. These compensating errors result in accurate ET simulations. The underestimate in modelled leaf area during the monsoon could either be: i) incorrect timing of LAI growth for either grasses or shrubs and an underestimate of peak LAI; and/or ii) due to the fact the static vegetation fractions prescribed in the model do not allow for an increase in vegetation cover during the wet season (e.g. the model lacks the ability to grow grass in interstitial bare soil areas).”

We have also added the following sentence in the abstract:
“However, discrepancies in the timing of the transition from minimum T/ET ratios during the hot, dry May-June period to high values during the summer monsoon period in July-August could point towards incorrect simulations of seasonal leaf phenology. ”

In terms of winter values at US-Fuf (and now US-Vcp), my co-author (Russ Scott) left out months where GPP is very low because both estimation procedures rely on the relationship between ET and GPP, very low and low variability GPP (in the winter) results in a poor relationship between these two quantities. We have added the following sentence explaining this into Section 2.3.1:

“T/ET ratio estimates are omitted in certain winter months when very low GPP and limited variability in GPP results in poor regression relationships.”

Thus, the data-derived estimates are not given for US-Fuf during the winter months when there is a lot of snow so we are not relying on the T/ET estimates for this period. And we agree with the reviewer that the model is likely right on simulating low T/ET values during this period.

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.

We were initially reluctant to add these snow test results because a) we didn't show the results of the other snow-related tests we did (described in the original lines 436-438) and b) because there are already a lot of figures in this paper and the figure for this snow forcing test was deemed to be of lower importance. However, we have now added this test to the supplementary (Figure S5 - please see below). Please note also that we have slightly lengthened and added to the description of these snow-related results in Section 3.2 (and changed Figure 5) following some suggestions from Reviewer 1. We hope that the description and discussion of these particular results is more detailed and nuanced. The edited text is:

"In contrast, the temporal mismatch between the observations and the model in the uppermost layer is higher at the forest sites. The US-Fuf and US-Vcp 11LAY simulations appear to compare reasonably well with observations in the upper 2cm of the soil from June through to the end of November (end of September in the case of US-Vcp) (Fig. 4). However, in some years the model appears to overestimate the VWC at both sites during the winter months (positive model-data bias), and underestimate the observed VWC during the spring months (negative model-data bias), particularly at US-Fuf. Although US-Fuf and US-Vcp are semi-arid sites, their high-elevation means that during winter precipitation falls as snow; therefore, these apparent model biases may be related to: i) the ORCHIDEE snow scheme; ii) incorrect snowfall meteorological forcing; and/or iii) incorrect soil moisture measurements under a snow pack. During the early winter period the model soil moisture increases rapidly as the snowpack melts and is replenished by new snowfall, whereas the observed soil moisture response is often slower (Fig. 5a and b light blue zones). This often coincides with periods when the surface temperature in the model is below 0°C (Fig. 5 bottom panel), suggesting that in reality soil freezing may be negatively biasing the soil moisture measurements. An alternative explanation is that ORCHIDEE overestimates snow cover (and therefore snow melt and soil moisture) at the forest sites because it is assumed that snow is evenly distributed across the grid cell, whereas in reality the snow mass/depth is lower under the forest canopy than in the clearings.
At US-Fuf, it appears that the model melts snow quite rapidly after the main period of snowfall (Fig. 5a light green zones). Once all the snow has melted, the model soil moisture also declines; however, the observed soil moisture often remains high throughout the spring – causing a negative model-data bias (Fig. 5a). Unlike US-Fuf, a similar negative model-data

bias at US-Vcp often coincides with periods when snow is still falling, although the amount is typically lower (Fig. 5b light green zones); however, the model does not always simulate a high snow mass during these periods. These periods coincide with rising surface temperature above 0°C. Although snow cover, mass, or depth data have not been collected at these sites, snow typically remains on the ground until late spring after winters with heavy snowfall, suggesting that the continued existence of a snow pack and slower snow melt that replenishes soil moisture until late spring when all the snow melts. Therefore, the lack of a simulated snow pack into late spring could explain the negative model-data soil moisture bias. To test the hypothesis that the model melts or sublimates snow too rapidly, thereby limiting the duration of the snowpack and also allowing surface temperatures to rise, we altered the model to artificially increase snow albedo and decrease the amount of sublimation; however, these tests had little impact on the rate of snow melt or the duration of snow cover (results not shown). Aside from model structural or parametric error, it is possible that there is an error in the meteorological forcing data. Rain gauges may underestimate the actual snowfall amount during the periods when it is snowing (Rasmussen et al., 2012; Chubb et al., 2015). If the snowfall is actually higher than is measured, it may in reality lead to a longer lasting snowpack than is estimated by the model. To test this hypothesis, we artificially increased the meteorological forcing snowfall amount by ten times and re-ran the simulations. Although this artificial increase is likely exaggerated, the result was an improvement in the modelled springtime soil moisture estimates at US-Fuf (Fig. S5). However, the same test increased positive model-data bias in the early winter increased at US-Fuf, and degraded the model simulations at US-Vcp. This preliminary test suggests that inaccurate snowfall forcing estimates may play a role in causing any negative model-data bias spring soil VWC but more investigation is needed to accurately diagnose the cause of the springtime negative model-data bias."

