

Dear authors,

Two reviewers have given supportive feedback on your manuscript and a great deal of constructive advice. I agree with the reviewers that the manuscript is of interest to the readership of HESS, although it still needs some work. You have already addressed the scientific comments in the extensive discussion.

I would like to emphasize the comment of Reviewer 2 who states that the manuscript sometimes seems to lose focus and objectivity, which is an observation I share. The new title is a good step in that direction, because it is focussed on water storage and water fluxes. Additionally to the propositions already made in the discussion, I suggest adding quantitative means for comparison of modeled and observed data (correlation and for ET also bias, variation ratio) as it would help to objectively assess the two setups.

Also, reading the discussion with both reviewers: I think a table summarizing the differences between the two model setups would help to navigate the paper.

Lastly, I would like to extend many thanks to both authors and reviewers for a truly in depth exchange on this manuscript! I am confident this review round will contribute to an improved manuscript. Please send a step by step response and a manuscript with highlighted changes. I am looking forward to the new version.

Sincerely,  
Anke Hildebrandt

Dear Professor Hildebrandt,

Thank you very much for your response and decision on our paper. We are grateful to both you and the two reviewers for your detailed and useful suggestions. We believe that by providing the detailed point by point response to the reviewers and associated edits to the manuscript, as well as by following your suggestions as detailed below, we have strongly improved the focus, clarity, and objectivity of the manuscript.

We agree with you and reviewer 2 that the original manuscript did lack a more detailed and objective comparison between the two model configurations. We had attempted to perform such a detailed comparison for the evapotranspiration but had not done so for the soil moisture, for reasons we have given in our response to reviewer 2. However, we have accounted for this issue in two ways: 1) as per reviewer 2's suggestion, by adding observations of upper layer soil moisture to Figures 2 and S2; and 2) as per your suggestion, by adding the correlation between the re-scaled upper layer soil moisture and the observations to our original Table 3 (now table 4) that contains R, RMSE and bias metrics for ET. Both the changes in Figs 2 and S2 and in Table 3 (now 4) have also been accompanied by additions to the text in Section 3.1 that describes the changes between the 2 and 11layer configurations.

Table 3 originally had the R, RMSE and mean absolute bias between the daily ET model and observations. As mentioned above we have now added the R between the upper layer soil moisture and observations, as you suggested and we have added the following sentence into Section 3.1 to describe those results: "This improvement in upper layer soil moisture temporal dynamics is also indicated by the strong increase in correlation at all sites between the re-scaled modelled and observed 11LAY upper layer soil moisture compared to the 2LAY (increases in R ranged from 0.1 to 0.48 – Table 4)."

We have also added to Table 3 (now table 4) the ratio of the modelled to observed ET standard deviation as a measure of their relative variability, as you suggested and have added the following sentence into Section 3.1: "The ratio of modelled to observed standard deviation in ET,  $\alpha$ , is also provided as a measure of relative variability in the simulated and observed values (Table 4). With the exception of US-Fuf,  $\alpha$  values tend closer to 1.0 in the 11LAY simulations compared to the 2LAY – highlighting again that the 11LAY version does a better job of capturing the daily variability. The higher ET model-data bias and  $\alpha$  at US-Fuf is mostly due to model discrepancies in spring (Fig. S2a), which we discuss further in Section 3.3."

We acknowledge that we did not originally state in the methods that we were going to calculate these model evaluation metrics; therefore, a reader would not know to look out for such a table. We have now added the following sentence in to Section 2.4 that describes the simulation setup and postprocessing: "To evaluate the two model configurations we calculated the Pearson correlation coefficient between the simulated and observed daily time series for both the upper layer soil moisture (with the model re-scaled according to the linear CDF matching method given in Section 2.3.2) and ET. We also calculated the RMSE, mean absolute bias, and a measure of the relative variability,  $\alpha$ , between the modelled and observed ET. The latter is calculated as the ratio of model to observed standard deviations ( $\alpha = \sigma_m/\sigma_o$ ) based on Gupta et al. (2009)."

We strongly hope these modifications bolster our argument that the 11layer version does a better job at capturing the temporal dynamics of the upper layer soil moisture, which in turn improves the model predictions of ET when compared with the observations. The addition of the soil moisture observations to Figures 2 and S2 are very helpful in seeing the improvement in the 11layer soil moisture. And the additional model evaluation metrics (the soil moisture R and ET relative variability) in Table 3 (now 4) also further help our argument. Thank you both for these suggestions.

Finally, we have included the second table you suggested summarizing the differences between the two configurations, and we wholeheartedly agree that this will be very useful to the reader. I am not sure why we didn't think to do that before. This is now the new Table 2 and we have included the following sentence at the end of the main model description in Section 2.2.1 to reference this table: "The main differences between the two ORCHIDEE configurations used in this study are described in the sections below and are summarised in Table 2."

In this submission we provide our detailed responses to all reviewer and editor comments as well as a track change version of the manuscript with all associated edits that we described in the response to reviewers. We hope that you will find this is a significant improvement on the original submission.

Yours sincerely,  
Natasha MacBean (on behalf of all authors)

## 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](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](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](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 *this 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  $f_v^j = f^j (1 - e^{-(k_{ext} \cdot LAI_j)})$ , 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  $k_{ext}$  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 CO<sub>2</sub> 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 CO<sub>2</sub> fluxes. As mentioned in our initial information response to reviewer 1 we are investigating model representation of CO<sub>2</sub> fluxes in a separate study. The issues related to CO<sub>2</sub> 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 title. "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  $\theta$  ( $m^3m^{-3}$ ) 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  $ET \neq 0$  (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.

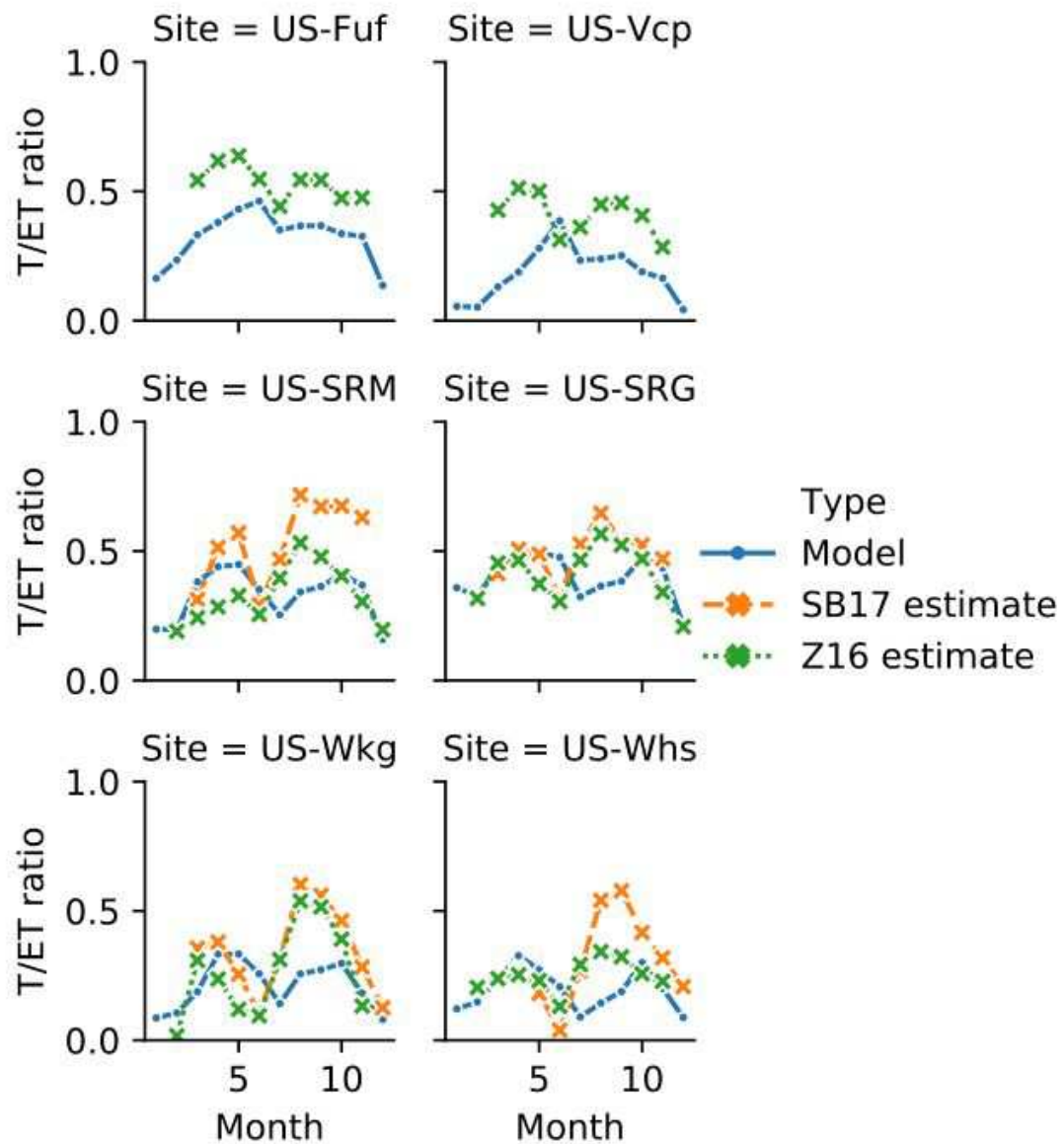
As the reviewer says below, there are a number of algorithms in the literature and it is hard to validate them. At the forested sites we only keep the Zhou et al. estimates for the reasons given above and at the lower elevation grass and shrub sites we now give estimates from both Zhou et al. (2016) 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 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 ( $uWUE_a$ ) to the potential  $uWUE$  ( $uWUE_p$ ).  $uWUE_a$  is calculated based on a linear regression between ET and  $GPP.VPD_{0.5}$  at observation timescales for a given site, whereas  $uWUE_p$  was calculated based on a quantile regression between ET and  $GPP.VPD_{0.5}$  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  $ET \neq 0$  when  $GPP = 0$  (Biederman et al., 2016). The Z16 method also assumes the  $uWUE_p$  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 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”.

- I.255: what is the subscript  $j$  on  $c_j$ ?

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.

- I.255ff:  $R(z)$  is explained but not  $n_{\text{root}}$ . If  $n_{\text{root}}$  were explained then one does not have to (confusingly) start the sums from 2 because  $n_{\text{root}}=0$  in  $v=1$  and  $i=1$ .

$n_{\text{root}}$  is explained in the original manuscript on lines 257-258 (directly after explaining  $R(z)$ ): “In 11 LAY, a related variable is  $n_{\text{root}}(i)$ , quantifying the mean relative root density of each soil layer  $i$ , so that  $\sum n_{\text{root}}(i) = 1$ ”.

- I.265ff: Why is the relative water content weighted with  $n_{\text{root}}$ ? 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 Ronsay, 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.

- I.268: Should  $W$  be in  $\text{kg/m}^3$  instead of  $\text{kg/m}^2$ ? Why is  $W$  used and not volumetric soil moisture  $\theta$ ?

The units are correct here (kg/m<sup>2</sup>). This takes into account the total water content in each layer of different thickness.

- I270ff: 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 Alpatov (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](https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2)

Alpatov, 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.

- I.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  $G_s$  (also following discussion with collaborators on this work). See for example the following text in the 2013 paper: “The results are consistent with other studies 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 photosynthetic rates. Photosynthesis under drought would be greatly overestimated if the decline in apparent  $V_{cmax}$  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  $g_1$ ) alone (see Figs. 1 and 2 in Appendix B).”

And from the 2014 paper “We found consistency among the drought responses of  $g_1$ ,  $g_m$ ,  $V_{cmax}$  and  $J_{max}$ , 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.

- I.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).

- I.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/m<sup>3</sup>, then Upper SM and Total SM could have the same scale. If kg/m<sup>2</sup>, they should be scaled according to layer depth.

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

The units are kg/m<sup>2</sup>, not kg/m<sup>3</sup>. 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 m<sup>3</sup>/m<sup>3</sup> (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 participated, 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.

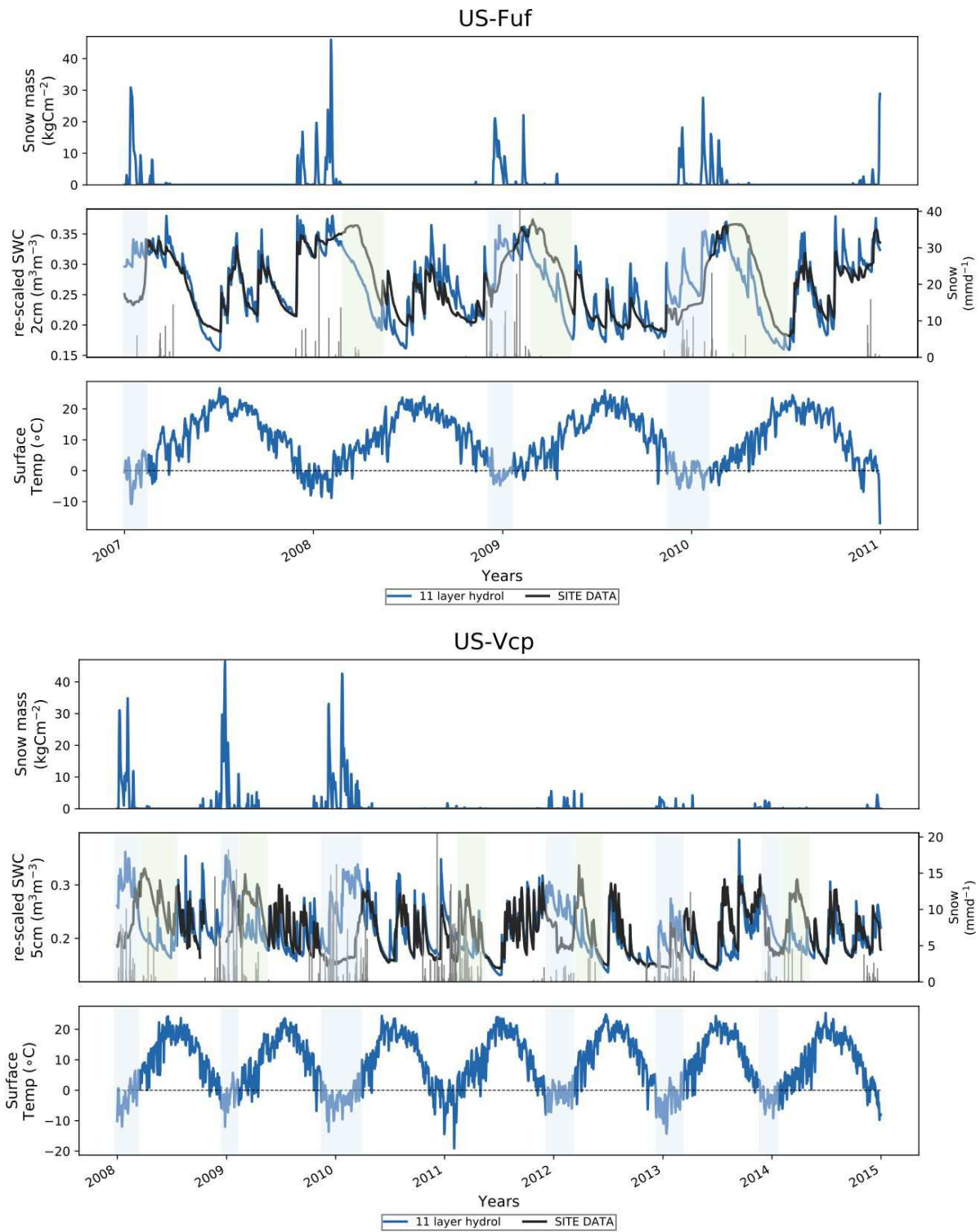
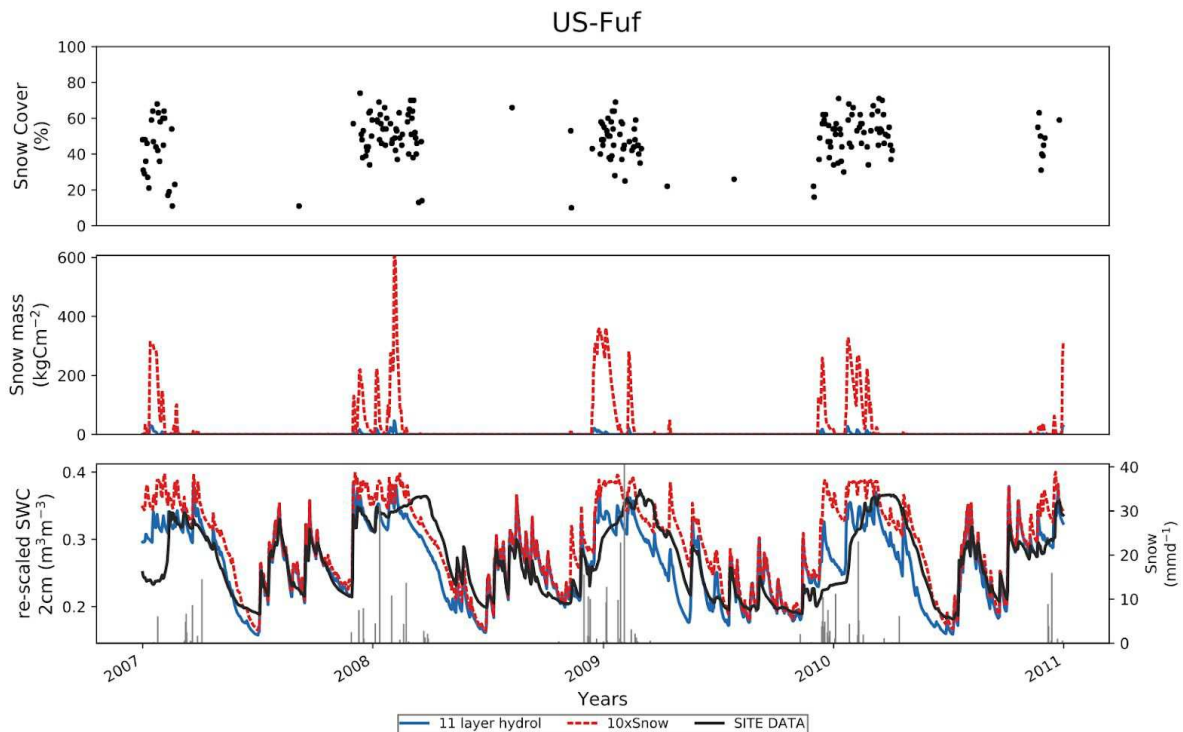


