Response to Reviewer #1

MacBean and colleagues compare the land surface model ORCHIDEE against six semi-arid flux sites, using the old 2-layer soil hydrology scheme and the new 11-layer scheme of ORCHIDEE.
The study is certainly done correctly and the comparisons are fine. Specific remarks and questions are below.

We thank anonymous reviewer 1 for providing us with such a thoughtful and useful review. We provide more detailed comments to all of their comments and suggestions 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.

However, one asks him/herself why one needs another validation of a Richards model in an LSM, showing that it performs better than on old bucket or 2-bucket version? Specifically the multi-layer soil model of ORCHIDEE was tested quite a number of times already.

We agree to a certain extent with the reviewer's comment and we initially addressed this in our interactive informal response to this review: https://editor.copernicus.org/index.php/hess-2019-598-SC1.pdf?_mdl=msover_md&_jrl=13&_lcm=oc108lcm109w&_acm=get_comm_file&_ms=81557&c=175959&salt=225704252184511132. We respond with some updated comments here.

It is true that most land surface models do now have a more mechanistic Richards' equation-type approach to modeling soil moisture dynamics. It's also the case that it is hard to compare the 2-layer and 11-layer approach given how different the representation of soil hydrology is and that it is very difficult to compare the 2-layer version to observations (much harder than the 11-layer).

However, despite these considerations we decided to keep the 2-layer vs 11-layer comparison in the first part of the results for this paper following reasons: firstly, we are expecting that not all readers are land surface modelers and that some of those people might either not be familiar with simple bucket models, or they might be users of hydrological or other types of models that still use a simple bucket scheme. For these readers, we wanted to show for a range of semi-arid sites that the bucket model really does not represent the temporal dynamics of the soil moisture or ET well; therefore, they should likely not trust ET predictions in semi-arid from any model that uses these types of soil hydrology schemes.

Secondly, the ORCHIDEE model CMIP5/IPCC AR5 simulations were based on the 2-layer version of the hydrology model. While this was a long time ago now and the CMIP6 simulations are being released, many people are still using CMIP5 to study various aspects of earth system processes, climate change impacts, or to understand model deficiencies. Given the fact that CMIP6 results are ~1 year delayed, we expect that people will continue to use CMIP5 simulations for at least another year. Therefore, we explicitly wanted to mention

that the ORCHIDEE CMIP5 ET predictions might not be as accurate as previously thought for semi-arid regions, with consequences for predictions of other variables.

Finally, we asked anonymous reviewer #2 what they thought about the 2 vs 11 layer comparison and, given the comments of this review, whether they would also be inclined to suggest keeping or discarding the comparison. See our initial interactive response to reviewer #2 here: https://editor.copernicus.org/index.php/hess-2019-598-SC2.pdf?_mdl=msover_md&_jrl=13&_lcm=oc108lcm109w&_acm=get_comm_file&_ms=81557&c=175961&salt=1073010281052178988. Reviewer #2 replied that they disagree with removing the 2 vs 11 layer comparison. Their reasoning can be read here: https://editor.copernicus.org/index.php/hess-2019-598-RC3-print.pdf?_mdl=msover_md&_jrl=13&_lcm=oc108lcm109w&_acm=get_comm_print_file&_ms=81557&c=176142&salt=1113425543317000663.

Bearing all these points in mind, we choose to keep the comparison between the 2 vs 11 layer, but in our revised manuscript we propose outlining our reasoning for this comparison more clearly by including the following statement in the introduction (after original lines 120-122):

"Although there have been many previous studies comparing simple bucket schemes versus mechanistic multi-layer hydrology based on the Richards equation, we include such a comparison in the first part of our analysis for the following reasons: a) the simple bucket schemes were the default hydrology in some CMIP5 model simulations and these simulations are still being widely used to understand ecosystem responses to changes in climate; b) variations on the simple bucket schemes are still implemented by design in various types of hydrological models (Bierkens et al., 2015); c) there has not yet been extensive comparisons of these two types of hydrology model for semi-arid regions, and especially not for the SW US; and d) so that the 2LAY can serve as a benchmark for the 11LAY scheme."

Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, Water Resources Research, 51(7), 4923–4947, doi:10.1002/2015wr017173, 2015.

We hope this satisfies both reviewers.

But semi-arid ecosystems are interesting because quite a few model assumptions of LSMs get challenged there. Unfortunately the paper does not talk about it nor tries to advance in this direction.

We absolutely agree with the reviewer that many interesting aspects related to hydrology of heterogeneous semi-arid ecosystems were not either a) detailed in the model description and/or b) not elaborated on in the discussion. We have address this issue in detail for each of reviewer #1's comments below.

For example, ORCHIDEE uses tiles or fractions to deal with different land cover within one grid cell. To my knowledge, if a grid cell is vegetated then there is only transpiration (T). Evaporation (E) is from a special bare soil fraction only. There is no below-canopy E, which experiences lower wind speed, higher humidity and a litter layer compared to bare soil. This might have changed in the 11-layer version. Would be interesting to know. If the bare soil fraction mimics below-canopy E, then it is just a modelling concept and should be treated like this.

Reviewer #1 is right that if the grid cell is vegetated then there is only transpiration - but t*his is only the case* for the 2 layer scheme and not for the 11 layer. In the 11-layer scheme, soil evaporation *is allowed* from each PFT, proportionate to the effective bare fraction, which decreases when LAI increases. The effective vegetated fraction is calculated as an exponential function of LAI, and the effective bare fraction is the complement. The same roughness is used in both the effective bare and vegetated fractions, so reviewer 1 is right that in ORCHIDEE the soil evaporation does not depend on below-canopy conditions (i.e. there is no below canopy E).

In the initial manuscript we did mention the first point (that the bare soil fraction increases as LAI decreases) but we only made this point in the discussion (original lines 572 to 575 in section “Issues with modelling vegetation dynamics in semi-arid ecosystems”). However, it was not described as explicitly as we do here and we did not describe it in the model description. Therefore, in the revised manuscript we include the following lines at the end of Section 2.2.1 (the general model description) after we talk about the vegetation soil tiles in the model (original line 190):

“In the 11-layer scheme, both T and E occur in the vegetated soil tiles. T occurs over the effective vegetated fraction, which increases as LAI increases, whereas E occurs at low LAI over the effective bare soil fraction. The effective vegetated fraction is calculated following a modified Beer-Lambert equation describing attenuation of light penetration through a canopy f_v^j= f^j (1-e^((-k_ext 〖LAI〗_j))), where f^j is the fraction of the grid cell covered by PFT j (i.e. the unattenuated case), f_v^j is the fraction of the effective fraction of the grid cell covered by PFT j and kext is the extinction coefficient and is set to 1.0. The effective bare soil fraction f_b^j is the complement to f_v^j.”