To better match this text we have updated Figure 5 to only include the pertinent variables (and have added surface temperature) and we have added an extra supplementary figure (S5) to show the results of the increased snow forcing (as per a comment from Reviewer 2):

Figure 5: a) US-Fuf and b) US-Vcp 11LAY (blue curve) daily time series (2007-2010) of model versus re-scaled (via linear CDF matching) observed volumetric soil water content (middle panel SWC – m3m-3) (black curve), compared to simulated snow mass (top panel) and surface temperature (bottom panel). Snowfall is also shown as grey lines in the SWC time series. In the bottom panel the grey horizontal dashed line shows 0°C threshold.

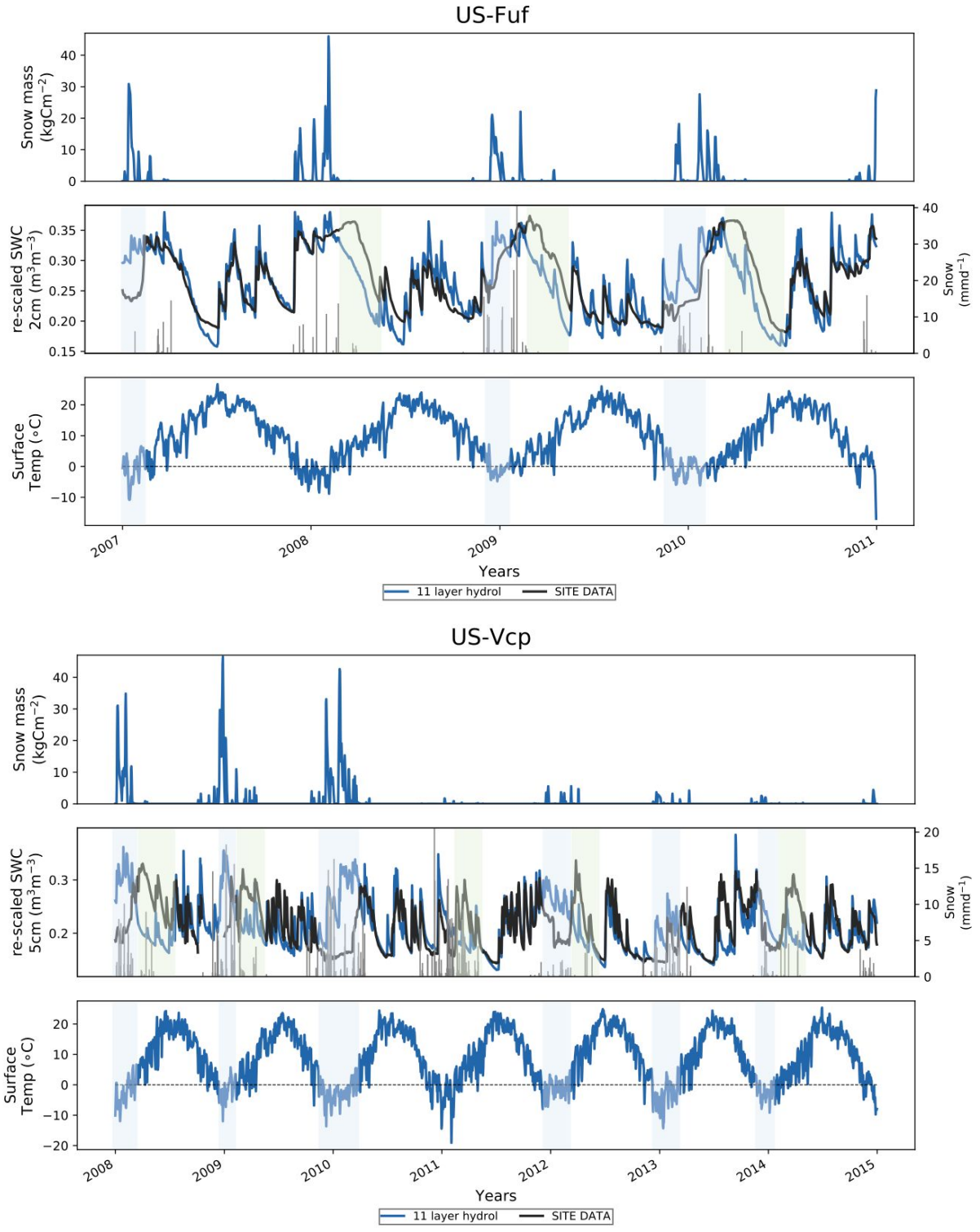
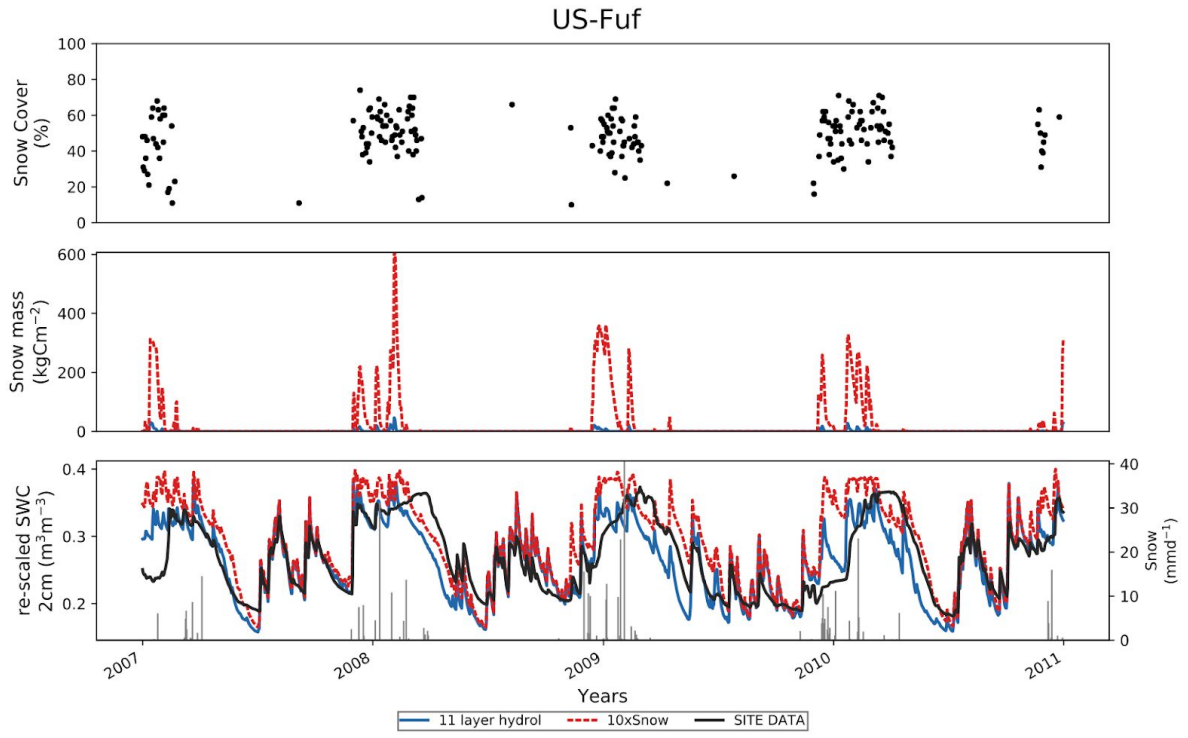