Figure S5: Linear regressions between spring (March-April) mean monthly LAI ( $m^2m^{-2}$ ) and spring mean monthly ET ( $mmmonth^{-1}$ ) model-data misfits for each site. The dominant PFT is given in brackets for each site. See Table 1 for PFT acronyms.



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.”

• I.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 omission. 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 compensate 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).

• I.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  $\geq 2.4\text{mm d}^{-1}$  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 CO<sub>2</sub> 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 priori predictions for the CO<sub>2</sub> response 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, Ducharme 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

• I.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 ( $K_s$ ) with depth, which is an exponential increase in  $K_s$  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  $K_s$  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  $k_s$ , which results in a slightly faster decrease in the  $K_s$  downwards through the soil profile with the equation that accounts for decrease in  $K_s$  with soil compaction. However, this would counter the effect of changes in  $K_s$  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

• I.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.

## 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  $c_j$  (in  $m^{-1}$ ) is the thethe root density decay factor for PFT  $j$ . For a 2m soil profile,  $c_j$  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  $ET \neq 0$  (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 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 ( $uWUE_a$ ) to the potential  $uWUE$  ( $uWUE_p$ ).  $uWUE_a$  is calculated based on a linear regression between ET and  $GPP.VPD_{0.5}$  at observation timescales for a given site, whereas  $uWUE_p$  was calculated based on a quantile regression between ET and  $GPP.VPD_{0.5}$  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  $ET \neq 0$  when  $GPP = 0$  (Biederman et al., 2016). The Z16 method also assumes the  $uWUE_p$  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  $\sim 2.4$ . 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 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.

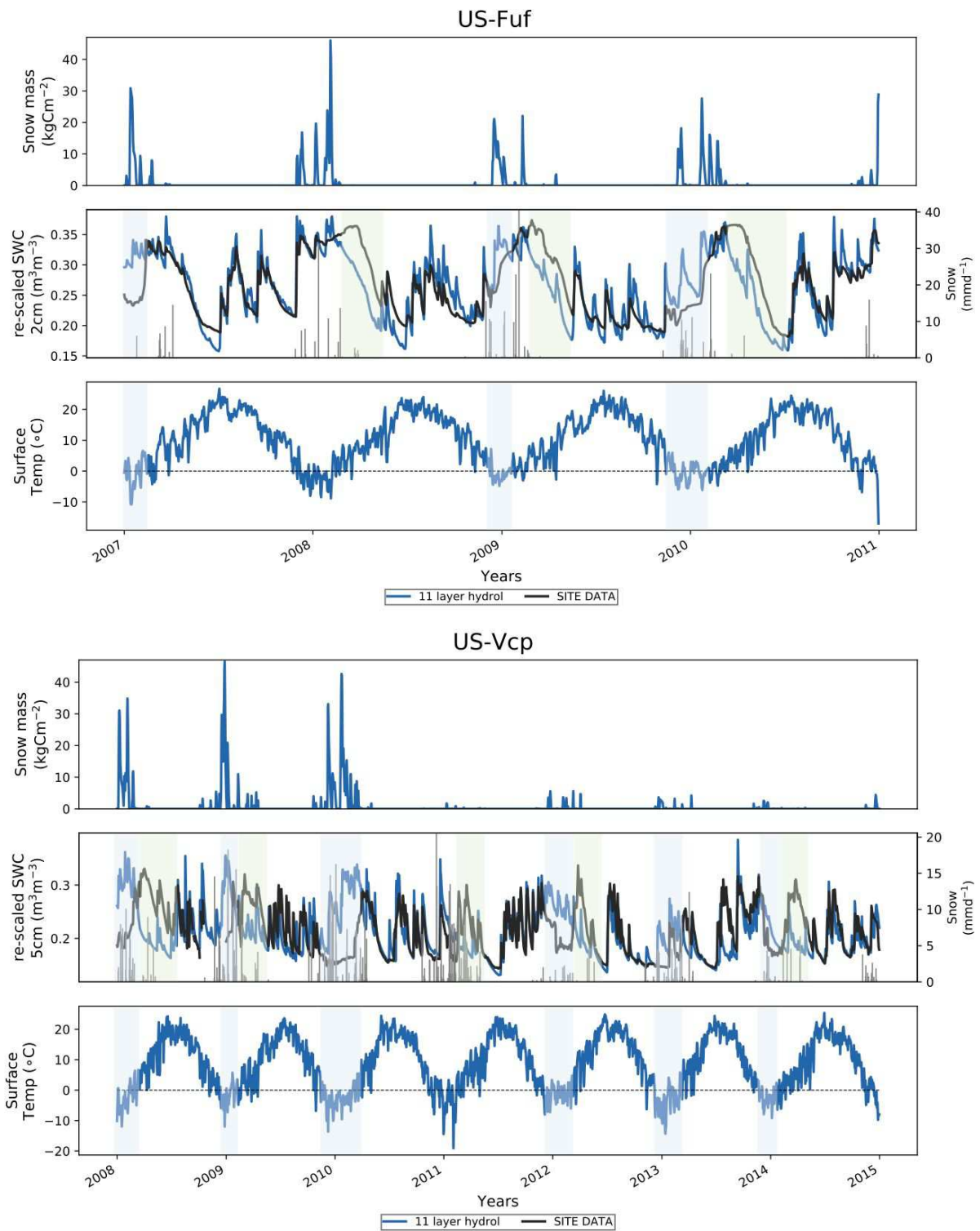
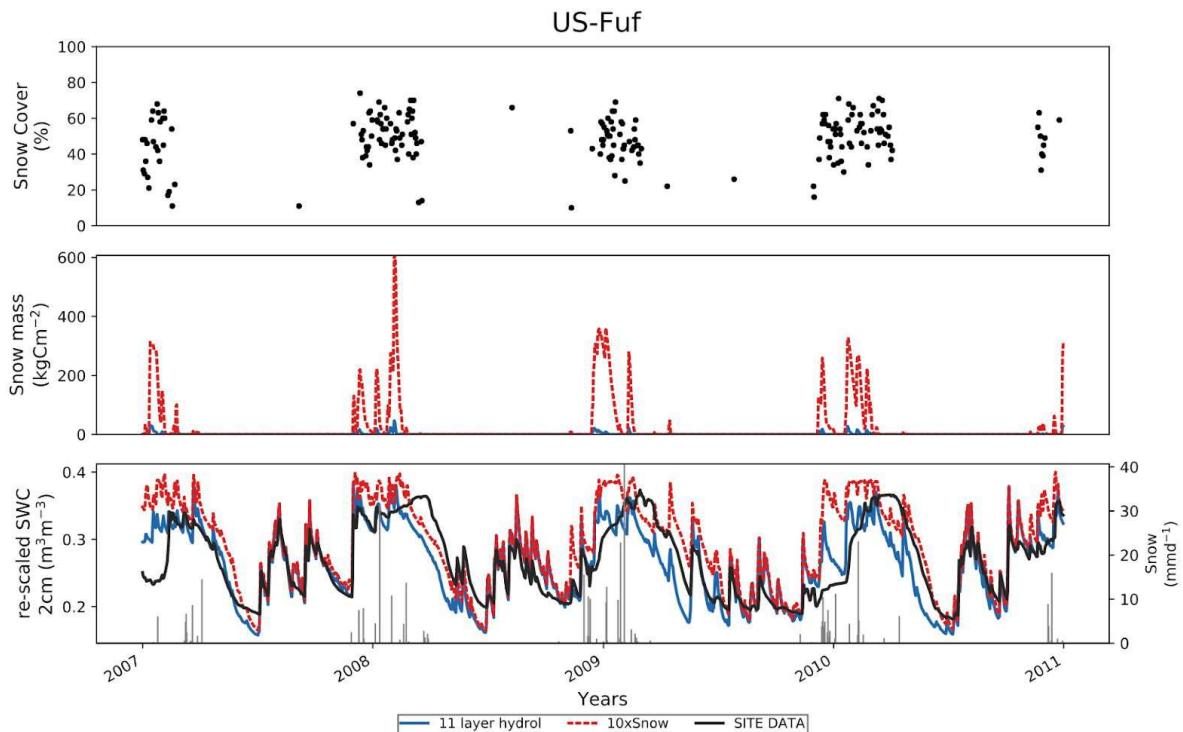


Figure S5: Linear regressions between spring (March-April) mean monthly LAI ( $\text{m}^2\text{m}^{-2}$ ) and spring mean monthly ET ( $\text{mmmonth}^{-1}$ ) model-data misfits for each site. The dominant PFT is given in brackets for each site. See Table 1 for PFT acronyms.





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 on to 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 ( $\geq 25\text{km}$ ). 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.

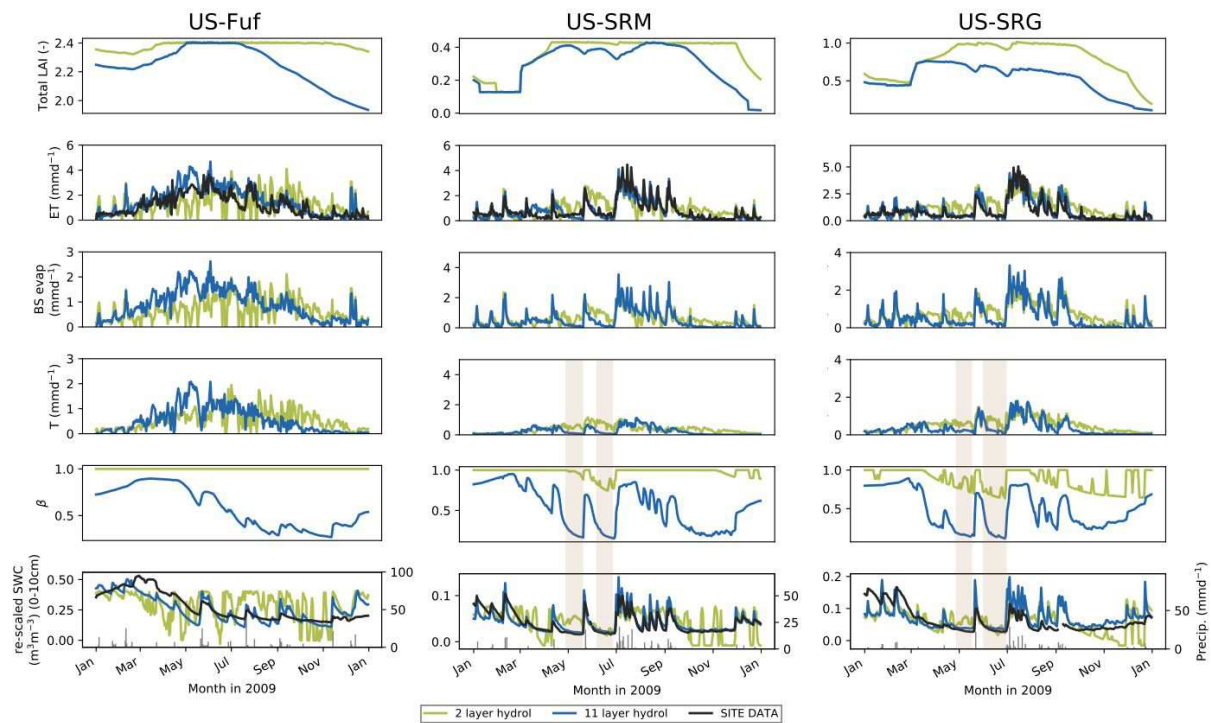
Similarly, a large part of paragraph 3.1 gives a description on the differences between the two model set-ups, and discusses Figure 1. Nevertheless, without any idea on how reality looks like, it is hard to really get an understanding on what is actually better. So I am not sure if this part of the paragraph 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., & Saxton, 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 Ducharme 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  $d$