We further add at the end of Section 2.2.3 (Bare soil evaporation and additional resistance term) that there is no belowground E in ORCHIDEE:

“Note that there is no representation of below canopy E in ORCHIDEE and the same roughness is used for both the effective bare ground and vegetated fractions.”

We also add a reference to the relevant model description sections when we discuss this issue in the first section of the discussion (“Issues with modelling vegetation dynamics in semi-arid ecosystems”):

“The connection between vegetation fractional cover and LAI is also a particular issue in sparsely vegetated regions when low LAI effectively means more bare soil is coupled with

the atmosphere *and E increases*. To account for this in ORCHIDEE, the bare soil fraction is slightly increased when LAI is low following a Beer-Lambert law approximation *(see section 2.2.1)*, which is often the case at these sites; however, there are only limited observations to support this model specification."

We also address the issue of below canopy E in the discussion section "ET partitioning (T/ET ratio)" by adding the following after the original final sentence in that section (which was "Nevertheless, in spatially heterogeneous mixed shrub-grass ecosystems it seems likely that missing model processes will need to be accounted for before accurate simulations of T/ET ratios are achieved.")

"One example of this might be the need to include in the model a representation of shrub understory and below canopy E."

Semi-arid ecosystems are probably the only ecosystems where this model structure is valid for soil evaporation. However, the rest of the model structure with fractions comes to its limits. If there is a shrub-encroached grassland, the shrubs (trees in this study) get all crammed into a small tile, shading each other and competing for soil moisture. Or is there a gap fraction in ORCHIDEE? Does it allow for shrub (tree) roots to forage in the grass tile? The grass in semi-arid ecosystems dies off during the year. This changes the LAI as discussed in the paper. But does the grass fraction stay constant? Should LAI rather stay constant in the grass tile but the tile should shrink, leading to more bare soil fraction? I think that one cannot discuss semi-arid ecosystems without talking about vegetation (dynamics). The CO2 fluxes could be interesting in this respect as well. They are omitted in the current paper.

The reviewer is absolutely right that the complexity of semi-arid vegetation dynamics are not well represented in this version of the model - resulting in weaknesses beyond the implementation of the hydrological scheme. No there is no gap fraction in this version of the model and no cross-foraging of tree roots in the grass tile etc. The fraction of vegetation stays constant in the model. All these points are severe limitations and changing these aspects of the model would indeed affect the hydrology. Unfortunately it would not be trivial to change these vegetation dynamics in the model and therefore we have not attempted to do so here. We did investigate the impact of reducing the bare soil fraction. This simple test was in place of having a more dynamic grass vs bare soil cover that changes over the course of the year (which is trickier to implement in ORCHIDEE although we are looking into it). In other words, this lower bare soil fraction test represents the other bookend of two possible ratios of grass to bare soil fraction. The reviewer is also right that this will affect CO2 fluxes. As mentioned in our initial information response to reviewer 1 we are investigating model representation of CO2 fluxes in a separate study. The issues related to CO2 fluxes are greater than can be fixed by changing the soil hydrology and therefore we have separated out these analyses into a separate, forthcoming paper. For this future paper we are also investigating the best way to implement more dynamic seasonal changes in grass cover but it is an ongoing study that is outside the scope of this current study. However, we have added the following sentence into Section 2.4 describing the simulations set-up so as to explain the reasoning for the reduced bare soil fraction test:

"Tests 3 and 5 (reduced bare soil fraction) are designed to account for the fact that grass cover is highly dynamic at intra-annual timescales at the low-elevation sites and therefore during certain seasons (e.g. the monsoon) the grass cover will likely be higher than is represented in the model."

Furthermore, while we did discuss all these issues of vegetation dynamics in the original manuscript discussion (section entitled "Issues with modelling vegetation dynamics in semi-arid ecosystems"), we appreciate that we could have been clearer about these particular issues. Therefore, we have changed the first sentence of that section to:

"Our analysis has suggested that that biases in low-elevation shrub and grassland site ET might be due to incorrect simulations of seasonal vegetation dynamics; therefore, in order to obtain realistic estimates of ET and its component fluxes, it is important that the model can accurately simulate seasonal changes in leaf area and/or grass versus bare soil fractional cover."

And we have added the following sentence later in the paragraph after the original sentence "While not tested in this study, it is also possible that LSMs contain an inaccurate representation of different semi-arid vegetation *phenology*, including drought-deciduous shrubs and annual versus perennial C4 grasses". The new sentence is:

"The model does yet discern between perennial grasses and annual C4 grasses that only grow during warmest, wettest periods (Smith et al., 1997). It is possible that LSMs need new phenology models that account for annual C4 grass strategies in order to obtain accurate simulations of semi-arid water and carbon fluxes."

Developing new models that account for annual C4 grasses is also beyond the scope of this study unfortunately. We need to conduct separate analyses to develop such models, which will take some time (but we are working on it).

The paper discusses quite a few shortcomings of ORCHIDEE, or even LSMs in general. But there is no assessment of the importance of each point. They all seem to be similar important. I would have loved to see either prioritisation for model development or at least a guidance to the reader how to evaluate model shortcomings. The model might already be fine from an atmospheric perspective, or it might lead to a wet bias in spring.

The reviewer makes a good point here; however, it is hard to know how to prioritize model shortcomings. We did attempt to highlight issues that perhaps haven't been raised before in the final sentence of the conclusion (and this has been further adapted based on changes to the revised version):

"We recommend that future work on improving LSM semi-arid hydrological predictions focuses not only on issues highlighted in previous studies such as dynamic root zone moisture uptake, inclusion of ground water, lateral and vertical redistribution of moisture (e.g.

Whitley et al., 2016; 2017; Grippa et al., 2017) but also on: i) multi-variable calibration of vegetation and hydrology-related parameters across all sites; ii) more data to test modelled snow mass or depth at high elevation sites; iii) more data to better estimate and evaluate the seasonal trajectory of LAI across all sites and the vegetation fractional cover and LAI magnitudes at low elevation sites; and iv) testing of a more mechanistic description of resistance to bare soil evaporation."

We've discussed these points extensively above. We feel that these are the main contributions from this particular study and therefore serve as somewhat of a priority list, but we cannot evaluate how important they are compared to other issues that have been highlighted (e.g. the need for groundwater, dynamic root zone moisture uptake and lateral and vertical redistribution of moisture - which we also mention in the discussion) because we have not evaluated those components; indeed, they are not all implemented in the models yet. This is an age old issue in modeling - knowing which of the issues to focus on - and we appreciate it is frustrating.