Figure S5: Linear regressions between spring (March-April) mean monthly LAI (m2m-2) and spring mean monthly ET (mmmonth-1) model-data misfits for each site. The dominant PFT is given in brackets for each site. See Table 1 for PFT acronyms.



We have also modified this sentence in the discussion:
"More specifically, more information on snow cover, depth or mass, particularly under closed forest canopies, would be useful to ~~test if the precipitation data measured by the meteorological stations accurately captures the right amount of snowfall~~ diagnose potential sources of bias in the snowfall simulations. "

We have also added this sentence into the abstract:
"Biases in winter and spring soil moisture at the forest sites could be explained by inaccurate soil moisture data during periods of soil freezing and underestimated snow forcing data."

Finally, we also updated a sentence in the conclusions to reflect both the negative and positive model-data biases in soil moisture at the forested sites could be related to snowfall issues:

"Remaining discrepancies in both overestimated and underestimated winter and spring soil moisture at high-elevation semi-arid forested sites might be respectively related to issues with soil moisture data during periods of soil freezing and underestimated snowfall forcing data causing a limited duration snowpack, with consequent implications for predictions of water availability in regions that rely on springtime snowmelt."

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.

We did mistakenly say that the overestimation of spring ET was for Fuf *and* Vcp - we have corrected that now to only refer to Fuf. But we agree with the reviewer that it is an incomplete explanation (and doesn't help to explain Vcp). We did try to emphasize this in the original text by moving on later in the paragraph to use T/ET ratios to try to explain all ET issues at both sites. At both these sites the T/ET ratios are lower than the estimated values (we see this now we have Vcp included in these estimates); thus, we go to on say that this could be due to a lack of T or a possible overestimation of E at both sites due to the lack of the bare soil resistance term and/or issues with LAI and the phenology. We test the former further hypothesis in Section 3.4. So, while it was our intention in Section 3.2 to say the underestimate in spring *soil moisture* at Fuf and Vcp was due to incorrect snowfall (and we have updated that text - see above), the link between an underestimated snowpack and overestimated spring ET at Fuf in Section 3.3 is just one of the hypotheses we put forward for the errors in spring ET. It was not our intention to say that snowfall *is* definitively the factor that contributes to overestimated spring ET at Fuf - but more that it is one possible explanation (and we believe it does not read that way given we say "The lack of a persistent snowpack in the model during this period could explain the positive bias in spring ET because in reality the presence of snow would suppress bare soil evaporation"). We did not give this as a definitive cause of the ET in the abstract and conclusions. Many interacting factors likely go into why spring ET is overestimated at Fuf (and indeed, not at Vcp), which we try to emphasize. Unfortunately it is difficult to test all of these hypotheses - we have tested one in Section 3.4. We have added this sentence in the conclusions (after the sentence about the possible role of snowfall issues in soil moisture model-data biases detailed in the answer to the previous comment) to clarify that there are multiple possible reasons why there are ET discrepancies at the forest sites, not just reasons related to snowfall:

"However, biases in soil moisture at both the forested sites do not translate into the same biases in modelled ET at the forest sites, suggesting other factors such as issues in evergreen phenology/LAI simulations or the lack of resistance to bare soil evaporation may also play a role."

Do the two model set-ups use the same snow module and are the parameterizations the same for the different sites?

Yes, they are. But even though the snow model is the same in the 2 configurations, the different hydrology simulations at the two sites then impacts the soil thermal processes differently because the soil properties (heat capacity and thermal conductivity) depend on soil water content. Also we don't expect complete snow coverage all the time at each site (we can see in Fig. 5 snow comes and goes throughout the winter period), so the overall surface temperature may be different, leading to different snow melt, snowpack etc at the two 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.

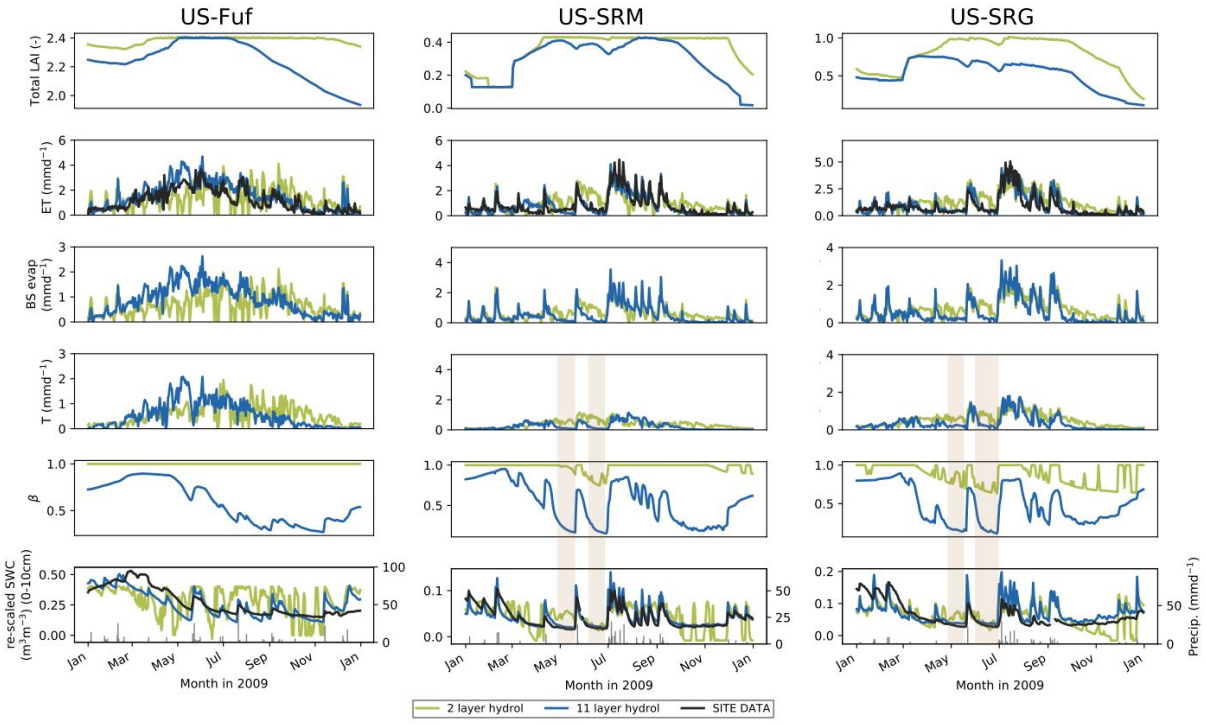
We would love to have data that could help us test our hypotheses the model is underestimating snow pack. Indeed we said this in original lines 470-471 (directly after the sentence quoted above): “To accurately diagnose this issue, we would need further information on snow mass or depth”. However, as we mentioned, we considered that we would need information on snow mass or depth (to validate the top panel in Figs. 5 a and b), not snow cover (given these are site simulations and we’re not examining spatial heterogeneity), which is what the MOD10A product is. There are satellite products of snow water equivalent that might be more useful in validating snow mass/depth but as far as we understand these products are only available at very coarse spatial resolutions (>=25km). However, after considering the reviewer’s suggestion we agreed that the MOD10A snow cover at 500m, while not helping us with snow amount, would help with evaluating snow duration and indeed it did to some extent. Therefore, we have included it in the new Fig S5 which also shows the result of the increased snow forcing test we did, which we have now included as per the reviewer’s justified request.

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.