$h_d^{\dagger}$  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

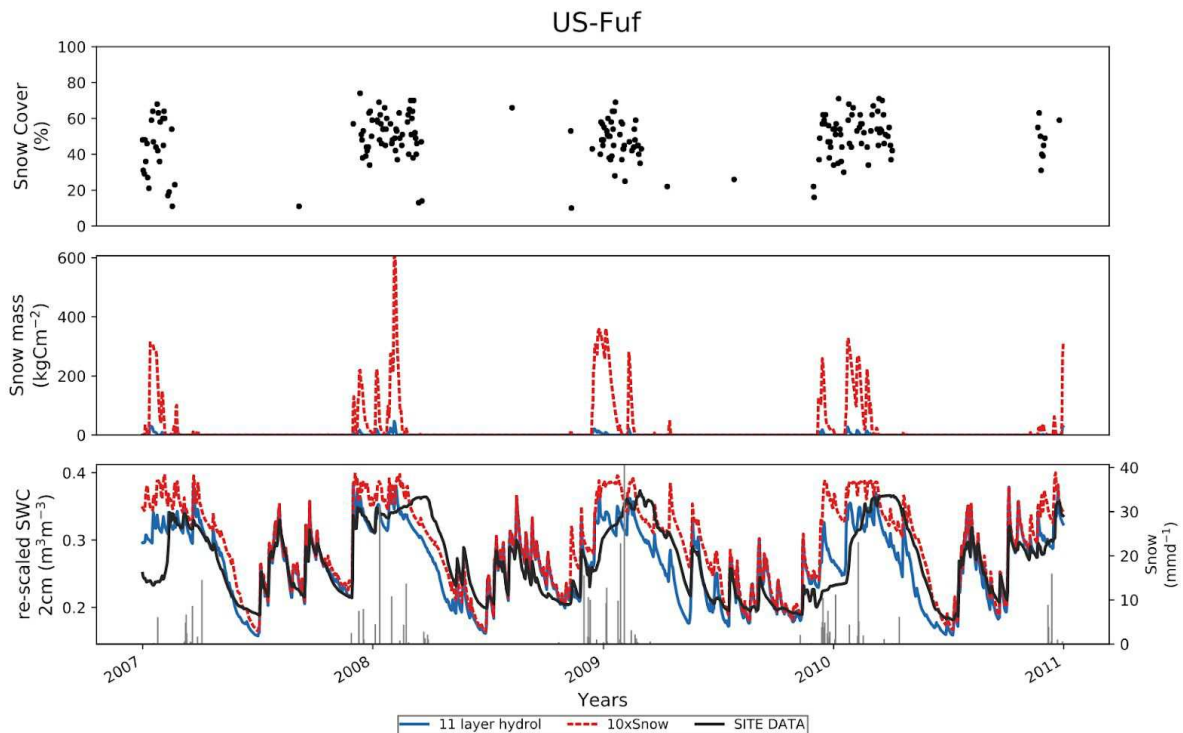
Changed - thank you.

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 E<sub>i</sub> 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<sup>-1</sup>)" 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).

Interactive comment on “Multi-variable, multi-configuration testing of ORCHIDEE land surface model water flux and storage estimates across semi-arid sites in the southwestern US” by Natasha MacBean et al.

Anonymous Referee #2

Received and published: 20 February 2020

I would like to thank the authors for their openness and the discussion, below I tried to reply to their questions in the informal response.

[Thank you very much for considering our informal response and getting back to us so quickly. These additional comments helped to clarify our edits to the manuscript and our responses to the formal reviews.](#)

I can see the difficulties the authors raise with regard to comparing the 2LAY and 11LAY soil moisture values, and also understand why the soil moisture values are not compared to observations for the 2LAY-model. For me, it is not a problem that you cannot use the 2LAY-values, but I just wonder what the point is of comparing 11LAY- results with soil moisture if you cannot do the same for the 2LAY-model. This also depends on the goal of the comparison, because you cannot use it to assess which of the models is better (which I believe is the main goal of the paper, and also how I interpreted this section). I believe it could serve as an explanation why the ET-values are better, but some textual changes may be needed to clarify this. In the current version, this comparison seems rather important, and relates to some conclusions, whereas it is merely an additional and supportive explanation for some other more important findings.

[We agree with the reviewer and in fact we did originally have a comparison to the 2LAY moisture but removed it for reasons we detail in the response to their formal \(original\) review. We have now proposed adding in a comparison to soil moisture observations to Figure 2, which compares the 2 vs 11LAY model. Please see our full response and updated figure in the response to the formal review.](#)

Regarding the second point of the authors, and I am sorry for not making it easier, but I strongly disagree with reviewer 1 that you should remove the 2-layer versus the 11-layer comparison. This is for me the key-point of the manuscript, and this relates also to my comment in my review that the authors sometimes show already a preference for the 11-layer model. It is not carved in stone that a more detailed model is better, and it should objectively be assessed which one is better. Even though reviewer 1 points out that more detailed Richards' equation approaches often improve LSMs, there is also an important reason bucket-type models are still often used especially in catchment hydrology. The Richards' equation approach does not include macro-pores, which in more sloped areas plays an important role. In addition, the parameterization often assumes a homogeneous soil, which is also not true. The fact that LSMs often perform better with Richards' approach also relates to how they are parameterized, bucket-type models need actual calibration as the parameters are less physically based, whereas the Richards' approach uses more physically

based soil parameters that are often measured. In general, the hydrological schematization in LSMs is in my view still rather poor, even with more detailed Richards' equation approaches, whereas it actually has a strong influence on the outcomes of the models, so I believe it is important that the authors show this. In addition, for a strong modelling experiment, you always need a benchmark, which is here the 2-layer model. Leaving it out leads to a manuscript that is just a model application, and the reader can never see what the 11-layers actually add.

We thank the reviewer for outlining further reasoning for keeping the 2 vs 11 layer comparison. To address both reviewers concerns/suggestions on this matter, 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 do completely agree that it is not necessarily the case that a more complex model is needed. We hope that by addressing the reviewer's original comments and suggestions (including adding soil moisture observations to figure 2) that we have made our case for why we think the 11LAY does a better job at capturing the temporal dynamics of the upper layer (root zone) soil moisture and evapotranspiration. We have tried explicitly not to go beyond that specific conclusion regarding any preference for the 11 layer. We also hope that it is clear from our analyses on remaining model discrepancies, as well as other topics that we have highlighted in the discussion, that we agree that there are still many issues (missing or inadequately represented processes) in the more mechanistic versions included in LSMs that still need to be addressed. For example we have mentioned, as the reviewer discussed above, the fact that soil texture and most hydraulic parameters are fixed both vertically in the soil column and that spatial heterogeneity is not well captured. We also mention the need for parameter calibration. We did not include an exhaustive list of all the LSM hydrology model issues simply because these comparisons are still point-based, whereas many of the issues that remain are related to modeling spatially distributed hydrological budgets, which is beyond the scope of our present study.

I hope my thoughts are useful, even though it is probably not making it easier. I still look forward to an improved manuscript and hope the authors find a good way to address all the issues of myself and reviewer 1.

We sincerely thank the reviewer for their second round of feedback. It was certainly both insightful and useful in making a decision on this issue.



**Deleted:** Multi-variable, multi-configuration testing of ORCHIDEE land surface model water flux and storage estimates across semi-arid sites in the southwestern US

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

5

Natasha MacBean<sup>1\*</sup>, Russell L. Scott<sup>2</sup>, Joel A. Biederman<sup>2</sup>, Catherine Ottlé<sup>3</sup>, Nicolas Vuichard<sup>3</sup>, Agnès Ducharne<sup>4</sup>, Thomas Kolb<sup>5</sup>, Sabina Dore<sup>6</sup>, Marcy Litvak<sup>7</sup>, David J.P. Moore<sup>8</sup>.

<sup>1</sup>Department of Geography, Indiana University, Bloomington, IN 47405, USA.

<sup>2</sup>Southwest Watershed Research Center, United States Agricultural Department, Agricultural Research Service, Tucson, AZ 85719, USA.

<sup>3</sup>Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, F-91191, France.

<sup>4</sup>UMR METIS, Sorbonne Université, CNRS, EPHE, Paris, F-75005, France

<sup>5</sup>School of Forestry, Northern Arizona University, Flagstaff, AZ, 86011, USA.

<sup>6</sup>Hydrofocus, Inc., Davis, CA, 95618, USA.

<sup>7</sup>Department of Biology, University of New Mexico, Albuquerque, NM, 87131, USA.

<sup>8</sup>School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, 85721, USA.

\*Correspondence to: Natasha MacBean ([nmacbean@gmail.com](mailto:nmacbean@gmail.com))

20

### Abstract

Plant activity in semi-arid ecosystems is largely controlled by pulses of precipitation, making them particularly vulnerable to increased aridity expected with climate change. Simple bucket-model hydrology schemes in land surface models (LSMs) have had limited ability in accurately capturing semi-arid water stores and fluxes. Recent, more complex, LSM hydrology models have not been widely evaluated against semi-arid ecosystem *in situ* data. We hypothesize that the failure of older LSM versions to represent evapotranspiration, ET, in arid lands is because simple bucket models do not capture realistic fluctuations in upper layer soil moisture. We therefore predict that including a discretized soil hydrology scheme based on a mechanistic description of moisture diffusion will result in an improvement in model ET when compared to data because the temporal variability of upper layer soil moisture content better corresponds to that of precipitation inputs. To test this prediction, we compared ORCHIDEE LSM simulations from 1) a simple conceptual 2-layer bucket scheme with fixed hydraulic parameters; and 2) a 11-layer discretized mechanistic scheme of moisture diffusion in unsaturated soil based on Richards equations, against daily and monthly soil moisture and ET observations, together with data-derived estimates of transpiration/evapotranspiration, T/ET, ratios, from six semi-arid grass, shrub and forest sites in the southwestern USA. The 11-layer scheme also has modified

**Deleted:** hydrological

**Deleted:**

**Deleted:** ration

40 calculations of surface runoff, resistance to bare soil evaporation, E, and water limitation to be compatible with the more complex hydrology configuration. To diagnose remaining discrepancies in the 11-layer model, we tested two further configurations: i) the addition of a term that captures bare soil evaporation resistance to dry soil; and ii) reduced bare soil fractional vegetation cover. We found that the more mechanistic 11-layer model results in a better representation of the daily and monthly ET observations. We show that is likely because of improved simulation of soil moisture in the upper layers of soil (top 5cm). Some discrepancies between observed and modelled soil moisture and ET may allow us to prioritize future model development and the collection of additional data. 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. Although ET is generally well captured by the 11-lay model, modelled T/ET ratios were generally lower than estimated values across all sites, particularly during the monsoon season. Adding a soil resistance term generally decreased simulated bare soil evaporation, E, and increased soil moisture content, thus increasing transpiration, T, and reducing the negative bias between modelled and estimated monsoon T/ET ratios. This negative bias could also be accounted for at the low elevation sites by decreasing the model bare soil fraction, thus increasing the amount of transpiring leaf area. However, adding the bare soil resistance term and decreasing the bare soil fraction both degraded the model fit to ET observations. Furthermore, remaining discrepancies in the timing of the transition from minimum T/ET ratios during the hot, dry May-June period to high values at the start of the monsoon in July-August may also point towards incorrect modelling of leaf phenology and vegetation growth in response to monsoon rains. We conclude that a discretized soil hydrology scheme and associated developments improves estimates of ET by allowing the modelled upper layer soil moisture to more closely match the pulse precipitation dynamics of these semi-arid ecosystems; however, the partitioning of T from E is not solved by this modification alone.

Deleted:

Deleted: and T/ET ratios

Deleted: model-data bias

Deleted: By reducing the bare soil fraction in the model, we illustrated that modelled leaf T is too low at sparsely vegetated sites.

Deleted: bare soil evaporation

## 60 1 Introduction

Semi-arid ecosystems – which cover ~40% of the Earth's terrestrial surface and include rangelands, shrublands, grasslands, savannas, and seasonally dry forests – are in zones of transition between humid and arid climates and are characterized by sparse, patchy vegetation cover and limited water availability. Moisture availability in these ecosystems is therefore a major control on the complex interactions between vegetation dynamics and surface energy, water, and carbon exchange (Biederman et al., 2017; Haverd et al., 2016). Given the sensitivity to water availability, semi-arid ecosystem functioning may be particularly vulnerable to projected changes in climate (Tietjen et al., 2009; Maestre et al., 2012; Gremer et al., 2015). IPCC earth system model (ESM) projections and observation-based datasets indicate these regions will likely experience more intense warming and droughts, increases in extreme rainfall events, and a greater contrast between wet and dry seasons in the future (IPCC, 2013; Donat et al., 2016; Sippel et al., 2017; Huang et al., 2017).

Deleted: E

Deleted: S

Deleted: M

70 To simulate the impact of climate change on semi-arid ecosystem functioning, it is essential that the land surface model (LSM) component of ESMs accurately represent semi-arid water flux and storage budgets (and all associated processes). In the last

two to three decades, LSM groups have progressively updated their hydrology schemes from the more simplistic “bucket” type models included in earlier versions (Manabe, 1969). The resulting schemes typically include more physically-based representations of vertical diffusion of water in unsaturated soils (Clark et al., 2015). In addition to increasing the complexity of soil hydrology, several studies have attempted to address the issue that models tend to miscalculate partitioning of evapotranspiration (ET) into transpiration (T) and bare soil evaporation (E), with models systematically underestimating T/ET ratios (Wei et al., 2017; Chang et al., 2018). One such mechanism that models have introduced is an evaporation resistance term that reduces the rate of water evaporation from bare soil surfaces (Swenson and Lawrence, 2014; Decker et al., 2017). The development of these more mechanistic soil hydrology schemes should mean that LSMs better capture high temporal frequency to seasonal and long-term temporal variability of water stores and fluxes. However, it is not always apparent that increasing model complexity provides more accurate representations of reality (as encapsulated by observations of different variables at multiple spatio-temporal scales). Further, increasing model complexity comes at a cost of increased computational resources and unknown parameters. Therefore, it is imperative that we test models of increasing complexity against multiple types of observations at a variety of sites.

New generation LSM water flux and storage estimates have been extensively tested at multiple scales from the site level to the globe (Abramowitz et al., 2008; Dirmeyer et al., 2011; Guimberteau et al., 2014; Mueller et al., 2014; Best et al., 2015; Ukkola et al., 2016b; Raoult et al., 2018; Scanlon et al., 2018, 2019). Model-data biases are observed across all biomes; however, a key finding common to these studies is that models do not capture seasonal to inter-annual water stores and fluxes well during dry periods and/or at drier sites (Mueller et al., 2014; Swenson et al., 2014; De Kauwe et al., 2015; Best et al., 2015; Ukkola et al., 2016a; Humphrey et al., 2018; Scanlon et al., 2019). Mueller et al. (2014) showed that CMIP5 models overestimated multiyear mean daily ET in many regions, with the strongest bias in dryland regions (particularly western North America). Likewise, Grippa et al. (2011) and Scanlon et al. (2019) demonstrated that LSMs underestimate seasonal amplitude of total water storage in semi-arid (and tropical) regions. However, compared to more mesic ecosystems, semi-arid ecosystem LSM water flux and storage simulations have rarely been tested extensively against *in situ* observations, apart from a few exceptions (Hogue et al., 2005; Abramowitz et al., 2008; Whitley et al., 2016; Grippa et al., 2017). Whitley et al. (2016) compared carbon and water flux simulations from six LSMs at five OzFlux savanna sites. Their study highlighted two key deficiencies in modeling water fluxes: i) modeled C4 grass T is too low; and ii) models with shallow rooting depths typically underestimate woody plant dry season ET. As part of a model inter-comparison for West Africa (the AMMA LSM Intercomparison Project – ALMIP), LSM water storage, fluxes, runoff, and land surface temperature were evaluated against *in situ* and remote sensing data in the Malian Gourma region of the central Sahel (Boone et al., 2009; De Kauwe et al., 2013; Lohou et al., 2014; Grippa et al., 2011; Grippa et al., 2017). These studies highlight that temporal characteristics of water storage and fluxes in this monsoon-driven semi-arid region are captured fairly well by models; however, the studies also point to various model issues, including: difficulties in simulating bare soil evaporation response to rainfall events (Lohou et al., 2014); underestimation of dry season ET (Grippa et al., 2011); the need for greater water and energy exchange sensitivity to different vegetation types and soil characteristics (De Kauwe et al., 2013; Lohou et al., 2014; Grippa et al., 2017); and overestimation of surface runoff

Deleted: been

(Grippa et al., 2017). How models prescribe or predict leaf area index (LAI) has also been highlighted as a driver of hydrological model-data differences (Ha et al., 2015; Grippa et al., 2017).