Specific remarks are:
• I would change the tile. "Multi-variable" and "flux and storage" is tautologic. "Multi-configuration" is a bit much for two configurations.

The lead author admits she is not the best at formulating manuscript titles and thus agrees with the reviewer on this point. In response to the reviewer's comment, we suggest the title could be changed to:

"Testing water fluxes and storage from two hydrology configurations within the ORCHIDEE land surface model across US semi-arid sites"

• You should only cite one paper in preparation for CMIP6 and not once Ducharne et al. (in prep.) and once Peylin et al. (in prep.).

We have dropped the reference to the Peylin et al. paper in prep. The Ducharne paper is the relevant one for the hydrology.

• There are three personal communications, which are all from co-authors. Which co-author talked to which co-author?

It was the site PIs communicating with NM. However, we agree that given they are all co-authors these "pers. comms." are not needed so we have removed them.

• The description of "Richards and Darcy's equation" is strange. Darcy is part of Richards. The description is strange at two places (l.110 and l.211ff). I think that Richards equation is

known sufficiently so it is only interesting which form is solved, the saturation-based or the head-based form.

Agreed. We have removed the reference to Darcy and instead referred to it as the Richards equation around line 110 (introduction) and changed the sentence around line 211 to:

“The scheme implemented in ORCHIDEE relies on the one-dimensional Richards equation, combining the mass and momentum conservation equations, but is in the form of a Fokker-Planck equation that uses volumetric water content θ (m3m−3) as a state variable instead of pressure head.”

• If LAI was identified to be important why is no local LAI data used? I found local LAI data in Scott and Biederman (2017) for some of the sites.

Actually the LAI data in Scott and Biederman (2017) are from the MODIS satellite with a 1km resolution. Indeed we would love to have local LAI data to validate the model, and it is something we are looking into with a PhD student at the University of Arizona. As we explain in the discussion section on “Issues with modelling vegetation dynamics in semi-arid ecosystems” there are unfortunately no local LAI timeseries we can use at these sites - all the data in the associated papers are derived from satellite measurements, and given the spatial heterogeneity at the site is it impossible to say which vegetation type is dominating the signal at this resolution as LAI doesn’t scale linearly (i.e. you can’t unmix the signal based on % cover type, and in fact, estimates of % cover type are uncertain given the heterogeneity):
“Similarly, there are not many LAI measurements for grasses and shrubs in these ecosystems; therefore, we have relied on estimating the LAImax parameter from MODIS LAI data. While different satellite LAI products often correspond well to each other in terms of temporal variability, there is often a considerable spread in their absolute LAI values (Garrigues et al., 2008; Fan et al., 2013); therefore, the MODIS LAI data may not be accurate for these ecosystems. In any case, the satellite LAI values represent a mix of different vegetation types and unlike satellite reflectance data it is not possible to linearly unmix the satellite LAI estimates based on fractional cover. More field LAI measurements are needed from different vegetation types (especially annual versus perennial grasses and shrubs) to verify what the likely maximum LAI is for each PFT. ” Therefore, unfortunately at this time we cannot use local LAI data. We will revisit this in future studies if (hopefully, when) we get time series of field LAI data.

• Why are different T/ET algorithms used for different sites?

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.

However, we appreciate that our original manuscript lacked a lot of detail and explanation when it came to the T/ET ratio estimates: we did not explain why there are two methods, we 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.

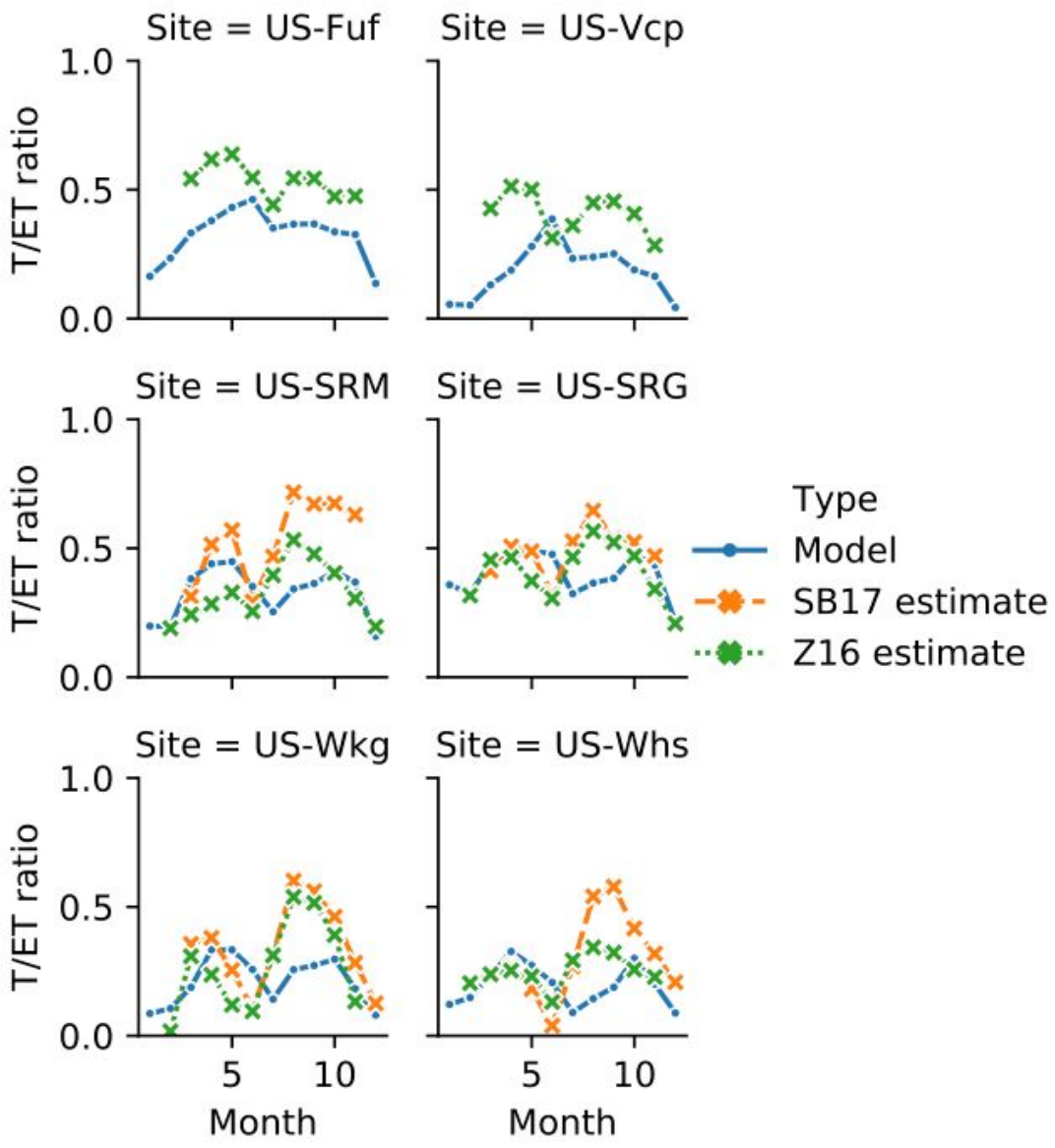
As the reviewer says below, there are a number of algorithms in the literature and it is hard to validate them. At the forested sites we only keep the Zhou et al. estimates for the reasons given above and at the lower elevation grass and shrub sites we now give estimates from both Zhou et al. (2016) 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.