We appreciate the reviewer’s comments. Indeed, we debated whether or not to add observations to Figure 2 when comparing the 2 vs 11 layer upper soil moisture, and in an earlier version we did have such a comparison. For the 2LAY version we only have the option to compare either the upper layer moisture (0-10cm) or the total column of 2m (or bottom 1.9m). In figure 2 we really wanted to look at the upper layer moisture given this is predominantly a plot about what is happening for ET (and its component fluxes and relevant

processes) and the upper layer moisture comes from. However, the issue with the 2LAY upper layer moisture is that that layer can disappear entirely (as we describe in Section 2.2.2), which is why it has this very noisy temporal profile that can decrease to zero. Even by looking at the temporal dynamics of the 2 vs 11 layer upper layer compared to the ET we can see that the ET temporal dynamics are mostly related to this upper layer moisture. We also thought of having a further soil moisture plot that shows the much smoother total column soil moisture temporal dynamics (to highlight again that the ET temporal dynamics are dominated by the upper layer, and not the total column, but that would have added a 7th panel to Figure 2, which we felt was too much (but we can add that in if the reviewer thinks that would help explain this point.

With all this in mind, we thought a) that comparing the 2 layer upper moisture to the observations is tricky because of this issue that the layer can disappear entirely, and b) that it takes away from the main point, which is that the temporal dynamics of the 2lay reflect this issue that the layer can disappear entirely and are therefore almost by design not realistic when comparing to observations. Furthermore, because of the reasons we highlight in Section 2.3.2, we do not wish to compare absolute soil moisture values in any given layer directly to observations (and we are happy the reviewer understands these points), therefore we need to do the linear CDF matching. We can absolutely do this for the 2 layer upper layer soil moisture (as well as the 11 layer equivalent) and compare to the observations that most represent this 0-10cm interval (see the adapted figure below), but we want to stress that this is not as direct a comparison as we make for the 11-layer comparison in Figure 4 and is somewhat more subject to the issues we describe above (essentially, less of an “apples to apples” comparison than we have for Figure 4). However, we have done the CDF matching and adapted the figure, and we agree with the reviewer that it certainly helps to make our case that the 11 layer does indeed better capture the observed temporal characteristics (the fluctuations are much more realistic). We hope they see our point more clearly now and we propose keeping this revised figure 2 and will make any necessary changes in the text.



However, we choose not to add observations into Figure 1 because this is the only figure we have where we look at the overall changes in the *absolute values* of total soil moisture (as opposed to re-scaling the model to match the observations using linear CDF matching (in original equation 5 and as explained in Section 2.3.2). As we also explain in Section 2.3.2 the observations come from different depths at each site and it is hard to know over which depth the different soil moisture probes measure. Furthermore, we do not have observations below 75cm (and much shallower at some sites - Table 2). Therefore, we do not have estimates of how much water content there is in the total soil column and thus we cannot put the observed total column moisture in Figure 1. Even if we were to convert the total column soil moisture to volumetric soil moisture, we still do not know a column average volumetric soil moisture content. We think it would be heavily biased if we were to simply average over the limited depths we have. We have added the following sentence into Section 2.3.2 for further clarification of this point:

"Given the maximum depth of the soil moisture measurements is 75cm (and is much shallower at some sites) we cannot use these measurements to estimate a total 2m soil column volumetric soil moisture content."

If we do add the re-scaled soil moisture observations into the upper layer soil moisture comparison plot in Fig. 2 (bottom panel - see above), we will also modify the following sentence of this section:

"Instead, we only used these measurements to evaluate the 11LAY model and 2LAY upper layer soil moisture (calculated for 0-10cm) because, unlike the 2LAY model, with the 11LAY version of the model we have model estimates of soil moisture at discrete soil depths."

We have also added this sentence to the caption of Figure 1:

"For soil moisture, the absolute values of total water content for the upper layer and total 2m column are shown for both model versions, i.e. the simulations have not been re-scaled to match the temporal dynamics of the observations (as described in Section 2.3.2); therefore, soil moisture observations are not shown. Observations are only shown for ET."

Finally, we agree with the reviewer about the claim that surface runoff is more realistic, given we do not actually show any data in the manuscript. The fact that claim appears to be overstated is perhaps more due to our lack of properly articulating what we meant here, and the lack of referencing other studies when discussing the runoff and drainage results (although data is still limited). We did refer to two studies from US-Fuf and US-SRM that discuss low drainage results and the fact that Precipitation is mostly accounted for by ET.

In the revisions (and in response to another comment by Reviewer #1 about whether limited drainage at the forested sites was plausible), we have also added this sentence: "In general, all these semi-arid sites have very little precipitation that is not accounted for by ET at the annual scale (Biederman et al., 2017 Table S1)."

Table S1 in Biederman also shows that most precipitation is accounted for by ET across all these sites; therefore, although we don't explicitly have runoff and drainage data we feel these data do serve to highlight that the original 2LAY estimates of total runoff were *likely* too high and that the 11LAY values appear to be more plausible.