The aim of this study was to contribute a new LSM hydrology model evaluation in a semi-arid region not previously investigated: the monsoon-driven semi-arid southwestern United States (hereafter, the SW US). The density and diversity of ecosystem research sites in the SW US provides a rare opportunity to test an LSM across a range of semi-arid ecosystems. The semi-arid SW US has also been identified as one of the key regions of global land-atmosphere coupling (Koster et al., 2004) and the most persistent climate change hotspot in the US (Diffenbaugh et al., 2008; Allen et al., 2016). Expected future soil moisture deficits in this region will result in strong atmospheric feedbacks, with consequent high temperature increases (Senerivatne et al., 2013) and a potential weakening of the terrestrial biosphere C sink (Berg et al., 2016; Green et al., 2019). Several studies based on model predictions, instrumental records, and paleoclimatic data analyses have suggested that over the coming century the risk of more severe, multi-decadal drought in the SW US will increase considerably (Ault et al., 2014, 2016; Cook et al., 2015). In fact, models suggest that a transition to drier conditions is already underway (Seager et al., 2007; Archer and Predick, 2008; Seager and Vecchi, 2010). Investigating how well LSMs capture hydrological stores and fluxes in this region therefore provides a crucial test for how well models can produce accurate global climate change projections.

Here, we tested the ability of the ORCHIDEE (ORganizing Carbon and Hydrology in Dynamic EcosystEms) LSM to simulate multiple water flux and storage related variables at six SW US semi-arid Ameriflux eddy covariance sites spanning forest, shrub- and grass-dominated ecosystems (Biederman et al., 2017). We tested two versions of the ORCHIDEE LSM with hydrological schemes of differing complexity: 1) a simple 2-layer conceptual bucket scheme (hereafter, 2LAY) with constant water holding capacity (de Rosnay and Polcher 1998); and 2) a 11-layer mechanistic scheme (hereafter, 11LAY) based on the

Richards equation, with hydraulic parameters based on soil texture (de Rosnay et al., 2002). Besides the change in the soil hydrology between the 2LAY and 11LAY versions, several other hydrology-related processes have also been modified due to increases in the complexity of the model. In the 2LAY scheme, runoff occurred when the soil reached saturation; whereas in the 11LAY scheme, surface runoff and drainage are treated more mechanistically (see Section 2.2.2). In the 2LAY scheme, there was an implicit resistance to bare soil evaporation based on the depth of the dry soil for the bare soil plant functional type (PFT). In the 11LAY scheme, there is an optional bare soil evaporation resistance term based on the relative soil water content of the first four soil layers, based on the formulation of Sellers et al. (1992) – (see Section 2.2.3). Both resistance terms aim to describe the resistance to evaporation exerted by a dry mulch soil layer. Similarly, the calculation of moisture limitation on stomatal conductance has changed. In the 2LAY version, moisture limitation depended on the dry soil depth of the upper layer; whereas in the 11LAY version, the limitation is based on plant water availability for root water uptake throughout the soil column. Finally, in the 2LAY scheme there is no E from the vegetated portion of the grid cell (only T); whereas, in the 11LAY scheme both E and T occur (see Section 2.2.1). The 2LAY scheme was used in the previous CMIP5 runs, whereas the 11LAY scheme is the default scheme in the current version of ORCHIDEE that is used in the ongoing Coupled Model Intercomparison Project (CMIP6) simulations (Ducharme et al., in prep). Although there have been many previous studies comparing simple bucket schemes versus mechanistic multi-layer hydrology, we include such a comparison in the first part of our analysis for

**Deleted:** combination of

**Deleted:** and Darcy's

**Deleted:** s

**Deleted:** that are still being used to understand ecosystem responses to changes in climate

155 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.

160 In testing the ORCHIDEE against *in situ* semi-arid water stores and fluxes, a novel component of our study was to investigate whether some of the site-scale semi-arid LSM hydrology model discrepancies outlined above (e.g. underestimation of C4 grass T, weak dry season ET and therefore low T/ET ratios, ET issues related to incorrect representation of leaf area, and overestimation of surface runoff) are improved with recent ORCHIDEE model developments. Where the model does not capture observed patterns, we attempted to investigate which model processes and parameterizations are responsible for model-

165 data discrepancies. First, we evaluated how changing from the conceptual 2LAY bucket model to the physically-based 11LAY soil hydrology scheme – and all associated modifications – has influenced the high temporal frequency and seasonal variability of semi-arid ecosystem soil moisture, ET (and its component fluxes), runoff, drainage, and snow mass/melt. Second, we evaluated the temporal dynamics of the 11LAY model against observations at three specific soil depths (shallow:  $\leq 5$ cm; mid: 15-20cm; deep:  $\geq 30$ cm) to assess whether the physically-based discretized scheme accurately captures moisture transport

170 down the soil profile. Note that when evaluating the 11LAY model soil moisture against observations, our primary focus was on the temporal dynamics – rather than the absolute magnitude – given the difficulty of comparing absolute values of volumetric water content between the models and the data (see Section 2.3.2 for more detail). Therefore, in the model-data comparison, we scale the observations to the 11LAY model simulations via linear CDF matching. Finally, having evaluated the standard (default) 11LAY model, we investigated which model processes or mechanisms in the 11LAY scheme might be

175 responsible for remaining model-data discrepancies in the water stores and fluxes. In particular, we assessed the impact of a) decreasing the bare soil fraction (thus, increasing leaf area); and b) including the optional bare soil resistance term into the 11LAY scheme (Ducharme et al., in prep.). Given the sparsely-vegetated nature of the low-elevation semi-arid grass- and shrub-dominated sites in our study it is possible that inclusion of this term may counter any dry season ET underestimate. Throughout, we explored if there are any discernible differences across sites due to elevation and vegetation composition.

180 Section 2 describes the sites, data, model and methods used in this study; Section 3 details the results of the two-part model evaluation (as outlined above); and Section 4 discusses how future studies may resolve remaining model issues in order to improve LSM hydrology modeling in semi-arid regions.

Deleted: (

Deleted: )

Deleted: seasonality

Deleted: s

Deleted: abovementioned hydrological variables

Deleted: , soil type, or

Deleted: d

Deleted: s

## 2 Methods and Data

### 2.1 Southwestern US study sites

185 We used six semi-arid sites in the SW US that spanned a range of vegetation types and elevations (Biederman et al. 2017). The entire SW US is within the North American Monsoon region; therefore, these sites typically experience monsoon rainfall

195 during July to October, preceded by a hot, dry period in May and June. Table 1 describes the dominant vegetation, species and soil texture characteristics at each site, together with the observation period. 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). These four sites are split into pairs of grass- and shrub-dominated systems: US-SRG and US-SRM are located at the Santa Rita Experimental Range ~60km south of Tucson, AZ, whilst US-Whs and US-Wkg are located at the Walnut Gulch Experimental Watershed ~120km to the southeast of Tucson, AZ. 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). The US-Fuf (Flagstaff Unmanaged Forest) and US-Vcp (Valles Caldera Ponderosa) sites are at higher elevations (2215m and 2501m). Both high elevation sites experience cooler mean annual temperatures of 7.1 and 5.7°C, respectively, and are dominated by 205 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. US-Fuf is located near the town of Flagstaff in northern AZ, whilst US-Vcp is located in the Valles Caldera National Preserve in the Jemez Mountains in north-central New Mexico. Groundwater depths across all sites are typically 10s to 100s metres. Flux tower instruments at all six sites collect half-hourly measurements of meteorological forcing data and eddy covariance measurements of net surface energy and carbon exchanges (see Section 2.3.1). 210

Deleted: a

Deleted: , respectively

## 2.2 ORCHIDEE Land Surface Model

### 2.2.1 General model description

The ORCHIDEE land surface model (LSM) forms the terrestrial component of the French IPSL ESM (Dufresne et al., 2013), 215 which contributes climate projections to IPCC Assessment Reports. ORCHIDEE has undergone significant modification since the 'AR5' version (Krinner et al., 2005), which was used to run the CMIP5 (Coupled Model Inter-comparison Project) simulations included in the IPCC 5<sup>th</sup> Assessment Report (IPCC, 2013). Here, we use ORCHIDEE v2.0 that is used in the ongoing CMIP6 simulations. The hydrology scheme in ORCHIDEE v2.0 is described in detail in Ducharne et al. (in prep.). ORCHIDEE simulates fluxes of carbon, water and energy between the atmosphere and land surface (and within the sub- 220 surface) on a half-hourly time step. In uncoupled mode, the model is forced with climatological fields derived either from climate reanalyses or site-based meteorological forcing data (2m air temperature, rainfall and snowfall, incoming long and shortwave radiation, wind speed, surface air pressure, and specific humidity). As in most LSMs, all vegetation is grouped into broad plant functional types (PFTs) based on physiology, phenology, and for trees, the biome in which they are located. In ORCHIDEE, by default, there are 12 vegetated PFTs plus a bare soil PFT. A prognostic leaf area is calculated based on 225 phenology schemes originally described in Botta et al. (2000) and further detailed in MacBean et al. (2015 – Appendix A).

Deleted: (Peylin et al., in prep).

230 The albedo is calculated based on the average of the defined albedo coefficients for vegetation (one coefficient per PFT), soil  
(one value for each grid cell, referred to as background albedo) and snow weighted by their fractional cover. Snow albedo is  
also parameterized according to its age, which varies according to the underlying PFT. The albedo coefficients for each PFT  
and background albedo have recently been optimized within a Bayesian inversion system using the visible and near infrared  
MODIS white sky albedo product at 0.5x0.5° resolution for years 2000-2010 (Bastrikov et al., in prep). The prior background  
(bare soil) albedo values were retrieved from the MODIS data using the EU Joint Research Center Two Stream Inversion  
235 Package (JRC-TIP).

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). There are two soil hydrology models implemented in ORCHIDEE: one based on a  
2-layer conceptual model, the other on a physically based representation of moisture redistribution across 11-layers. These two  
240 schemes were described and compared in the Amazon basin by Guimberteau et al. (2014), and are described further in the  
following sections. 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).

Independent water budgets are calculated for each "soil tile", which define separate water columns within a grid cell. In the 2-  
layer scheme, soil tiles correspond to PFTs; therefore, a separate water budget is calculated for each PFT within the grid cell.  
245 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. In the 2-layer scheme there is no E from the vegetated tiles (only transpiration). 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  
250 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  $k_{ext}$  is the extinction coefficient and is set to 1.0. The effective bare soil fraction  $f_b^j$  is the complement  
to  $f_v^j$ . The grid cell water budget is calculated by vegetation fraction weighted averaging across all soil tiles (Guimberteau et  
al., 2014). Soil texture classes and related parameters are prescribed based on the percentage of sand, clay and loam. The main  
differences between the two ORCHIDEE configurations used in this study are described in the sections below and are  
255 summarised in Table 2.

Deleted: , E,

Deleted: Separate

Deleted: )

### 2.2.2 Soil Hydrology

#### *2-layer conceptual soil hydrology model*

260 In the 'AR5' version of ORCHIDEE used in the CMIP5 experiments, the soil hydrology scheme consisted of a conceptual 2-  
layer (hereafter, 2LAY) model based on Choissnel et al. (1995). The depth of the upper layer is variable up to 10 cm and

265 changes with time depending on the balance between throughfall and snowmelt inputs, and outputs via three pathways: i) bare  
soil evaporation, limited by a soil resistance increasing with the dryness of the topmost soil layer; ii) root water extraction for  
transpiration, withdrawn from both layers proportionally to the root density profile; and iii) downward water flow (drainage)  
to the lower layer. If all moisture evaporates or if the entire soil saturates, the top layer can disappear entirely. Three empirical  
parameters govern the calculation of the drainage between the two layers, which depends on the water content of the upper  
layer and takes a non-linear form, so drainage from the upper layer increases considerably when the water content of the upper  
270 layer exceeds 75% of the maximum capacity (Ducharne et al., 1998). Transpiration is also withdrawn from the lower layer via  
water uptake by deep roots. Finally, runoff only occurs when the total soil water content exceeds the maximum field capacity,  
set to 150 kg.m<sup>-2</sup> as in Manabe (1969). It is then arbitrarily partitioned into 5% surface runoff to feed the overland flow and  
95% drainage to feed the groundwater flow of the routing scheme (Guimberteau et al., 2012b), which is not activated here.

#### 275 *11-layer mechanistic soil hydrology model*

The 11LAY scheme was initially proposed by de Rosnay et al. (2002) and simulates vertical flow and retention of water in  
unsaturated soils based on a physical description of moisture diffusion (Richards, 1931). The scheme implemented in  
ORCHIDEE relies on the one-dimensional Richards equation, combining the mass and momentum conservation equations,  
but is in its saturation form that uses volumetric water content  $\theta$  (m<sup>3</sup>m<sup>-3</sup>) as a state variable instead of pressure head (Ducharne  
280 et al., in prep). The two main hydraulic parameters (hydraulic conductivity and diffusivity) depend on volumetric soil moisture  
content defined by the Mualem–van Genuchten model (Mualem, 1976; van Genuchten, 1980). The Richards equation is solved  
numerically using a finite-difference method, which requires the vertical discretisation of the 2m soil column. As described by  
de Rosnay et al. (2002), 11 layers are defined: the top layer is ~0.1mm thick and the thickness of each layer increases  
geometrically with depth. The fine vertical resolution near the surface aims at capturing strong vertical soil moisture gradients  
285 in response to high temporal frequency (sub-diurnal to few days) changes in precipitation or ET. De Rosnay et al. (2000) tested  
a number of different vertical soil discretization and decided 11 layers was a good compromise between computational cost  
and accuracy in simulating vertical hydraulic gradients. The mechanistic representation of redistribution of moisture within  
the soil column also permits capillary rise, and a more mechanistic representation of surface runoff: The calculated soil  
hydraulic conductivity determines how much precipitation is partitioned between soil infiltration and runoff (d'Orgeval et al.,  
290 2008). Drainage is computed as free gravitational flow at the bottom of the soil (Guimberteau et al., 2014). The USDA soil  
texture classification, provided at 1/12-degree resolution by Reynolds et al. (2000), is combined with the look-up pedotransfer  
function tables of Carsel and Parrish (1988) to derive the required soil hydrodynamic properties (saturated hydraulic  
conductivity Ks, porosity, Van Genuchten parameters, residual moisture), while field capacity and wilting point are deduced  
from the soil hydrodynamic properties listed above and the Van Genuchten equation for matric potential, by assuming they  
295 correspond to potentials of -3.3m and -150m respectively (Ducharne et al., in prep). 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).

**Deleted:** at al.

**Deleted:** a one-dimensional differential equation, combining the one-dimensional Darcy equation, which describes the rate of flow of a fluid (soil moisture) within a permeable medium, with the mass conservation equation (the complete formulation is more generally called the saturation-based Richards equation).