Figure 6: Comparison of modelled and data-derived estimates of mean monthly T/ET ratios for each site. Forest site (US-Fuf and US-Vcp) T/ET estimates are derived using the method

of Zhou et al. (2016 – Z16 – green curve). Monsoon low-elevation grass- and shrub-dominated site T/ET estimated are based on both Zhou et al. (2016) and Scott and Biederman (2017 – SB17 – orange curve). Blue curves show the model ratios at each site. Please see Section 2.3.1 for details on methods for data-derived T/ET estimates.



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. "

• T/ET is seen as a measurement in the manuscript. But it is not. Any validation is missing in the Scott and Biederman (2017) paper, because it is pretty impossible to validate it. So T/ET should be seen only as an estimate. There are quite some algorithms in the literature to calculate T/ET and it is hard to tell why one should be more correct than the other.

We agree and shouldn't have ever referred to the T/ET ratios as "observations" we have changed all the text throughout to refer to these as "estimates" or "data-derived estimates".

• l.255: what is the subscript j on cj?

Thank you for spotting this. It refers to the PFT. We have added this into the manuscript. We have also changed all other subscripts referring to PFT to *j* and not *v* as was in the original manuscript.

• l.255ff: R(z) is explained but not nroot. If nroot were explained then one does not have to (confusingly) start the sums from 2 because nroot=0 in v=1 and i=1.

nroot is explained in the original manuscript on lines 257-258 (directly after explaining R(z): "In 11 LAY, a related variable is nroot(i), quantifying the mean relative root density of each soil layer i, so that ∑ nroot(i) = 1".

• l265ff: Why is the relative water content weighted with nroot? This formulation is an empirical observation and the beta term is never weighted by root length density (or similar) in the data papers (e.g. Keenan et al. (Biogeosci 2009)).

The exponential dependence of beta to soil moisture in the 2 layer scheme can be related to the convolution of SM and root density controls, as demonstrated by de Ronsay et al 1998. The root density control component was then extended by de Ronsay et al 2002 to the multi-layer scheme. Whilst it may not be in the data papers, we believe that an exponential decay of root density must be a common assumption, and therefore that convolution of SM and root density controls for plant water uptake are reasonable formulations. It is certainly a common approach in other LSMs (e.g. De Kauwe et al., 2015). These papers are already cited elsewhere in the model description section, particularly the De Kauwe paper in the new discussion section "Implications for modelling plant water stress" (see comment below) and we also highlight the need for calibrating water stress function parameters as well as parameters related to root zone uptake. But we can add a sentence clarifying this at this point in the manuscript if needed.

De Kauwe, M. G., Zhou, S.-X., Medlyn, B. E., Pitman, A. J., Wang, Y.-P., Duursma, R. A. and Prentice, I. C.: Do land surface models need to include differential plant species responses to drought? Examining model predictions across a mesic-xeric gradient in Europe, Biogeosciences, 12(24), 7503–7518, doi:10.5194/bg-12-7503-2015, 2015.
de Rosnay, P. and Polcher, J.: Modelling root water uptake in a complex land surface scheme coupled to a GCM, Hydrol. Earth Syst. Sci., 2, 239–255, https://doi.org/10.5194/hess-2-239-1998, 1998.

• l.268: Should W be in kg/m3 instead of kg/m2? Why is W used and not volumetric soil moisture theta?

The units are correct here (kg/m2). This takes into account the total water content in each layer of different thickness.

• l270ff: Why is p% = 0.8? There is quite some literature that it should be around 0.4 (e.g. Granier et al. (AFM 2007)), at least for forests?

The water stress function of the 11-layer hydrology scheme was inspired by the bucket model, of Manabe (1969), who used a value of 0.75 for the equivalent parameter to p%, and mentioned a plausible range of 0.7-0.8 based on Alpatev (1954).

A quick look at the literature shows that the range of values that is effectively used in LSMs is between 0.4 and 1 for the place in the WP-FC range at which the water stress function becomes 1 (corresponding no unstressed transpiration), regardless of the shape of the function (see for instance the review by Mahfouf et al 1998, or Verhoef and Gregorio, 2014).

MANABE, S., 1969: CLIMATE AND THE OCEAN CIRCULATION. Mon. Wea. Rev., 97, 739–774, https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2

Alpatev, A. M., “Vlagooborot kul’turnykh rastenil,” (Moisture Exchange in Crops), Gidrometeoizdat, Leningrad, 1954, 247 pp.

Mahfouf JF, Ciret C, Ducharne A, Irannejad P, Noilhan J, Shao Y, Thornton P, Xue Y, Yang ZL (1996). Analysis of transpiration results from the RICE and PILPS Workshop, Global and Planetary Change , 13, 73-88, doi:10.1016/0921-8181(95)00039-9

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.

As described in other responses to both reviewers, for many other parameters in this model we use the default values to test the default behavior (also to allow a comparison to forthcoming CMIP6 results), and have not performed a full calibration of all these parameters as this would take too long and is therefore outside the scope of this study. In the discussion we have discussed the need for parameter calibration, including the need to optimize “water-limitation parameters”. p% also is a universal parameter and not PFT-dependent. We have not investigated the need for PFT-dependence of this parameter but again we would take that into account when doing a parameter calibration.

• l.276f: The references are missing. And only the Keenan et al. paper actually supports this claim. The Zhou et al. papers do something very different and act only on stomatal conductance.