Given these points, we could modify this sentence in the conclusions in the following way:

"Associated changes in the calculations of runoff, soil moisture infiltration, and bottom layer drainage also appear to result in more plausible (lower) estimates of total runoff (surface runoff plus drainage) at the forest sites given that across all these semi-arid sites, most precipitation is accounted for by ET at the annual scale."

However, if the reviewer feels this is still too exaggerated a claim for the conclusions we will remove the sentence entirely.

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.

We are glad the reviewer finds the manuscript interesting and appreciate their thoughtful comments. We hope that by addressing their comments (above and below) we have helped to clarify our objectives, to better align the results with those objectives, and to provide conclusions that better support the results. In particular, we hope that the modifications we propose to figure 2 help to support one of our main conclusions that the 11LAY does a better

job in terms of capturing the ET temporal dynamics, and that it is not simply that we prefer the 11LAY version. We also hope that the discussion we provided serves to highlight that we realize there are many remaining caveats (model issues, missing processes) in how we currently model hydrology using the mechanistic 11LAY model, but that dealing with all of these issues is beyond the scope of the current paper.

Minor comments

P1.L36. Results better → results in a better?

Changed - thank you.

P2.L62. A evaporation → an evaporation

Changed - thank you.

P3.L79 have been rarely been → have rarely been

Changed - thank you.

P4.L115. Define PFT

Done - thank you.

P6. L187. What do you mean with soil tile? The spatial distribution of different soils within a grid cell?

No it corresponds to the number of water columns for which each separate water flux is calculated. We have modified this sentence to read: "Independent water budgets are calculated for each "soil tile", which define separate water columns within a grid cell.". We hope that with this modification and the original following sentences of "In the 2-layer scheme, soil tiles correspond to PFTs; therefore, a separate water budget is calculated for each PFT within the grid cell. In the 11-layer scheme there are three soil tiles: one gathering all tree PFTs, one gathering grasses and crops, and the third as bare soil." that the meaning of soil tile is now clearer.

P6.L189. "all three PFT's" –> It is mentioned before that there are 12, so why three now?

This actually reads "all **tree** PFTs"

P6.L191. Related parameters) → remove ")"

Changed - thank you.

P7.L210. At al → et al

Changed - thank you.

P7.L217. At al → et al

Changed - thank you.

P8.L227. Seems a bit arbitrary to me, why these numbers?

These are very classical values, often given as -33kPa and -1500kPa, or -0.33 and -15 bars, see for instance Rawls et al. (1982) and Verhoef and Gregorio (2014)

Rawls, W. J., Brakensiek, D. L., & Saxtonn, K. E. (1982). Estimation of soil water properties. Transactions of the ASAE, 25(5), 1316-1320. Cited 1894 times according to Google Scholar.

Verhoef, A., and Gregorio, E. (2014). Modeling plant transpiration under limited soil water: Comparison of different plant and soil hydraulic parameterizations and preliminary implications for their use in land surface models, Agricultural and Forest Meteorology, 191, 22-32, https://doi.org/10.1016/j.agrformet.2014.02.009.

Bonan (2002) gives the same potential for wilting point, but -1m for field capacity (which is very close to -3.3m given the wide range of soil water potential in an unsaturated soil: from 0 to much less than -150m).

Bonan, G. (2015). Ecological climatology: concepts and applications. Cambridge University Press. Cited 1285 times in Google Scholar.

We have added a reference to Ducharne et al. (in prep.) here that explains this point (extensive description of the latest version of the ORCHIDEE soil hydrology). We can add references to the above if needed.

P8.L229. Has been test → have been tested
Changed - thank you.

P8.L256. The the root density → the root density
Changed - thank you.

P8.L256-257. Why these values? What are they based on? Eq3. Please define and describe also h_t and d

h_d^t is one variable that is “dry soil height of the topmost soil layer” (originally defined in lines 259-260).

These values were selected to get a higher root density for forests than for low vegetation and have not been calibrated against field data. We have added/modified lines in the discussion on the need for calibration of these parameters, e.g.:

In the discussion section on Issues with modelling vegetation dynamics in semi-arid ecosystems:

“Alternatively, it may be that other model parameters and processes involved in leaf growth – for example phenology, root zone plant water uptake, and photosynthesis-related parameters – are inaccurate and in need of statistical calibration (e.g. MacBean et al., 2015).
”

And in the discussion section on bare soil evaporation: “
“It is possible that the bare soil resistance is only part of the solution, and that the simulation of ET and its component fluxes could be fixed with both a more realistic representation of semi-arid phenology or vegetation fractional cover at both grass and shrub dominated sites (as discussed above) and/or a statistical calibration of relevant vegetation, root density, soil hydraulic parameters (e.g. Shi et al., 2015).”

However, there are limited root density decay factor parameters, so these parameters would have to be calibrated by means of indirect observations (such as ET). Although it is beyond the scope of the present study, we do plan to conduct future studies. Calibration of such parameters has not yet been attempted, and based on the data assimilation experience of NM, investigating how best to optimize new processes/parameters will take time. Thus, these studies will be presented in future papers.