**Deleted:** at al.



305 The 11LAY soil hydrology scheme has been implemented in the ORCHIDEE Trunk since 2010, albeit with various  
modifications since [that time](#), as described in this study. [The most up-to-date version of the model is described in Ducharme et  
al. \(in prep\)](#). Similar versions of the 11LAY scheme have been tested against a variety of hydrology-related observations in  
the Amazon Basin (Guimberteau et al., 2012a; Guimberteau et al., 2014), for predicting future changes in extreme runoff  
events (Guimberteau et al., 2013) and against a water storage and energy flux estimates as part of ALMIP in West Africa (as  
310 detailed in Section 1 – d’Orgeval et al., 2008; Boone et al., 2009; Grippa et al., 2011; Grippa et al., 2017).

Deleted: s

### 2.2.3 Bare soil evaporation and additional resistance term

The computation of bare soil evaporation,  $E_g$ , in both versions is implicitly based on a supply and demand scheme. In the  
2LAY version,  $E_g$  decreases when the upper layer gets drier, owing to a resistance term that depends on the height of the dry  
315 soil in the bare soil PFT column (Ducoudré et al., 1993). In the 11LAY version,  $E_g$  proceeds at the potential rate  $E_{pot}$  unless  
the water supply via upward diffusion from the water column is limiting, in which case  $E_g$  is reduced to correspond to the  
situation in which the soil moisture of the upper 4 layers is at wilting point. However, since ORCHIDEE v2.0 (Ducharme et  
al., in prep.),  $E_g$  can also be reduced by including an optional bare soil evaporation resistance term,  $r_{soil}$ , which depends on the  
relative water content and is based on a parameterization fitted at the FIFE grassland experimental site at Konza Prairie Field  
320 Station in Kansas (Sellers et al., 1992):

$$r_{soil} = \exp(8.206 - 4.255 W_1) \quad (1)$$

where  $W_1$  is the relative soil water content of the first four layers (2.2cm – Table S1).  $W_1$  is calculated by dividing the mean  
soil moisture across these layers by the saturated water content. The calculation for  $E_g$  then becomes:

$$E_g = \min(E_{pot}/(1 + r_{soil}/r_a), Q) \quad (2)$$

325 where  $E_{pot}$  is the potential evaporation,  $r_a$  the aerodynamic resistance,  $Q$  the upward water supply from capillary diffusion  
through the soil, and  $r_{soil}$  the soil resistance to this upward exfiltration. In all simulations, the calculation of  $r_a$  includes a  
dynamic roughness height with variable LAI, based on a parameterization by Su et al. (2001). By default, in the 11LAY version  
there is no resistance ( $r_{soil} = 0$ ). [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.](#)

330

### 2.2.4 Empirical plant water stress function, $\beta$

The soil moisture control on transpiration is defined by an empirical water stress function, called  $\beta$ . 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  
335  $c_j$  (in  $m^{-1}$ ) is [the root density decay factor for PFT  \$j\$ . In both model versions for a 2m soil profile,  \$c\_j\$  is set to 4.0 for grasses, 1.0](#)

Deleted: it

Deleted: the the

Deleted: F

for temperate needleleaved trees and 0.8 for temperate broadleaved trees. In 1LAY, a related variable is  $n_{root}(i)$ , quantifying the mean relative root density  $R(z)$  of each soil layer  $i$ , so that  $\sum n_{root}(i) = 1$ .

In the 2LAY version,  $\beta$  is calculated as an exponential function of the root decay factor  $c_j$  and the dry soil height of the topmost soil layer ( $h^d$ ):

$$\beta = \exp(-c_j h^d) \quad (3)$$

In the 1LAY,  $\beta$  is rather based on the available moisture across the entire soil moisture profile and is calculated for each PFT  $j$  and soil layer  $i$ , and then summed across all soil layers (starting at the 2<sup>nd</sup> layer given no water stress in the 1<sup>st</sup> layer – a conservative condition that prevents T from inducing a negative soil moisture from this very thin soil layer):

$$\beta(j) = \sum_{i=2}^{11} n_{root}(i) \max\left(0, \min\left(1, \max\left(0, \frac{W_{Lp} - W_{wpt}}{W_{\%} - W_{wpt}}\right)\right)\right) \quad (4)$$

where  $W_i$  is the soil moisture for that layer and soil tile in  $\text{kgm}^{-2}$ ,  $W_{wpt}$  is the wilting point soil moisture, and  $W_{\%}$  is the threshold above which T is maximum – i.e. above this threshold T is not limited by  $\beta$ .  $W_{\%}$  is defined by:

$$W_{\%} = W_{wpt} + p_{\%}(W_{fc} - W_{wpt}) \quad (5)$$

Where  $W_{fc}$  is the field capacity and  $p_{\%}$  defines the threshold above which T is maximum.  $p_{\%}$  is set to 0.8 and is constant for all PFTs. This empirical water stress function equation means that, in the 1LAY,  $\beta$  varies linearly between 0 at the wilting point to 1 at  $W_{\%}$ , which is smaller or equal to the field capacity. LSMs typically apply  $\beta$  to limit photosynthesis ( $A$ ) via the maximum carboxylation capacity parameter  $V_{\text{cmax}}$ , or to the stomatal conductance,  $g_s$ , via the  $g_0$  or  $g_1$  parameters of the  $A/g_s$  relationship, or both (De Kauwe et al., 2013; 2015). In ORCHIDEE there is the option of applying  $\beta$  to limit either  $V_{\text{cmax}}$  or  $g_s$ , or both. In the default configuration used in CMIP6,  $\beta$  is applied to both (based on results from Keenan et al., 2010; Zhou et al., 2013; Zhou et al., 2014); therefore, this is the configuration we used in this study.

## 2.2.5 Snow scheme

ORCHIDEE contains a multi-layer intermediate complexity snow scheme that is described in detail in Wang et al. (2013). The new scheme was introduced to overcome limitations of a single layer snow configuration. In a single layer scheme, the temperature and vertical density gradients through the snowpack, which affect the sensible, latent and radiative energy fluxes, are not calculated. The single layer snow scheme does not describe the insulating effect of the snow pack, nor the links between snow density and changes in snow albedo (due to aging) in a physically mechanistic way. In this explicit snow scheme, there are three layers that each have a specific thickness, density, temperature and liquid water and heat content: These variables are updated at each time step based on the snowfall and incoming surface energy fluxes, which are calculated from the surface energy balance equation. The model also accounts for sublimation, snow settling, water percolation and refreezing. Snow mass cannot exceed a threshold of  $3000 \text{ kg.m}^{-2}$ . Snow age is also calculated and is used to modify the snow albedo. Default snow albedo coefficients have been optimized using MODIS data per the method described in Section 2.2.1. Snow fraction is

Deleted: full

Deleted: v

Deleted: v is the PFT index (starting at 2 given the 1<sup>st</sup> PFT is bare soil), i is the index for each soil layer (starting at the 2<sup>nd</sup> layer given no water stress in the 1<sup>st</sup> layer – a conservative condition that prevents T from inducing a negative soil moisture from this very thin soil layer),

calculated at each time step according to snow mass and density following the parametrization proposed by Niu and Yang (2007).

380

## 2.3 Data

### 2.3.1 Site-level meteorological and eddy covariance data and processing

Meteorological forcing and eddy covariance flux data for each site were downloaded from the AmeriFlux data portal (<http://ameriflux.lbl.gov>). Meteorological forcing data included 2m air temperature and surface pressure, precipitation, incoming long and shortwave radiation, wind speed, and specific humidity. To run the ORCHIDEE model, we partitioned the *in situ* precipitation into rain and snowfall using a temperature threshold of 0°C. The meteorological forcing data were gap-filled following the approach of Vuichard and Papale (2015), which uses downscaled and corrected ERA-Interim data to fill gaps in the site-level data. Eddy covariance flux data were processed to provide ET from estimates of latent energy fluxes. ET gaps were filled using a modified look-up table approach based on Falge et al. (2001), with ET predicted from meteorological conditions within a 5-day moving window. Previous comparisons of annual sums of measured ET with site-level water balance measurements at a few of these sites show an average agreement within 3% of each other, but could differ by -10 to +17% in any given year (Scott and Biederman, 2019). 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 (uWUE<sub>a</sub>) to the potential uWUE (uWUE<sub>p</sub>). uWUE<sub>a</sub> is calculated based on a linear regression between ET and GPP.VPD<sup>0.5</sup> at observation timescales for a given site, whereas uWUE<sub>p</sub> was calculated based on a quantile regression between ET and GPP.VPD<sup>0.5</sup> 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 uWUE<sub>p</sub> is when T/ET = 1, which rarely occurs in water-limited environments (Scott and Biederman, 2017). T/ET ratio estimates are omitted in certain winter months when very low GPP and limited variability in GPP results in poor regression relationships.

395

400

405

Deleted: estimates

Deleted: who

Deleted: during the summer growing season at the non-forested sites using eddy covariance fluxes.

### 2.3.2 Soil moisture data and processing

Daily mean volumetric soil moisture content (VWC,  $\text{m}^3\text{m}^{-3}$ ) measurements at several depths were obtained directly from the site PIs. For each site, [Table 3](#) details the depths at which soil moisture was measured. Soil moisture measurement uncertainty is highly site and instrument specific, but tests have shown that average errors are generally below  $0.04 \text{ m}^3 \text{ m}^{-3}$  if site specific calibrations are made. [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. 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. However, several factors mean that we cannot directly compare absolute values of measured versus modelled soil VWC, even though the 11LAY has discrete depths. First, site-specific values for soil saturated and residual water content were generally not available to parameterize the model \(see Section 2.4\); instead, these soil hydrology parameters are either fixed \(in the 2LAY\) or derived from prescribed soil texture properties \(in the 11LAY – see Section 2.2.2\). Therefore, we may expect a bias between the modelled and observed daily mean volumetric soil water content \( \$\theta\$ \). Second, while the soil moisture measurements are made with probes at specific depths, it is not precisely known over which depth ranges they are measuring VWC. 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. Instead, we focus solely on comparison between the modelled and observed soil moisture temporal dynamics.](#) To achieve this, we removed any model-data bias using a linear cumulative density function (CDF) matching function to re-scale and match the mean and standard deviation of soil moisture [simulations](#) to that of the [observations](#) for each layer where soil moisture is measured following the equation:

$$\theta_{Mod,CDF} = \frac{\sigma_{\theta,Obs}(\theta_{Mod} - \theta_{Mod})}{\sigma_{\theta,Mod}} + \theta_{Obs} \quad (6)$$

Raoult et al. (2018) found that linear CDF matching performed nearly as well as full CDF matching in capturing the main features of the soil moisture distributions; therefore, for this study we chose to simply use a linear CDF re-scaling function. [Note that while we do compare the re-scaled 2LAY and 11LAY upper layer soil moisture \(top 10cm\) to the observations \(see Section 3.1\), we cannot compare the total column soil moisture given our observations do not go down to the same depth as the model \(2m\). Also note that we chose to focus most of the model-data comparison on investigating how well the \(re-scaled\) 11LAY model captures the observed temporal dynamics at specific soil depths \(see Section 3.2\).](#)

### 2.4 Simulation set-up and post-processing

All simulations were run for the period of available site data (including meteorological forcing and eddy covariance flux data – see Section 2.3.1 and Table 1). Table 1 also lists: i) the main species for each site and the fractional cover of each model PFT that corresponds to those species; iii) the maximum LAI for each PFT; and iii) the percent of each model soil texture class that

Deleted: Table 2

Deleted: W

Deleted: is

Deleted: instead of

Deleted: ing

Deleted: 11LAY

Deleted: , we specifically investigate how well the model captured the temporal dynamics at specific soil depths

Deleted: observations

Deleted: model

Deleted: 5

455 corresponds to descriptions of soil characteristics for each site – all of which were derived from associated site literature  
detailed in [the references in Table 1](#). The PFT fractional cover and the fraction of each soil texture class are defined in  
ORCHIDEE by the user. The maximum LAI has a default setting in ORCHIDEE that has not been used here; instead, values  
based on the site literature [were](#) prescribed in the model (Table 1). Note that ORCHIDEE does not contain a PFT that  
specifically corresponds to shrub vegetation; therefore, the shrub cover fraction was prescribed to the forested PFTs (see Table  
1). [Due to the lack of available data on site-specific soil hydraulic parameters across the sites studied, we chose to use the  
460 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 also allows us to test the default behavior of the model.](#)  
At each site we ran five versions of the model: 1) 2LAY soil hydrology; 2) 11LAY soil hydrology with  $r_{soil}$  flag not set ([default  
model configuration](#)); 3) 11LAY soil hydrology with  $r_{soil}$  flag not set and with reduced bare soil fraction (increased C4 grass  
cover); 4) 11LAY soil hydrology with the  $r_{soil}$  flag set (therefore, Eqn. 2 activated); and 5) 11LAY soil hydrology with the  $r_{soil}$   
465 flag set and with reduced bare soil fraction. [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 was prescribed in the model based on average fractional cover values  
given in the site literature.](#) A 400-year spinup was performed by cycling over the gap-filled forcing data for each site (see Table  
1 for period of available site data) to ensure the water stores were at equilibrium. Following the spinup, transient simulations  
470 were run using the forcing data from each site. Daily outputs of all hydrological variables (soil moisture, ET and its component  
fluxes, snow pack, snow melt), the empirical water stress function,  $\beta$ , and LAI were saved for all years and summed or averaged  
to derive monthly values, where needed. For certain figures we show the 2009 daily time series because 2009 was the only  
year for which data from all sites overlapped and had a complete year of daily soil moisture observations. [To evaluate the two  
model configurations, we calculated the Pearson correlation coefficient between the simulated and observed daily time series  
475 for both the upper layer soil moisture \(with the model re-scaled according to the linear CDF matching method given in Section  
2.3.2\) and ET. We also calculated the RMSE, mean absolute bias, and a measure of the relative variability,  \$\alpha\$ , between the  
modelled and observed daily ET. The latter is calculated as the ratio of model to observed standard deviations \( \$\alpha = \frac{\sigma\_m}{\sigma\_o}\$ \) based  
on \[Gupta et al. \\(2009\\)\]\(#\).](#)

Deleted: s

Deleted: have been

### 3 Results

#### 480 3.1 Differences between the 2LAY and 11LAY model versions for main hydrological stores and fluxes

Increasing the soil hydrology model complexity between the 2LAY and 11LAY model versions does not result in a uniform  
increase or decrease across sites in either the simulated upper layer (top 10cm) and total column (2m) soil moisture ( $\text{kgm}^{-2}$ )  
(Fig. 1 2<sup>nd</sup> and 3<sup>rd</sup> panel; also see Fig. S1 for complete daily time series for each site). The largest change between the 2LAY  
and 11LAY versions in the upper layer soil moisture [were](#) seen at the high-elevation ponderosa forest sites (US-Fuf and US-

Deleted: amount

Deleted: a

Vcp – Fig. 1 and Figs. S1 a and b). 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. At US-Fuf, both the upper layer and total column soil moisture increase in the 11LAY simulations compared to the 2LAY, which corresponds to an increase in mean daily ET (Fig. 1 top panel) away from the observed mean, and a decrease in total runoff (surface runoff plus drainage – Fig. 1 bottom panel). In contrast, at US-Vcp while there is an increase in the upper layer soil moisture, which results in a slight increase in mean ET (and ET variability) that better matches the observed mean daily ET, and a decrease in total runoff, there is hardly any change in the total column soil moisture. Note that changes in maximum soil water holding capacity are due to how soil hydrology parameters are defined. In the 2LAY, a maximum capacity is set to 150kgm<sup>-2</sup> across all PFTs; whereas in the 11LAY, the capacity is based on soil texture properties and is therefore different for each site.