Thank you for pointing out the missing references. We have added these references in. However, we disagree that the Zhou et al. papers do something different and only act on Gs (also following discussion with collaborators on this work). See for example the following text in the 2013 paper: “The results are consistent with other stud- ies showing that both stomatal and non-stomatal processes are affected by drought (e.g. Egea et al., 2011; Keenan et al., 2010). Our analysis shows that non-stomatal limitation is considerable and has in general a greater impact than that of stomatal limitation on pho- tosynthetic rates. Photosynthesis under drought would be greatly overestimated if the decline in apparent Vcmax was not taken into account. Both assimilation rate and stomatal conductance decrease as pre-dawn leaf water potential declines, but assimilation rate usually decreases more – often many times more – than could be explained by a reduction in stomatal conductance (and g1) alone (see Figs. 1 and 2 in Appendix B).”
And from the 2014 paper “We found consistency among the drought responses of g1, gm, Vcmax and Jmax, suggesting that drought imposes limitations on Rubisco activity and RuBP regeneration capacity concurrently with declines in stomatal and mesophyll conductance”.
The beta functions are different in the Zhou studies (resulting in different shapes of water-limitation function).

Keenan, T., Sabate, S. and Gracia, C.: The importance of mesophyll conductance in regulating forest ecosystem productivity during drought periods, Global Change Biology, 16(3), 1019–1034, doi:10.1111/j.1365-2486.2009.02017.x, 2010.
Zhou, S., Duursma, R. A., Medlyn, B. E., Kelly, J. W. and Prentice, I. C.: How should we model plant responses to drought? An analysis of stomatal and non-stomatal responses to water stress, Agricultural and Forest Meteorology, 182-183, 204–214, doi:10.1016/j.agrformet.2013.05.009, 2013.
Zhou, S., Medlyn, B., Sabaté, S., Sperlich, D., Prentice, I. C. and Whitehead, D.: Short-term water stress impacts on stomatal, mesophyll and biochemical limitations to photosynthesis differ consistently among tree species from contrasting climates, Tree Physiology, 34(10), 1035–1046, doi:10.1093/treephys/tpu072, 2014.

• l.303f: I wondered if this claim means that you have a near perfect energy balance closure?

Energy balance closure at the low elevation sites is typically good, on the order of 10%. At the flagstaff site energy balance closure was 0.69 or greater for 30-minute values, and 0.81 or greater for daily values (Dore et al. 2010). But no, the close matching of annual ET with P indicates mainly these sites have very little runoff and drainage, i.e. most precipitation evaporates or transpires locally (also verified in the cited paper with additional hydrologic measurements).

• l.315f: why are there no site-specific soil characteristics? They must have been done at some point in the past.

In fact this sentence is misleading - these parameters have not all been measured at all sites. The parameters we need are mostly not available. No site has measured all the soil and hydraulic parameters we need (perhaps one or two) given the number and difficulty of

measuring them, and some sites don't have any measurements. So it makes it difficult to only use site-specific parameters for just a few of the values we need and not across all sites. We therefore have taken an approach that we only set site specific parameters if we have them for all sites and the rest we are effectively testing the default model parameters (which has the benefit that we're testing the default model behavior). We have added this sentence in to section 2.4 ("Simulation set-up and post-processing") and refer to this section around the lines the reviewer has highlighted in this comment.

"Due to the lack of available data on site-specific soil hydraulic parameters across the sites studied, we chose to use the default model values that were derived based on pedotransfer functions linking hydraulic parameters to prescribed soil texture properties (see Section 2.2.2). Using the default model parameters values also allows us to test the default behavior of the model."

However, as we point out in the end of results section 3.4, in the discussion section on Bare Soil Evaporation, and in the conclusions, it is possible that calibrating these hydraulic parameters at each site would be beneficial, as done in this study:

Shi, Y., Baldwin, D. C., Davis, K. J., Yu, X., Duffy, C. J. and Lin, H.: Simulating high-resolution soil moisture patterns in the Shale Hills watershed using a land surface hydrologic model, Hydrological Processes, 29(21), 4624–4637, doi:10.1002/hyp.10593, 2015.
We have added that reference to that sentence in the discussion section on bare soil evaporation.

It is also possible that further analyses using pedotransfer functions to determine soil hydraulic parameters from soil texture data at each site would be useful but we have not done this for this study - in part because the pedotransfer functions themselves are uncertain (Mermoud et al., 2006). Some of the authors are involved in ongoing investigations related to this topic. Taking all this into consideration, it's not clear that we would improve the accuracy or reliability of the model by using pedotransfer functions to derive these parameters, and as we said above it is useful (particularly considering ongoing CMIP6 experiments) to test the default behavior of the model. However, we have added the following sentence into the discussion section on bare soil evaporation (after adding the reference to Shi et al., 2015) to highlight that, along with statistical parameter calibration experiments, it may be possible (if needed) to better determine soil hydraulic properties following further investigation into the uncertainty surrounding available pedotransfer functions:

"Future studies could also investigate the impact of uncertainty in the use of pedotransfer functions (e.g. Mermoud et al., 2006) in deriving soil hydraulic parameters from soil texture information. "

Mermoud, A. and Xu, D.: Comparative analysis of three methods to generate soil hydraulic functions, Soil and Tillage Research, 87(1), 89–100, doi:10.1016/j.still.2005.02.034, 2006.

• Fig. 1: Where are the observations?

We have added ET observations but not the observations for soil moisture variables because in this plot these given as total water content (see comment below) to see overall mean changes in the amount of water in the upper and total soil column and therefore have not been re-scaled to match observations (as we outline in Section 2.3.2). Instead, we use the re-scaled soil moisture observations for all other plots. We also propose adding the following in the Figure 1 caption to make this point clear:

"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."

We have also changed the description of how we process soil moisture data in Section 2.3.2 to highlight this point:

"Therefore, with the exception of Fig. 1 in which we examine changes in total water content between the two model versions, for the remaining analyses we do not focus on absolute soil moisture values in the model – data comparison, we specifically investigate how well the model captured the temporal dynamics at specific soil depths."

• Fig. 1: Harmonise scales of ET, Runoff and Drainage, as well as of Upper SM and Total SM so that one can compare the fluxes/stocks. For example, why is Total SM up to 1000? If kg/m3, then Upper SM and Total SM could have the same scale. If kg/m2, they should be scaled according to layer depth.

We have harmonized the scales for all variables with the same units.

The units are kg/m2, not kg/m3. The total SM sums up SM over all the layers (0-2m - as in the y-axis title). The upper layer is only over the top 10cm. The max value shouldn't have been 1000 - this has been adjusted. We can convert these to m3/m3 (volumetric water content instead of total water content) if the reviewer would prefer so the upper layer and total column scales can be more comparable.

Fig. 1: Why is there (almost) no drainage at forested sites with the 11-layer version? Is this realistic? There is only a very small mention for US-Fuf in the text.