P8.L267. Is T here transpiration? Please define.
We do define that in Section 2.2.1. We can define it again here.

Eq5. Please define your variables
Done - thank you.

P12.L351. Higher compared to the other sites? It is not higher than the 11LAY-scheme.
We thank the reviewer for spotting the error in this sentence. It now reads: “In the 2LAY simulations, the upper layer soil moisture is similar across all sites; whereas, in the 11LAY simulations the difference between the high elevation forest sites and low elevation grass and shrub sites has increased.”

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.
The sentence actually refers to the upper layer (top 10cm) not the upper 2m. The whole soil column is 2m deep. The 2LAY upper layer (top 10cm) does decrease to 0 and this can be seen in Figure S2 (and Fig. 2, which we refer to in this sentence (not Fig. S1): “Whereas the 2LAY
upper layer soil moisture simulations at all sites fluctuate considerably between field capacity and zero throughout the year – including during dry periods with no rain – the temporal dynamics of the 11LAY upper layer moisture simulations correspond more directly to the timing of rainfall events (see Fig. 2 bottom panel for an example at 3 sites in 2009 and Fig. S2 for the complete time series for each site).”

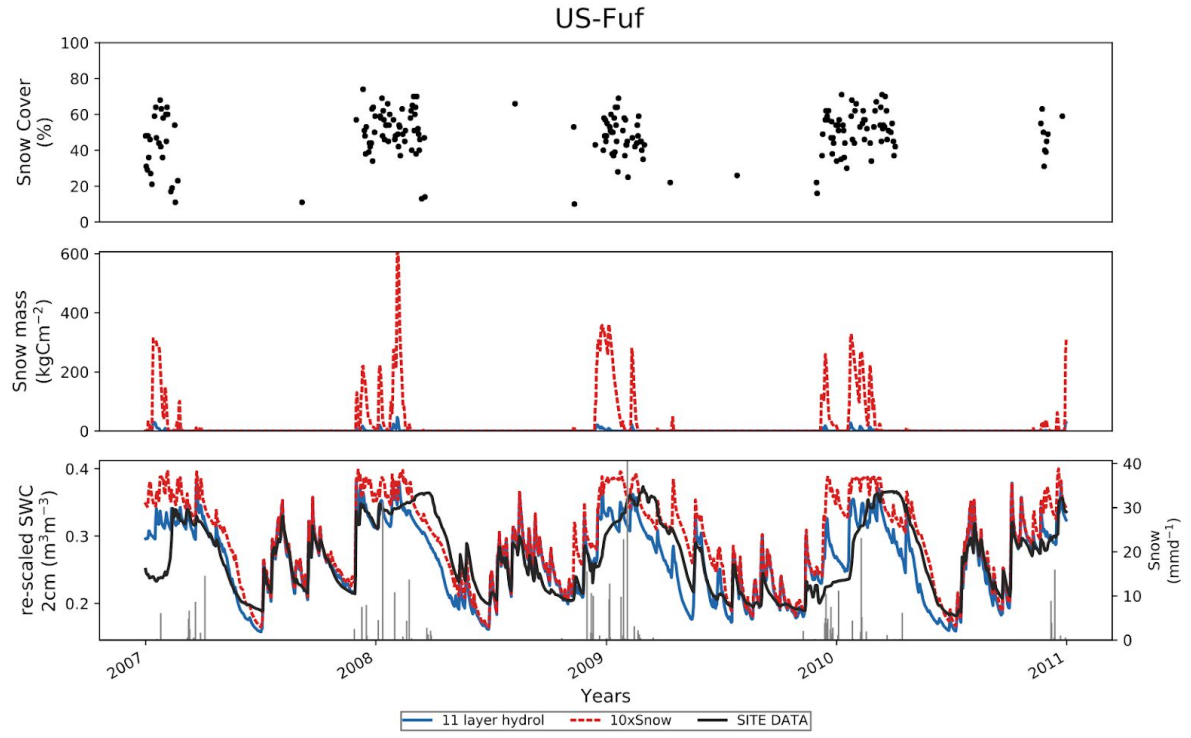
P12.L383-384. I do not think you can conclude 11LAY is better based on the data as shown, there are no observations shown of soil moisture in Fig. 2.

We have added observations into Fig. 2 (presented above) and we have also modified all the Fig. S2 plots. Please see the above discussion (in the reviewer’s main comments on this point).

P14.L421. Fig 4 →Fig. 4
Changed - thank you.

P14.L422. So which sites in fig4 do you mean? It's easier to add the names, then the reader knows where to look.
We have added in "(US-SRM, US-SRG, US-Whs, US-Wkg)" at the end of the sentence.

P14.L445-448. Where can I see this? Please make sure you back up your conclusions by showing the evidence.
The reviewer is right - we meant to add "(data not shown)" because the paper is already very dense and this is a relatively small test by comparison. But we agree that this should be shown and so we have added a figure to the supplementary material (new Fig S5). We also included the MOD10A snow cover product results for this site as per a suggestion from the reviewer (see above). Please note that we have also added further detail and nuance to the description of the forest site results related to snow as per a comment from Reviewer 1 - these can be found in the response to Reviewer 1.