At the low-elevation shrub and grass sites the differences between the two model versions for both the upper layer and total column soil moisture are much smaller (Fig. 1). Correspondingly, the changes in mean daily ET and total runoff are also marginal (although the mean total runoff is lower at Walnut Gulch: US-Wkg and US-Whs). Across all sites both model versions accurately capture the overall mean daily ET (Fig. 1). At Santa Rita (US-SRM and US-SRG), the 11LAY soil moisture is marginally lower than the 2LAY, whereas at the Walnut Gulch sites the 11LAY moisture is higher.

As described above, at all sites there is either no change between the 2LAY and 11LAY simulations (Santa Rita) or a decrease in total runoff (surface runoff plus drainage – Fig. 1 bottom panel). Across all sites, excess water is removed as drainage in the 2LAY simulations, with little to no runoff (Figs. S1 a-f 3<sup>rd</sup> panel); whereas in the 11LAY simulations excess water flows mostly as surface runoff, with more limited drainage (Figs. S1 a-f 2<sup>nd</sup> panel). This is explained by the fact that in the 2LAY scheme, the drainage is always set to 95% of the soil excess water (above saturation) and runoff can appear only when the total 2m soil is saturated. However, the 11LAY scheme also accounts for runoff that exceeds the infiltration capacity, which depends on the hydraulic conductivity function of soil moisture (Horton runoff). This means that when the soil is dry, the conductivity is low and more runoff will be generated. In the 11LAY simulations, the temporal variability in total runoff (as represented by the error bars in Fig. 1) has also decreased. As just described, in the 11LAY the total runoff mostly corresponds to surface runoff (Figs. S1 a-f). The lower drainage flux in the 11LAY simulations corresponds well to the calculated water balance at US-SRM (Scott and Biederman, 2019). 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). 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).

Across all sites, the magnitude of the total column VWC temporal variability (represented by the error bars in Fig. 1 3<sup>rd</sup> panel) only increases slightly between the 2LAY and 11LAY model versions. In the upper layer (top 10cm), the VWC temporal variability again onl increases marginally between the 2LAY and 11LAY for the high-elevation forest sites (Fig. 1 2<sup>nd</sup> panel error bars); however, the magnitude of variability decreases considerably in the 11LAY model for the low-elevation shrub and grass sites (also see Fig. S2). At all sites, the 2LAY upper layer soil moisture simulations fluctuate considerably between field

**Deleted:** higher

**Deleted:** at the ponderosa sites

**Deleted:** is reduced

**Deleted:** between the 2LAY and 11LAY simulations

**Deleted:** error bars

**Deleted:** ; although, the complete time series in Fig. S1 show that the 11LAY simulations appear to respond more dynamically to rain events...

**Deleted:** Whereas

**Deleted:** at all sites

capacity and zero throughout the year, including during dry periods with no rain. In the 11LAY however, the temporal dynamics of the 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). This results in much better fit of the 11LAY model to the temporal variability seen in the observations (Figs. 2 and S2). This improvement in upper layer soil moisture temporal dynamics is also indicated by the strong increase in correlation at all sites between the re-scaled modelled and observed 11LAY upper layer soil moisture compared to the 2LAY (increases in R ranged from 0.1 to 0.48 – Table 4). Note that not only is the upper high frequency temporal variability therefore arguably more realistic in the 11LAY version, the finer scale discretization of the uppermost soil layer in this version will also allow a much easier comparison with satellite-derived soil moisture products that can only “sense” the upper few cm of the soil (Raoult et al., 2018).

Deleted: –  
Deleted: –  
Deleted: 11LAY

A major and important consequence of the changes in the upper layer soil moisture temporal dynamics is a considerable improvement across all sites in the 11LAY simulated daily ET (Fig. 2 2<sup>nd</sup> panel, which shows 2009 for three sites; Figs. S2 a-f shows the complete time series for all sites). Across all sites, the 11LAY RMSE between daily modelled and observed ET has decreased in comparison to the 2LAY and the correlation has increased by a fraction of 0.3 to 0.4 (Table 4). With the exception of US-Vcp, the mean absolute daily ET model-data bias has increased slightly between the 2LAY and 11LAY versions (Table 4), which is due to the fact that the 2LAY version both underestimates and overestimates ET in the spring and summer respectively, resulting in a smaller mean absolute bias (Fig. S3). However, the 11LAY model only slightly underestimates mean daily ET at most sites, except at US-Fuf. In both model versions, the biases correspond to less than 10% of the mean daily ET across all low elevation sites. At the high elevation sites, the 11LAY bias corresponds to ~20% of the mean daily ET – an increase (decrease) compared to the 2LAY at US-Fuf (US-Vcp). The ratio of modelled to observed standard deviation in ET,  $\alpha$ , is also provided as a measure of relative variability in the simulated and observed values (Table 4). With the exception of US-Fuf,  $\alpha$  values tend closer to 1.0 in the 11LAY simulations compared to the 2LAY – highlighting again that the 11LAY version does a better job of capturing the daily variability. The higher ET model-data bias and  $\alpha$  at US-Fuf is mostly due to model discrepancies in spring (Fig. S2a), which we discuss further in Section 3.3. As previously discussed, the increase in 11LAY model upper layer moisture content at the high-elevation forest sites (Fig. 1 2<sup>nd</sup> panel and Fig. 2 bottom panel) have resulted in an increase in E and T at those sites, which in turn results in a lower ET RMSE between the model and the observations (Table 4, and see Figs. 2 and S2 2<sup>nd</sup> panel) if not a decrease in the mean ET bias for US-Fuf (Table 4 and Fig. 1). 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,  $\beta$  (Figs. 2 and S2 5<sup>th</sup> panel), which acts to limit both photosynthesis and stomatal conductance (therefore, T) during periods of moisture stress (Section 2.2.4). With the new calculation in the 11LAY version (see Section 2.2.4), we see a stronger, more rapid decrease in  $\beta$  (increased stress) during warm, dry periods that correspond to strong reductions in T (brown shaded zones in Fig. 2). Aside from T and E, the other ET components (interception and sublimation) did not change much between the two hydrology schemes (results not shown); therefore, these terms are not contributing to improvements between the 2LAY and 11LAY versions.

Deleted: variability  
Deleted: high frequency temporal dynamics  
Moved (insertion) [1]  
Deleted: Table 3  
Deleted: model-observed  
Deleted: Table 3  
Deleted: However, the mean absolute biases are only a small fraction of the daily ET.  
Deleted: T  
Deleted: only

Deleted:  
Deleted: better match  
Deleted: latest  
Deleted: are  
Deleted: ing

The improvement in daily ET temporal dynamics results in an 11LAY mean monthly ET that is also well captured by the model throughout the year, including both the warm, dry May-June period followed by monsoon summer rains, particularly for low-elevation grass and shrub sites (Fig. 3 and Fig. S3). As previously discussed, the improved, higher monthly ET in the 11LAY version during the period of maximum productivity (i.e. the spring and summer for the high-elevation sites, and the summer monsoon for the low-elevation sites – Fig. 3) is likely due to the increase in plant available water (Fig. 1 – 2<sup>nd</sup> and 3<sup>rd</sup> panels and Fig. S1). Despite the improvement in the 11LAY temporal variability at the high-elevation forest sites, there is still a bias in the mean monthly ET magnitude between the 11LAY model and observations: At US-Fuf there is a distinct overestimation of ET during the spring (Fig. S3a), whereas at US-Vcp there is a noticeable underestimation of ET during the spring and monsoon periods (Fig. S3b). We will return to these remaining 11LAY ET model-data discrepancies in Section 3.3 after having evaluated the 11LAY soil moisture against observations at different depths.

### 3.2 Comparison of 11LAY soil moisture against observations at different depths

Fig. 4 compares model versus observed daily volumetric soil water content time-series for 2009 at three different depths (see Fig. S4 for full time series at each site). The complete model time series were re-scaled via linear CDF matching to remove model-observation biases (see Section 2.3.2); however, the linear CDF matching preserves the mean and standard deviation of the temporal variability. As seen in Section 3.1 and Fig. 2 (bottom panels showing upper 10cm soil moisture), in Fig. 4 the high frequency temporal variability of the 11LAY soil moisture in the uppermost layer almost perfectly matches the observed, particularly at the low-elevation shrub- and grass-dominated sites (US-SRM, US-SRG, US-Whs, US-Wkg). At most of the low-elevation sites the soil moisture drying rates in the upper 20cm of soil are well captured by the model, with the small exception of the Santa Rita sites between January to March in which the model appears to dry down at a faster rate than observed (Fig. 4).

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 snow pack 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 soil temperature in the model is below 0°C (Fig. 5 bottom panel), suggesting that in the field 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

**Moved up [1]:** Across all sites, the 11 LAY RMSE between daily modelled and observed ET has decreased in comparison to the 2LAY and the correlation has increased by a fraction of 0.3 to 0.4 (Table 3). With the exception of US-Vcp, the mean absolute daily model-observed ET bias has increased slightly between the 2LAY and 11LAY versions (Table 3), which is due to the fact that the 2LAY version both underestimates and overestimates ET in the spring and summer respectively (Fig. S3). However, the mean absolute biases are only a small fraction of the daily ET. The 11LAY model only slightly underestimates mean daily ET at most sites, except at US-Fuf.

**Deleted:** The improvement in ET at high frequency timescales gives rise to a dramatic reduction in model-data misfit at monthly to seasonal timescales, particularly for low-elevation grass and shrub sites. Across all sites, the 11 LAY RMSE between daily modelled and observed ET has decreased in comparison to the 2LAY and the correlation has increased by a fraction of 0.3 to 0.4 (Table 3). With the exception of US-Vcp, the mean absolute daily model-observed ET bias has increased slightly between the 2LAY and 11LAY versions (Table 3), which is due to the fact that the 2LAY version both underestimates and overestimates ET in the spring and summer respectively (Fig. S3). However, the mean absolute biases are only a small fraction of the daily ET. The 11LAY model only slightly underestimates mean daily ET at most sites, except at US-Fuf. ¶

**Deleted:** mean

**Deleted:** T

**Deleted:** whole

**Deleted:** (particularly US-Vcp).

**Deleted:** even when comparing with the re-scaled observations, at both sites

**Deleted:** is large

**Deleted:** either

**Deleted:** or

**Deleted:** of peak snow fall



650 sites because it assumes 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.

655 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 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 a factor of ten and reran 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 the positive model-data bias in the early winter 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.

670 Overall, there is a decrease in the model ability to capture both high frequency and seasonal variability with increasing soil depth. At all sites the temporal dynamics of the deepest observations are not well represented in the model (Fig. 4 bottom panels for each site). At the high-elevation forest sites (US-Fuf and US-Vcp), the model does not capture the response of observed soil moisture in the deepest layer to summer storm events. In contrast, at the low-elevation shrub and grass sites the 11LAY VWC is far too dynamic in the deepest layer. 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. The poor model-data fit at lower depths may be related to the discretization of the soil column with a geometric increase of internode distance. Therefore, the soil layer thicknesses increase substantially beyond the ~2-4cm (7<sup>th</sup>

Deleted: Similarly, after

Deleted: the main period of snowfall

Deleted: again melts and sublimates

Deleted: snow

Deleted: ; o

Deleted: . H

Deleted: Although snow depth, mass or fractional cover data have not been collected at these sites, snow typically remains on the ground until late spring after winters with heavy snowfall (T. Kolb pers. comm.) suggesting that the continued existence of a snow pack and snow melt replenishes soil moisture until late spring when all the snow melts. Therefore, it is important to note that the rain gauges almost certainly underestimate the actual snowfall amount (Rasmussen et al., 2012; Chubb et al., 2015) and/or that the method of assigning precipitation data as snowfall when  $T < 0^{\circ}\text{C}$  (Section 2.3.1) results in inaccurate estimates of the amount or duration of snow cover.

Deleted: T

Deleted: ,

Deleted: ing

Deleted: be the main factor

Deleted: high-elevation

Deleted: model-data misfits

Deleted: at

Deleted: which increases

and 8<sup>th</sup> soil layers – Table S1). For the deeper soil moisture observations, it is therefore harder to match the depth of the observations with a specific soil layer. 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 lateral 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. In contrast to temperate trees, semi-arid trees and shrubs often have deep taproots for accessing groundwater. 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).

Deleted:

Deleted: changes with depth

Deleted: are

Deleted: ing

Deleted: Ducharme et al., in prep.).

### 3.3 Remaining discrepancies in ET and its component fluxes

Despite the improvement in seasonal ET temporal dynamics in the 11LAY model, particularly the timing of the reduction during the dry season, key model-data discrepancies in ET remain during spring (March–April) and monsoon (July–September) periods: i) At US-Fuf, the 11LAY observed ET is overestimated during the spring and early summer (Fig. S3a); ii) At US-Vcp, the model underestimates ET for much of the growing season, likely due to low LAI values in the earlier and later years of the simulation (Fig. S3b); iii) at US-SRM the 11LAY model overestimates springtime ET (in contrast to other low-elevation monsoon sites) (Fig. S3c); and iv) the 11LAY model still slightly underestimates peak monsoon ET at the low-elevation shrub sites (US-SRM and US-Whs– Figs. S3 c-d) as seen in a previous semi-arid model evaluation study (Grippa et al., 2011).

The model overestimate in spring ET at US-Fuf could be related to the snowfall issues that are causing the model to underestimate spring soil moisture during the same period (Figs. 4 and 5 and see Section 3.2). The lack of a persistent snow pack in the model during this period can explain the positive bias in spring ET because in reality the presence of snow would suppress bare soil evaporation. As discussed in Section 3.2, to accurately diagnose this issue we would need further information on snow mass or depth. Further support for the suggestion that modelled spring E is overestimated comes from comparing the model with 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 upper layer soil moisture (as previously described in Section 3.1 and Figs. 2 and S2a and b), 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 US-Fuf 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 explain why the model T/ET ratios do not increase as rapidly as estimated values at the start of the monsoon (Fig. 6). However, discrepancies in the timing of T/ET ratio peak and troughs between the model and data-derived

Deleted: Given the model tends to

Deleted: and US-Vcp

Deleted: , this bias is likely related to

Deleted: ould

Deleted: To

Deleted: ,

Deleted: examining the

Deleted: –

Deleted: –

Deleted: larger

Deleted: 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

Deleted: re

Deleted: is

Deleted: be due to the

Deleted: fact there is

765 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 only subject to leaf turnover as a result of leaf longevity, which may be an oversimplification.

770 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 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 (T/ET underestimated), again potentially due to the lack of the bare soil evaporation resistance term in the default 11LAY configuration. 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 appears to correlate well with spring mean LAI at US-SRM (Fig. S6).  
775 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).