It is unfortunate that we don't have more data on runoff and drainage across all these sites, as we mention in the discussion. We do have the following sentence for US-Fuf in the text as the reviewer points out: "The 11LAY limited drainage is also likely to be the case at US-Fuf given that nearly all precipitation at the site is partitioned to ET (Dore et al., 2012).". We don't have any corresponding data for US-Vcp unfortunately. However, in general these semiarid flux sites have very little precipitation that is not accounted for by ET, at the annual scale (i.e.

looking at ET:P ratios). This means that precip can be much higher than ET for some months (winter) but "catch up" during others (spring, early summer). See Biederman et al. (2017) Table S1. We have included this sentence where we talk about drainage:

"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)."

• Fig. 2: I think the titles of the y-axes of row 3 and 4 are swapped.

The y-axes labels of rows 3 and 4 are correct but the description in the caption is the wrong way round - thank you for spotting that. This was also wrong for Fig. S2 so we have corrected the captions for both figures.

• Fig. 4: please put the 2 cm, 20, cm and 50 cm plots on the same scales.

Done, thank you (and for Fig. S4).

• Fig. 5b: Data stays low during much of the snowfall period. This can happen if the data is measured inside a forest whereas the model assumes open space. Much of SnowMIP's model intercomparison, at which ORCHIDEE probably par- ticipated, focussed on open sites. We might not know well the behaviour of our models at forest sites.
It looks like that the data is even decreasing at the beginning of the snowfall period. This could point to soil freezing. Some soil moisture sensors measure only liquid water, so low values are measured during frozen soil conditions. So sites also do not include possible ice phases in their transformations from voltage to soil moisture.
Both processes were not discussed.

We thank the reviewer for pointing these processes out. We looked at the modeled surface temperatures and indeed found that the early winter positive model-data bias (model higher than the data) coincided with negative surface temperatures (which is therefore possibly related to instrument biases or the issue of open sites vs a closed forest setting as the reviewer mentioned). So we have increased the description and discussion of these results and made the paragraphs related to snow biases at the high elevation sites more nuanced. These paragraphs now read:

"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.

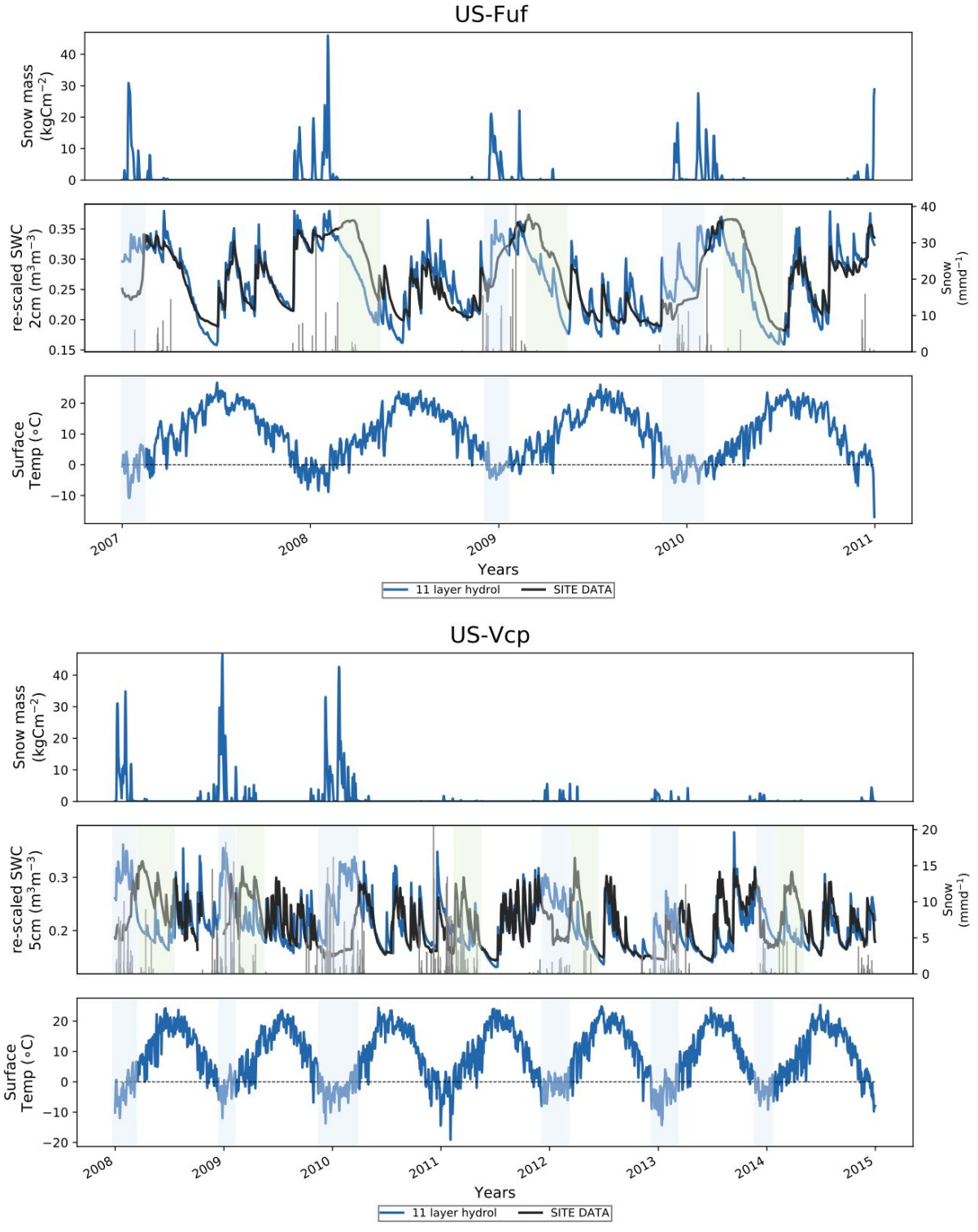
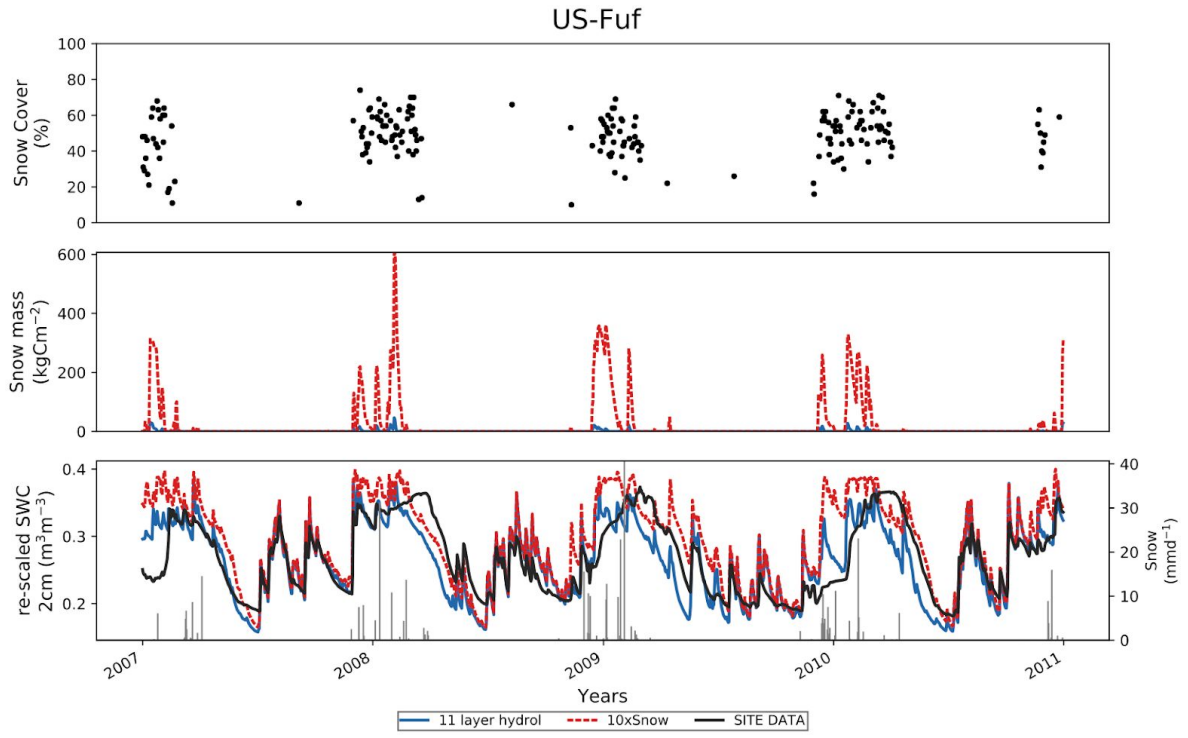