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.

We apologize, when we described what we include in ET in the original lines 182 to 184, we put the "E" in the wrong place (incorrectly placing it next to "evaporation from water intercepted by the canopy" instead of "bare soil evaporation". Those lines have been modified and now read: "Evapotranspiration, ET, in the model is calculated as the sum of

four components: 1) evaporation from bare soil, E; 2) evaporation from water intercepted by the canopy; 3) transpiration, T, (controlled by stomatal conductance); and 4) snow sublimation (Guimberteau et al., 2012b)."
We hope this is now clear, because when talking about plants we necessarily need to talk about plant transpiration, and therefore ET is not to be confused with E, which just refers to bare soil evaporation (at least, this is how we refer to it in LSMs that fully couple hydrology and biogeochemistry).
We have also made changes to the abstract to be clear as to what E and T refer to.

P15.L467. You mention before that US-Vcp underestimated ET, instead of overestimated.
Thank you for spotting this mistake. We had already spotted the incorrect inclusion of US-Vcp here and have removed it.

P16.L480. Are be responsible –> are responsible?
Changed - thank you.

P17.L517. You do not show that T/ET fractions are better with the reduced bare soil fraction.

The reviewer is right that we haven't shown these in Fig. 7 and that this statement is too broad and imprecise. We have now added the T/ET data-derived estimates into Fig. 7 and updated the caption. Given we now also have a more nuanced description of the use of the T/ET estimates in evaluating the model (described above), we have also further modified this sentence and included additional sentences to i) emphasize that we are talking about mean changes across all the sites; ii) to highlight differences between the spring and summer months; and most importantly, iii) to give further weight to the suggestion that the main issue might be more related to the model's ability to capture the seasonal changes in leaf area/vegetation cover (as opposed to just the amount of vegetation that is present throughout the year):

"However, although the T/ET ratios reduced the negative model biases compared to the data-derived estimates in the summer monsoon period, the model now overestimates ET (Figs. 7 and S8). However, while the decrease of the bare soil fraction (increase in C4 grasses) may have partially accounted for the negative bias in T/ET ratios at the start of the monsoon, the changes did not correct the phase discrepancy between the estimated and modelled T/ET seasonal trajectories: the estimated T/ET still declines to a minimum in June (as expected during the hot, dry period), whereas the model declines one month later. Furthermore, the spring ET model-data bias is further exacerbated by the increase in bare soil fraction and the mean spring estimated T/ET ratios and ET are a closer match to the original 11LAY version (Figs. 7 and S8). Putting the latter two points together, this new analysis gives further weight to the suggestion put forward in Section 3.3. that the model is not capturing the correct increase in leaf area at the start of the monsoon – not just that there is a lack in the overall amount of transpiring leaf area. Thus, there is potentially more of a problem with the model phenology schemes and/or the model's ability to capture dynamic changes in seasonal vegetation cover than there is with the prescribed fractional vegetation cover. We discuss these issues more in Section 4."

P17.L523. TeNE-forest?
We have removed the forest.

P17.L529. Spring → spring
Changed - thank you.

P19.L592. ORCHIEE → ORCHIDEE
Changed - thank you.

P21.L669. I am not sure how you can conclude this without runoff data and never evaluating it.

We have proposed a change to this sentence - please see the reviewer's general comments above.

Table3. Please note that RMSE also has a unit
Changed - thank you.

Figure 3. The unit is mmm-1, I believe you mean mm/month, but please make this clearer.
We have added "Units are mm per month (mmm^{-1})" to the caption and all other figures captions that have the same issue.

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?
We have changed this plot according to the discussion above in the reviewer's main comments, including adding the two methods and the addition of Vcp. We have also replied above about why there are no data for the first (winter) months. The 2LAY estimates are not included because this plot is referred to in the section in which we are describing remaining discrepancies between the 11LAY and the observations (and *not* the differences between the 2 and 11LAY); therefore, we only plot the 11LAY here for clarity in describing the results and to not have too many curves to distinguish between. The comparison between the 2LAY and 11LAY T/ET estimates are shown in Fig. S3.

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.
This was a collective decision on the part of the co-authors to show a summary here and then show each individual site in the supplementary (Fig. S8) to avoid an excessive number of figures in the main manuscript. We would like to stick with this decision.

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.
We assume the reviewer is talking about Figures S5 and S7, not S6? We have added units - thank you to the reviewer for pointing these out. However, in the original figures the data points were there - not just the regression lines, so we're not sure why these weren't

showing in the documents the reviewer received. We are not sure what the reviewer means by a legend as only one dataset is shown? We already have titles for each of the subplots (which correspond to each site).

Data availability:

Where are the model results shared?

The model results will be shared on my GitHub Page: https://github.com/nmacbean. We will wait to set up a specific repository for the article and to upload the simulations to the repository with a readme file (and eventually, the published paper) until the revisions have been accepted (in case we need to do more work on the paper).