780 During the monsoon at the low elevation grass- and shrub-dominated sites, both data-derived estimates of T/ET agree on the seasonality and, while different magnitudes, both are higher than the model T/ET values. Fig. 6 shows that the 11LAY model underestimates both Z16 and SB18 monthly monsoon period T/ET estimates across all low elevation sites. Given this agreement, both sets of estimated values can help to diagnose why the 11LAY model also underestimates monsoon peak ET at the low-elevation shrub sites (US-SRM and US-Whs– Figs. S3 c-d). The underestimate in modelled monsoon T/ET ratios across all grassland and shrubland sites could be either because 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. S8). 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 at the grass and shrub sites (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 grass and shrub 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 S3c-f). 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,

**Deleted:** We test this hypothesis in the following section.

**Deleted:** We can also glean some information on whether T or E (or both) are responsible for the 11LAY overestimate of springtime ET at US-SRM by comparing modelled T/ET ratios against data-derived estimates....

**Deleted:** 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). T

**Deleted:** observed spring T/ET at US-SRM is slightly underestimated by the model

**Deleted:** Given that T/ET ratios are underestimated by the model but ET is overestimated by the model,

**Deleted:** S

**Deleted:** could also

**Deleted:** be

**Deleted:** due

**Deleted:**

**Deleted:** S5

**Deleted:** ; however, the positive bias in E must be larger than the bias in T.

**Deleted:** (R.L. Scott – pers. comm.)

**Deleted:** D

**Deleted:** ratios also

**Deleted:** Fig. 6 shows that the 11LAY model also underestimates monthly T/ET ratios, and

**Deleted:** f

**Deleted:** that the model does not capture the correct temporal trajectory (Fig. 6).

**Deleted:** 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

**Deleted:** observed

**Deleted:** .

these results suggest that LAI is not increasing rapidly enough after the start of monsoon rains (see Fig. S7), resulting in negatively biased T/ET ratios in July. Meanwhile the increase in available moisture from monsoon rains, potentially coupled with a lack of bare soil evaporation resistance in the default 11LAY version, is causing a positively biased 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 leaf 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 (i.e. the model lacks the ability to grow grass in interstitial bare soil areas).

We attempted to address both hypotheses that can explain discrepancies in model ET and T/ET ratios (incorrect T due to lack of transpiring leaf area at low elevation grass and shrub sites, or overestimated E across all sites) with two further tests. First, given there is a >45% bare soil fraction at all four grass and shrub dominated sites, we increased the transpiring leaf area at these sites by increasing C4 grass cover at the expense of the bare soil. Second, given the default 11LAY does not include a bare soil evaporation resistance term, we then tested the inclusion of the optional 11LAY bare soil resistance term (described in Section 2.2.3). Third, we tested both options together. The results of these final simulations of this study are described in the following section.

#### 3.4 Testing decreased bare soil cover and the addition of the 11LAY bare soil resistance term

To further investigate the possibility that summer ET and T/ET ratios are underestimated at low-elevation sites because of a lack of transpiring leaf area, we reduced the bare soil fraction and increased C4 grass fraction to the maximum observed C4 grass cover under the most productive conditions. This decrease in bare soil fraction increased ET and T/ET ratios during the monsoon period at all sites (Fig. 7) and also increased ET during spring at the Santa Rita sites (Fig. S9). However, although the T/ET ratios reduced the negative model biases in the summer monsoon period when compared to the data-derived estimates, the model now overestimates ET in all seasons (Figs. 7 and S9). Furthermore, the spring ET model-data bias at US-SRM is further exacerbated by the decrease in bare soil fraction (Fig. S9) and the mean estimated T/ET ratios across all sites are a closer match to the original 11LAY version (Figs. 7). And finally, 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. Putting the latter points together, this new test 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 – i.e. the problem is 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.

- Deleted:** ¶  
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-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).
- Deleted:** monsoon period
- Deleted:** maximum
- Deleted:** for either grasses or shrubs
- Deleted:** e.g.
- Deleted:** growth in the model
- Deleted:** 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.
- Deleted:** of these possibilities
- Deleted:** G
- Deleted:** in the final set of simulations in this study we tested a) the
- Deleted:** reduction of the bare soil fraction by
- Deleted:** ; and b)
- Deleted:** this
- Deleted:** at the expense of bare soil fraction (R. L. Scott pers. comm.)...
- Deleted:** S8
- Deleted:** better matched
- Deleted:** values
- Deleted:** d

As described in Section 3.3, the remaining model ET issues (and its component fluxes) in both high-elevation forest sites and low-elevation shrub- and grass-dominated sites could also be due to the fact the model simulates too much bare soil evaporation. The 11LAY version has an optional bare soil evaporation resistance term that is not activated in the default version; therefore, the 11LAY simulations presented thus far have not included any such resistance term. We tested the application of the bare soil resistance term across all sites. The lack of bare soil at the high-elevation forested sites resulted in a higher sensitivity of the TeNE PFT T to the addition of the bare soil evaporation resistance term (Fig. 8 – left column). The reduction in E during the winter allowed for higher soil moisture content (Figs. S10 a and b) and therefore a greater T (and E) during the spring and summer. As a result, T/ET ratios were increased with the addition of the bare soil evaporation term, thus potentially partially resolving the issue of negatively biased T/ET ratio issue seen in the default 11LAY simulations (see Section 3.3). The increase in plant available moisture with the addition of the resistance term also led to a strong increase in LAI at US-Vcp from a mean around 0.5 to a mean around 2.1 (Fig. S10b), which is much closer to the observed LAI for the site. However, the dramatic increase in T resulted in a simulated ET at both forest sites that strongly overestimated the observations (Fig. 8 and Figs. S10a and b); therefore, overall the addition of the bare soil evaporation resistance term did not improve the ET model-data fit at these sites. As discussed in Section 3.2, spring ET may also be overestimated at these sites due to the lack of a persistent snowpack.

At all the low-elevation grass and shrub sites the addition of the bare soil resistance term resulted in a strong decrease in soil evaporation during the monsoon season, and a lesser, but non-negligible, decrease to almost zero evaporation during the winter (Fig. 8 – right column). Bare soil evaporation remained much the same during the spring and the hot, dry season months of May and June. As seen for the forest sites, the decline in bare soil evaporation during the monsoon period results in a slightly higher moisture storage (Figs. S10c-f), which in turn fractionally increases T throughout the year (Fig. 8). The net effect is a reduction in ET during summer and winter and an increase in spring and dry season ET (Fig. 8). However, as for the forested sites, this net effect in the simulated ET produces a worse model fit to the data. Therefore, the addition of this term does not resolve the ET issues documented in Section 3.3: A further positive bias in spring ET estimates is observed at US-SRM (Fig. S10c), and the underestimate in monsoon ET at US-SRM and US-Whs (Figs. S10c and d) is further exacerbated. Furthermore, the near zero evaporation in the winter months with the introduction of the bare soil resistance term results in an increase in winter T/ET ratios. Therefore, at the low-elevation sites the monthly seasonality of T/ET differs quite considerably from the default 11LAY model runs (Figs. S10c-f) and generally does not follow the seasonal trajectories estimated by either Zhou et al. (2016) or Scott et al. (2017) (Fig. 6).

In a final test, we combined both the decrease in bare soil fraction with the addition of the bare soil resistance term for the low elevation sites. The addition of the bare soil resistance term reduced the positive bias seen with the increase of C4 grass (decrease of the bare soil fraction) (Fig. S11). However, as seen in the bare soil resistance tests with the original vegetation and bare soil fractions, the addition of the resistance term increased spring T due to the higher spring soil moisture – thus exacerbating the positive bias in ET. It is clear that neither of these tests fully deal with remaining ET model-data biases in the 11LAY version – nor do they account for the issues in the model seasonality of T/ET ratios. The ET seasonal temporal

Deleted: forest

Deleted: S9

Deleted: S9

Deleted: However, as

Deleted: 3

Deleted: S

Deleted: . The increase in plant available moisture with the addition of the resistance term does lead to a strong increase in LAI at US-Vcp from a mean ~0.5 to a mean around 2.1 (Fig. S9b), which is much closer to the observed LAI for the site.

Deleted: S9

Deleted: transpiration

Deleted: ET

Deleted: S9

Deleted: summer temporal

Deleted: y

Deleted: observed

Deleted: S10

dynamics remain much the same in all tests. We point out, however that the model fit to ET observations was still greatly improved in the 11LAY version compared to the 2LAY and many of the remaining model-data discrepancies are less significant by comparison. It is therefore possible that some combination of the additional bare soil evaporation resistance term, decreased bare soil fraction, improved semi-arid leaf phenology schemes, and further calibration of hydrology, phenology, stomatal conductance, and water-limitation parameters would be able to resolve most, if not all, of the remaining model-data discrepancies in ET and T/ET estimates at these sites. This is beyond the scope of this study, but the options are discussed more in Section 4.

#### 4 Discussion

This study showed that in comparison to a simple bucket model (Manabe, 1969), a discretized soil hydrology scheme based on the Richards equation – and associated model developments – results in considerable improvements in simulated semi-arid site soil moisture temporal dynamics that exhibit a more realistic response to rainfall events (contrary to the model-data comparison of Lohou et al., 2014). As a result, we see dramatic improvements in high temporal frequency to seasonal ET simulations. In particular, there is a dramatic improvement in the model’s ability to capture the decline in ET during the hot, dry May-June period. Total runoff also decreased at forested sites owing to change in the calculation of soil moisture infiltration and partitioning between surface runoff and drainage. Such improvements might counter previous work highlighting that models tend to overestimate runoff (Grippa et al., 2017). Previous studies have also demonstrated that the more mechanistic descriptions of soil hydrology included in the latest LSM versions have resulted in improvements to surface latent and sensible heat fluxes (de Rosnay et al., 2002; Best et al., 2015); yet, few studies have specifically compared these two model versions across a range semi-arid ecosystems, as we have attempted in this study. However, there remain a number of missing hydrological processes that have not yet been incorporated into LSMs, and/or inadequate existing processes, which will clearly have an impact on semi-arid hydrological modeling (Boone et al., 2009; Grippa et al., 2017) and may resolve some of the remaining model-data discrepancies we were not able to address in this study. We highlight these in the sections below.

#### Issues with modelling vegetation dynamics in semi-arid ecosystems

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. The connection between vegetation fractional cover and LAI is 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. Similarly, there

Deleted: ;  
Deleted: ,  
Deleted: the  
Deleted: i  
Deleted: and  
Deleted: the ET seasonal temporal dynamics remain much the same in all tests.  
Deleted: se two configurations  
Deleted: plus  
Deleted: and  
Deleted: result in a near perfect fit to  
Deleted: the  
Deleted: data

Deleted: in vegetation  
Deleted: We have inferred that biases at low-elevation shrub and grassland site ET might be due to incorrect LAI simulations.  
Deleted: also  
Deleted: < 1

1000 are not many LAI measurements for grasses and shrubs in these ecosystems; therefore, we have relied on estimating the LAI<sub>max</sub>  
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 peak values 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  
1005 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.

Deleted: data

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 model does yet discern  
between perennial grasses and annual C4 grasses that only grow during warmest, wettest periods (Smith et al., 1997). It is  
1010 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. Finally, it is possible that incorrect seasonal LAI trajectories are also causing  
the issues in the T/ET seasonality seen at the higher elevation forested sites due to the lack of an evergreen phenology module  
in ORCHIDEE. Recently, a new evergreen phenology module has been implemented in ORCHIDEE (Chen et al., 2020);  
however, this scheme was developed for humid tropical forests. Testing it for evergreen trees in semi-arid regions is beyond  
1015 the scope of this study but will be investigated in future work. Again, seasonal LAI measurements of different high elevation  
semi-arid vegetation types would significantly help to improve or further develop semi-arid phenology models.

Deleted: phenological strategies

Alternatively, it may be that other model parameters and processes involved in leaf growth – for example phenology, root zone  
plant water uptake, water-limitation, and photosynthesis-related parameters – are inaccurate and in need of statistical  
calibration (e.g. MacBean et al., 2015). Incorrect representations of how we model low temperature and high VPD constraints  
1020 on stomatal conductance may also play a role. At the high-elevation sites, we assumed the ponderosa pine trees should be  
modelled as temperate needleleaved evergreen PFT. The default model parameters assigned to this PFT may not be appropriate  
for modelling this plant functional type in water-limited semi-arid environments. Another likely issue for modelling low  
elevation sparsely vegetated semi-arid ecosystems with ORCHIDEE is that there is no specific shrub PFT, although a recent  
ORCHIDEE version includes shrub PFTs for high latitude tundra ecosystems (Druel et al., 2017): In future work, we will adapt  
1025 similar shrub parameterizations for semi-arid environments.

Deleted: factor

Deleted: a

Deleted: study

The importance of vegetation cover and seasonal changes in leaf area for modeling hydrological fluxes – particularly T – is  
not a new observation (e.g. Ha et al., 2015; Grippa et al., 2017). Baldocchi et al. (2010) found that LAI was important at five  
Mediterranean sites in California and Europe for determining how much carbon is assimilated and how much water is lost.  
Hogue et al. (2005) also found the Noah LSM was not able to replicate monsoon period LE increases at the Walnut Gulch  
1030 sites, which they suggested may be related to inaccuracies in the satellite greenness fraction estimates that are used to run the  
model. Whitley et al. (2016 and 2017) also proposed that any improvements needed for terrestrial biosphere modelling of  
savanna ecosystems should include modifications to the phenology schemes and the split between fractional cover of trees and  
grasses.

1040 **ET partitioning (T/ET ratio)**

ORCHIDEE underestimated data-derived estimates of T/ET ratios across all sites, particularly during the monsoon (Fig. 6).

**Deleted:** t the low-elevation sites

The partitioning of ET between evaporation, E, and transpiration, T, is strongly related to presence of vegetative cover and plant physiological functioning: Whereas E results from physical processes and is linked to bare soil surface moisture and interception by the plant canopy, T is the result of leaf level biological processes regulated by LAI and stomatal conductance.

1045 Stomatal conductance is in turn modulated by root zone soil moisture availability, vapor pressure deficit (VPD) and atmospheric CO<sub>2</sub> concentrations (Novick et al., 2016). Both E and T are therefore governed by the fractional cover of vegetation, vegetation type, and changes in leaf area (i.e. phenology). and thus model testing and developments proposed in the previous section may be particularly important for accurate partitioning of T versus E. Diagnosing and addressing

**Deleted:** model-data

discrepancies between modelled and estimated T/ET is important, specifically for dryland ecosystems where increases in 1050 vegetation productivity and/or cover in response to rising atmospheric CO<sub>2</sub> appears to be driving higher T/ET rates (Lian et al., 2018).

**Deleted:** in

**Deleted:** observed

In agreement with this study, Lian et al. (2018) also show that CMIP5 models vastly underestimate T/ET ratios. They estimated a new global T/ET ratio of  $0.62 \pm 0.06$ , which is similar to the upscaled estimate of  $0.57 \pm 7\%$  of Wei et al. (2017), and suggest that model underestimates could be caused by misrepresentation of vegetation structure impacts on canopy light use, 1055 interception loss and root water uptake. Their conclusions lend further weight to our suggestion that further improvements in T/ET ratios may result from more accurate simulations of seasonal phenology and fractional vegetation cover (see previous section). Alternatively, Chang et al., (2018) have suggested that neglecting to account for lateral redistribution of moisture is responsible for model inability to capture T/ET partitioning. Current LSM versions do not simulate extensive shallow root systems that are typical of semi-arid vegetation that is more adapted to water limited conditions. However, they also mention

**Deleted:** estimates

1060 other LSM issues that might be affecting the T/ET ratio, such as the lack of root dynamics, vegetation shading, topographic effects and the representation of bare soil evaporation. In order to properly diagnose if discrepancies in modelled T/ET are caused by inaccurate representation of lateral moisture redistribution, we need to perform a comparison of a spatially distributed model simulation with a high-density network of hydrological observations. Nevertheless, in spatially heterogeneous mixed shrub-grass ecosystems it seems likely that missing model processes will need to be accounted for before 1065 accurate simulations of T/ET ratios can be achieved. One example of this might be the need to include in the model a representation of shrub understory and below canopy E.