Figure S5: Linear regressions between spring (March-April) mean monthly LAI (m2m-2) and spring mean monthly ET (mmmonth-1) model-data misfits for each site. The dominant PFT is given in brackets for each site. See Table 1 for PFT acronyms.



We have also 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."

• l.384f: This is a "false friend" to me. Evaporation is water vapour but the Richards equation (as used in ORCHIDEE) does not include vapour transport in the soil. So the model has to compensate for this omittance. This is one of the primary reasons why the Richards solvers need very thin layers at the top of the soil. These layers cannot be seen as physical layers because they have to compen- sate for all the model deficiencies on top of possible litter layers. It is thus doubtful that these first few layers should be compared satellite measurements.

To solve Richards equation, we need thin layers at the atmosphere interface, not because we need to compensate for model deficiencies in lacking vapor transport but because the

moisture gradients are larger (as we discussed in section 2.2.2). Models representing vapor transfer have even thinner discretization.
Concerning the comparison with satellite data, we agree that the non representation of vapor transfers, could lead to an overestimation of soil moisture in the surface layers but could be balanced by the fact that the satellite sounds also a deeper soil in dry soil conditions. But given the fact that the average sensing depth of the microwave instruments is of a few centimeters, the capacity of the model to represent thin layers compared to the 2LAY is a benefit. The challenges and benefits of how to compare model soil moisture with satellite soil moisture are discussed extensively in Raoult et al. (2018) (which we cite here).

• l.396: Isn't this a contradiction to Whitley et al. (2016). You state in the introduction that Whitley et al. (2016) found that T of the vegetation is mostly too low in the models. Is 2-layer ORCHIDEE different so that 11-layer ORCHIDEE can decrease T during the warm season?

Indeed this is a good point. It is not so much that the 2-layer and 11-layer are different here so much as the modification to the formulation of beta (water stress function) has allowed there to be a greater decrease in T during the hot, dry (water limited) periods - as highlighted by the brown shaded zones in Figure 2. We state this in the previous sentence at lines 392-394 in the original manuscript with the following sentence "At the low-elevation shrub and grass sites, the improvement in ET is also related to changes between the two versions in the calculation of the empirical water stress function, b (Figs. 2 and S2 5th panel), which acts to limit both photosynthesis and stomatal conductance (therefore, T) during periods of moisture stress (Section 2.2.4)."

However, we agree that it's worth noting what is, what isn't, similar to the findings of Whitley et al. (2016) in our study given we have highlighted that study in the introduction. Therefore, we have added an small extra section to the discussion with the following text (that also takes the opportunity to discuss more broadly about modeling plant response to water stress):

"**Implications for modelling plant water stress**

Similar to Whitley et al. (2016), the original 2LAY version of the model underpredicted wet monsoon season ET. The peak ET fluxes were generally much better captured in the 11LAY version. However, in contrast to Whitley et al. (2016), the 2LAY simulations overestimated ET during the hottest, driest period between May and June. Our results demonstrated that a modified empirical beta water stress function (used to downregulate stomatal conductance during periods of limited moisture) that takes into account available soil moisture and root density across the entire soil column (Section 2.2.4) helped to better capture dry season ET dynamics. These results are interesting in light of previous studies showing that LSMs employing empirical beta water stress functions show considerable differences in their simulated soil moisture response to during water stressed periods (Medlyn et al., 2016; De Kauwe et al., 2017). These studies argue for more evidence-based formulations of plant response to drought. De Kauwe et al. (2015) also highlight the need for models to

incorporate dynamic root zone soil moisture uptake down profile as the soil dries. It is therefore possible that while the modified beta function used in the 11LAY does help to capture seasonal water stress, as in this study, new mechanistic plant hydraulic schemes that can track transport of water through the xylem (e.g. Bonan et al., 2014; Naudts et al., 2015) may be needed when simulating plant response to prolonged drought periods. However, comparing beta functions versus plant hydraulic schemes under severe water stressed periods was not within the scope of this study. 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."

Bonan, G. B., Williams, M., Fisher, R. A. and Oleson, K. W.: Modeling stomatal conductance in the earth system: linking leaf water-use efficiency and water transport along the soil–plant–atmosphere continuum, Geoscientific Model Development, 7(5), 2193–2222, doi:10.5194/gmd-7-2193-2014, 2014.
De Kauwe, M. G., Zhou, S.-X., Medlyn, B. E., Pitman, A. J., Wang, Y.-P., Duursma, R. A. and Prentice, I. C.: Do land surface models need to include differential plant species responses to drought? Examining model predictions across a mesic-xeric gradient in Europe, Biogeosciences, 12(24), 7503–7518, doi:10.5194/bg-12-7503-2015, 2015.
De Kauwe, M. G., Medlyn, B. E., Walker, A. P., Zaehle, S., Asao, S., Guenet, B., Harper, A. B., Hickler, T., Jain, A. K., Luo, Y., Lu, X., Luus, K., Parton, W. J., Shu, S., Wang, Y. P., Werner, C., Xia, J., Pendall, E., Morgan, J. A., Ryan, E. M., Carrillo, Y., Dijkstra, F. A., Zelikova, T. J. and Norby, R. J.: Challenging terrestrial biosphere models with data from the long-term multifactor Prairie Heating and CO2 Enrichment experiment, Global Change Biology, 23(9), 3623–3645, doi:10.1111/gcb.13643, 2017.
Medlyn, B. E., Kauwe, M. G. D., Zaehle, S., Walker, A. P., Duursma, R. A., Luus, K., Mishurov, M., Pak, B., Smith, B., Wang, Y.-P., Yang, X., Crous, K. Y., Drake, J. E., Gimeno, T. E., Macdonald, C. A., Norby, R. J., Power, S. A., Tjoelker, M. G. and Ellsworth, D. S.: Using models to guide field experiments:a prioripredictions for the CO2response of a nutrient- and water-limited native Eucalypt woodland, Global Change Biology, 22(8), 2834–2851, doi:10.1111/gcb.13268, 2016.
Miller, G. R., Chen, X., Rubin, Y., Ma, S. and Baldocchi, D. D.: Groundwater uptake by woody vegetation in a semiarid oak savanna, Water Resources Research, 46(10), doi:10.1029/2009wr008902, 2010.
Naudts, K., Ryder, J., Mcgrath, M. J., Otto, J., Chen, Y., Valade, A., Bellasen, V., Berhongaray, G., Bönisch, G., Campioli, M., Ghattas, J., Groote, T. D., Haverd, V., Kattge, J., Macbean, N., Maignan, F., Merilä, P., Penuelas, J., Peylin, P., Pinty, B., Pretzsch, H., Schulze, E. D., Solyga, D., Vuichard, N., Yan, Y. and Luyssaert, S.: A vertically discretised canopy description for ORCHIDEE (SVN r2290) and the modifications to the energy, water

and carbon fluxes, Geoscientific Model Development, 8(7), 2035–2065, doi:10.5194/gmd-8-2035-2015, 2015.
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

• l.448ff: There also seems to be a problem with infiltration. At the model attenu- ates precipitation peaks too much at forest sites, while it is almost not attenuating at the grassland sites. Could you explain that please. There seems to be a differ- ence in the model why water can flow quickly to deep layers in grassland but not in forests. Or is it the bare soil fraction?

We thank the reviewer for raising this issue because in fact we omitted one important change of saturated hydraulic conductivity (Ks) with depth, which is an exponential increase in Ks towards the surface to account for the effect of increased soil porosity due to bioturbation by roots. Given tree roots are deeper this increase towards the surface starts lower in the profile, and as Ks increases towards the surface, so does the infiltration capacity. Therefore, infiltration under the forests is likely to be quicker, which we believe explains the smoother profiles at depth under the forested sites (although looking at the full timeseries in Fig. S4a the model doesn't do a bad job at US-Fuf of capturing the largest swings in soil moisture in the deepest layer - the smooth model temporal profile at depth is more of an issue at US-Vcp). This explains the difference in the model behavior between the forest and grass sites; however, it doesn't explain why the model simulations don't capture the observed soil moisture dynamics as well at depth. One reason may be that in the absence of PFTs defined specifically for semi-arid ecosystems we are essentially modeling the trees and shrubs in these ecosystems as temperate trees. One parameter that might be very different is the root density decay factor. Semi-arid shrubs and trees tend to have deeper tap roots than their temperate counterparts to account for limited water availability. Often they also have extensive shallow root systems, which is not something we can account for in ORCHIDEE. And this still doesn't explain the model-data differences at depth at the grass sites.

The forested sites also tend to be silt or clay loams, whereas the grass and shrub sites are more sandy loams. The latter has a higher ks, which results in a slightly faster decrease in the Ks downwards through the soil profile with the equation that accounts for decrease in Ks with soil compaction. However, this would counter the effect of changes in Ks with depth described above due to root zone bioturbation and we expect the effect to be much smaller at depth.

Despite not having a clearer answer to the reviewer's question, we agree we failed to explain or discuss any of the above mentioned points. Including these points would greatly aid a reader in understanding this issue. Therefore, we have adapted the manuscript text in several places to account for this.
In the description of the 11 layer hydrology in Section 2.2.2 we have added the following sentence:

“Ks increases exponentially with depth near the surface to account for increased soil porosity due to bioturbation by roots, and decreases exponentially with depth below 30cm to account for soil compaction (Ducharne et al., in prep).”

In addition, where we initially described the results of the differences in soil moisture at depth (original submission line 452 - Section 3.2), we have added the following:

“The smoother model temporal profile at depth at the forest sites compared to the sites with higher grass fraction is likely related to impact of rooting depth on exponential changes in Ks towards the surface (see Section 2.2.2). As the forests have deeper roots, the increase in Ks starts from a lower depth in the soil profile than the more grass-dominated sites, which in turn allows for a quicker infiltration of moisture to deeper layers and decreased simulated soil moisture temporal variability. However, this description of the model behaviour does not explain the model-data discrepancies.”

And at the end of the same paragraph we have modified the original text so it now reads:

“Alternatively, 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, changes in soil texture that in reality may occur much deeper in the soil could alter hydraulic conductivity parameters; in the model however, hydraulic conductivity only changes exponentially with depth owing to soil compaction (see Section 2.2.2).”

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 model-data discrepancies in soil moisture at lower depths

• l.454: I was wondering why the model was not tested with more layers, say 100?

11-layers is a compromise between computational cost vs accuracy. In the initial development of the model they tested several different vertical soil discretizations and found that 11 layers was a good compromise for unsaturated soils (de Rosnay et al., 2000). In Campoy et al. (2013) they tested the effect of alternative soil bottom boundary conditions (including impermeable soil bottom and prescribed water table depth). This led them to increase the number of layers to 20 to describe the hydraulic gradients with enough accuracy; however, this is not necessary for unsaturated soil and additional layers

significantly increase the CPU requirement. Also, given 11-layers is the default version used in CMIP6, we based all our simulations on that version and chose not to test a different number of layers. We suggest adding in the following sentence to Section 2.2.2 (describing the 11-layer hydrology) for clarification on this point:

“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.