**Deleted:** are

**Bare soil evaporation**

1070 Following the results of our analysis, we surmised that bare soil evaporation might be overestimated, particularly at grassland sites that showed a good match between modelled and observed ET, but negatively biased T/ET ratios. The addition of a term that simulates bare soil evaporation resistance to dry soil served to reduce E in the summer and winter and thus to increase



(and improve) the T/ET ratios (compared to data-derived estimates). However, resulting changes in modelled ET provided a worse fit to the observations. 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). 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. Alternatively, the relatively simple implementation of a bare soil resistance term (Eq. 2 – Section 2.2.3) might need to be adapted to include bare soil evaporation resistance across a litter or biocrust layer. Decker et al. (2017) noted a reduction in positive biases in evaporation after updating the soil evaporation scheme in the CABLE LSM to limit bare soil evaporation based on the physics of evaporation from porous media across a viscous sublayer (analogous to a dry matter litter layer). Similarly, Swenson and Lawrence (2014) accounted for estimated positive biases in semi-arid bare soil evaporation fluxes (and too high fluctuations in modelled ET) in the Community Land Model (CLM v4.5) by replacing the existing empirical evaporation resistance function with a more mechanistic scheme. In the more mechanistic scheme, bare soil evaporation is formulated as a diffusion of moisture through a dry surface layer (again, approximating a litter layer). Despite the fact that site-based soil textural properties were set for all simulations, it is also possible that modelled bare soil evaporation is too high because the model lacks vertical soil texture variability. The low-elevation sites typically have a very cobbly, rocky soil surface that is not accounted for in ORCHIDEE. Including soil texture variability with different soil horizons could further improve ORCHIDEE's capability to capture the correct E, ET and T/ET ratios. Furthermore, at the sparsely vegetated grass and shrub dominated sites in southern Arizona litter is barely present; instead, biological soil crusts (biocrusts) composed of assemblages of lichens, bryophytes, cyanobacteria, algae and microbes form across much of the bare soil surface (Belnap et al., 2016). Biocrust layers may significantly alter bare soil evaporation (and other aspects of ecosystem ecology and functioning – Ferrenberg et al., 2017) in sparsely vegetated regions in ways that have not yet been considered in any LSM bare soil evaporation scheme. Therefore, it is possible that in addition to a more mechanistically-based formulation of bare soil evaporation through a litter layer (as per Swenson and Lawrence, 2014 or Decker et al., 2017), separate formulations of evaporation through biocrust/mulch layers may need to be developed (e.g. Saux-Picard et al., 2009).

Deleted: re-

Deleted: and

#### High-elevation model snowpack and snow melt predictions

The model also needs to be tested at other high-elevation semi-arid mountainous sites (such as the Sierra Nevada mountains in California) for which spring snowmelt is the predominant (and controlling) annual source of moisture. More specifically, more information on snow cover, depth or mass, particularly under closed forest canopies, would be useful to diagnose potential sources of bias in the snowfall simulations. It is crucial that LSMs accurately capture semi-arid high-elevation snowfall temporal dynamics if we are to have unbiased projections in future moisture availability and productivity for these regions.

Deleted: test if the precipitation data measured by the meteorological stations accurately captures the right amount of snowfall...

### 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 the findings of 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 response to 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 seen across sites 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  $\geq 2.4\text{mm d}^{-1}$  in Fig. 4g of Wang et al., 2018). However, the fact LSMs typically do not include adequate descriptions of groundwater could impact their ability to simulate semi-arid ecosystem water uptake in the dry season given that drought deciduous shrubs 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.

### 5 Conclusions

These results strongly suggest that a more complex, process-based hydrology model – in particular one which contains fine scale discretization of the upper soil moisture layers and associated improvements in bare soil evaporation and plant water stress functions – improves daily to seasonal predictions of the upper layer root-zone soil moisture dynamics and ET (as seen in de Rosnay et al., 2002). 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. ORCHIDEE CMIP5 simulations used the 2-layer conceptual bucket scheme of Manabe (1969); therefore, ORCHIDEE CMIP5 predictions of semi-

**Deleted:** limitation of stomatal conductance

**Deleted:** high frequency

**Deleted:** resulted

**Deleted:** realistic

**Deleted:** surface

155 arid water availability and consequent impacts on ecosystem functioning and feedback to climate were likely inaccurate. Despite the appeal of simplicity and low calculation costs, 2-layer simple bucket hydrology models are likely unsuitable for accurate water flux simulations in the semi-arid SW US. The forthcoming ORCHIDEE CMIP6 simulations will likely provide more accurate and reliable results of semi-arid soil moisture availability and evapotranspiration.

160 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. However, biases in soil moisture at both the forested sites do not translate into the same biases in modelled ET, 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.

165 The addition of an empirical bare soil evaporation resistance term by itself did not improve estimates of ET in these ecosystems, although T/ET ratios were increased, potentially reducing the negative biases in the monsoon season when comparing to data-derived T/ET estimates. The increase in transpiring leaf area (from a reduction in bare soil fraction) at the low elevation forest sites also could account for the same monsoon season T/ET bias. However, issues in the timing of the simulated transition from low to high T/ET ratios at the start of the monsoon remain. Our analysis shows that remaining discrepancies semi-arid site ET simulations (and its constituent fluxes) might therefore be related to a combination of factors impacting both the amount and timing transpiring leaf area and resistance to bare soil evaporation. 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 as well as the vegetation fractional cover and peak LAI magnitude at low elevation sites; and iv) testing of a more mechanistic description of resistance to bare soil evaporation.

1175

#### Code availability

The ORCHIDEE model code and documentation are publicly available via the ORCHIDEE wiki page (<https://forge.ipsl.jussieu.fr/orchidee/browser>) under the CeCILL license (<http://www.cecill.info/index.en.html>). ORCHIDEE model code is written in Fortran 90 and is maintained and developed under an SVN version control system at the Institute Pierre Simon Laplace (IPSL) in France.

1180

Deleted: of semi-arid soil moisture availability

Deleted: inaccurate

Deleted: of

Deleted: and improved

Deleted: with the addition of this term

Deleted: in sparsely vegetated grass and shrub dominated

Deleted: amount of

Deleted: at fully documented

Deleted: ing the

Deleted: and

Deleted: amount of

Deleted: ii

Deleted: s

### **Data availability**

1195 Meteorological forcing and evapotranspiration data for each are available from the Ameriflux site: <https://ameriflux.lbl.gov>.  
Soil moisture was obtained directly from site PIs. Vegetation and soil texture characteristics were derived from the published  
literature, as specified in Table 1, and from site PIs. Model simulations are provided on NM's GitHub repository:  
<https://github.com/nmacbean>.

### **\*\*\*Appendices\*\*\***

#### **1200 Author contribution**

NM, RLS, JAB and DJPM designed the overall study. NM carried out the model simulations, post-simulation analysis and  
figure plotting. CO, NV and AD provided detailed inputs on model description/code and recommendations for further tests to  
diagnose model-data deficiencies. NV provided scripts to gap-fill the meteorological data. JAB gap-filled the ET data. RLS,  
JAB, TK and ML provided gap-filled soil moisture data and information on site characteristics and typical behaviour of  
1205 seasonal vegetation cover, LAI, and snowfall. NM wrote the manuscript. All co-authors provided detailed comments,  
suggestions, and edits on the first and second drafts of the manuscript.

#### **Competing Interests**

The authors declare that they have no conflict of interest.

#### **Acknowledgements**

1210 Funding for AmeriFlux data resources and data collection at US-SRM, US-SRG, US-Wkg, and US-Whs was provided by the  
U.S. Department of Energy's Office of Science and the USDA. Data collection at US-Fuf supported by grants from the North  
American Carbon Program/USDA CREES NRI (2004-3511115057), the U.S. National Science Foundation MRI Program,  
Science Foundation Arizona (CAA 0-203-08), the Arizona Water Institute, and the Mission Research Program, School of  
Forestry, Northern Arizona University (McIntire-Stennis/ Arizona Bureau of Forestry). The US-Vcp site funded by U.S. DOE  
1215 Office of Science through the AmeriFlux Management Project (AMP) at Lawrence Berkeley National Laboratory (Award  
#7074628) and Catalina-Jemez Critical Zone Observatory (NSF EAR 1331408). NM was funded by the US National Science  
Foundation Award Numbers 1065790 (Emerging Frontiers Program) and 1754430 (Division of Environmental Biology  
Ecosystems Program). We would like to thank the ORCHIDEE team for development and maintenance of the ORCHIDEE  
code and for providing the ORCHIDEE version used in this study.

1220

## References

Abramowitz, G., Leuning, R., Clark, M., & Pitman, A.: Evaluating the performance of land surface models. *Journal of Climate*, 21(21), 5468-5481, 2008.

Allen CD (2016) Chapter 4 - Forest ecosystem reorganization underway in the Southwestern US: A preview of widespread forest changes in the Anthropocene? In: RP Bixler and C Miller (eds) Forest Conservation and Management in the Anthropocene: Adaptation of Science, Policy and Practice. University Press of Colorado, Boulder, Colorado, pp 57-79

Ault, T. R., Mankin, J. S., Cook, B. I. and Smerdon, J. E.: Relative impacts of mitigation, temperature, and precipitation on 21st-century megadrought risk in the American Southwest, *Science Advances*, 2(10), doi:10.1126/sciadv.1600873, 2016.

Baldocchi, D. D., Ma, S., Rambal, S., Misson, L., Ourcival, J.-M., Limousin, J.-M., Pereira, J. and Papale, D.: On the differential advantages of evergreenness and deciduousness in mediterranean oak woodlands: a flux perspective, *Ecological Applications*, 20(6), 1583–1597, doi:10.1890/08-2047.1, 2010.

Bastrikov, V., Peylin, P. and Ottlé, C: Optimizing albedo computation in ORCHIDEE land surface model by assimilating MODIS satellite observation data. In prep.

Belnap, J., Weber, B., and Büdel B.: Biological soil crusts: an organizing principle in drylands, Springer., 2016.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., B. J. J. Van Den Hurk, Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L. and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, *Journal of Hydrometeorology*, 16(3), 1425–1442, doi:10.1175/jhm-d-14-0158.1, 2015.

Berg, A., Findell, K., Lintner, B., Giannini, A., Seneviratne, S., van den Hurk, B., Lorenz, R., Pitman, A., Hagemann, S., Meier, A., Cheruy, F., Ducharne, A., Malyshev, S., and Milly, P.C.D.: Land-atmosphere feedbacks amplify aridity increase over land under global warming, *Nature Climate Change*, 6, 869–874, doi:10.1038/nclimate3029, 2016.

Biederman, J. A., Scott, R. L., Goulden, M. L., Vargas, R., Litvak, M. E., Kolb, T. E., Yezek, E. A., Oechel, W. C., Blanken, P. D., Bell, T. W., Garatuza-Payan, J., Maurer, G. E., Dore, S. and Burns, S. P.: Terrestrial carbon balance in a drier world: the effects of water availability in southwestern North America, *Global Change Biology*, 22(5), 1867–1879, doi:10.1111/gcb.13222, 2016.

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

[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.](#)

**Deleted:** Ahlström, A., Raupach, M.R., Schurgers, G., Smith, B., Arneth, A., Jung, M., Reichstein, M., Canadell, J.G., Friedlingstein, P., Jain, A.K. and Kato, E.: The dominant role of semi-arid ecosystems in the trend and variability of the land CO<sub>2</sub> sink. *Science*, 348(6237), pp.895-899, 2015\*

- Boone, A., de Rosnay, P. D., Balsamo, G., Beljaars, A., Chopin, F., Decharme, B., Delire, C., Ducharne, A., Gascoïn, S., Grippa, M., Guichard, F., Gusev, Y., Harris, P., Jarlan, L., Kergoat, L., Mougïn, E., Nasonova, O., Norgaard, A., Orgeval, T., Ottlé, C., Pocard-Leclercq, I., Polcher, J., Sandholt, I., Saux-Picart, S., Taylor, C. and Xue, Y.: The AMMA Land Surface Model Intercomparison Project (ALMIP), *Bulletin of the American Meteorological Society*, 90(12), 1865–1880, doi:10.1175/2009bams2786.1, 2009.
- 1260 Botta, A., Viovy, N., Ciais, P., Friedlingstein, P. and Monfray, P.: A global prognostic scheme of leaf onset using satellite data, *Global Change Biology*, 6(7), 709–725, doi:10.1046/j.1365-2486.2000.00362.x, 2000.
- Campoy, A., Ducharne, A., Cheruy, F., Hourdin, F., Polcher, J. and Dupont, J. C.: Response of land surface fluxes and precipitation to different soil bottom hydrological conditions in a general circulation model, *Journal of Geophysical Research: Atmospheres*, 118(19), doi:10.1002/jgrd.50627, 2013.
- 265 [Carsel, R. F. and Parrish, R. S.: Developing joint probability distributions of soil water retention characteristics, \*Water Resources Research\*, 24\(5\), 755–769, doi:10.1029/wr024i005p00755, 1988.](#)
- Chang, L.-L., Dwivedi, R., Knowles, J. F., Fang, Y.-H., Niu, G.-Y., Pelletier, J. D., Rasmussen, C., Durcik, M., Barron-Gafford, G. A. and Meixner, T.: Why Do Large-Scale Land Surface Models Produce a Low Ratio of Transpiration to Evapotranspiration?, *Journal of Geophysical Research: Atmospheres*, 123(17), 9109–9130, doi:10.1029/2018jd029159, 2018.
- 270 [Chen, X., Maignan, F., Viovy, N., Bastos, A., Goll, D., Wu, J., Liu, L., Yue, C., Peng, S., Yuan, W., Conceição, A. C., O'sullivan, M. and Ciais, P.: Novel Representation of Leaf Phenology Improves Simulation of Amazonian Evergreen Forest Photosynthesis in a Land Surface Model, \*Journal of Advances in Modeling Earth Systems\*, 12\(1\), doi:10.1029/2018ms001565, 2020.](#)
- 1275 Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R. P., Kumar, M., Leung, L. R., Mackay, D. S., Maxwell, R. M., Shen, C., Swenson, S. C. and Zeng, X.: Improving the representation of hydrologic processes in Earth System Models, *Water Resources Research*, 51(8), 5929–5956, doi:10.1002/2015wr017096, 2015.
- Cook, B. I., Ault, T. R. and Smerdon, J. E.: Unprecedented 21st century drought risk in the American Southwest and Central Plains, *Science Advances*, 1(1), doi:10.1126/sciadv.1400082, 2015.
- 1280 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.
- 285 [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.](#)

1290 De Kauwe, M. G., Taylor, C. M., Harris, P. P., Weedon, G. P. and Ellis, R. J.: Quantifying Land Surface Temperature Variability for Two Sahelian Mesoscale Regions during the Wet Season, *Journal of Hydrometeorology*, 14(5), 1605–1619, doi:10.1175/jhm-d-12-0141.1, 2013.

Decker, M., Or, D., Pitman, A. and Ukkola, A.: New turbulent resistance parameterization for soil evaporation based on a pore-scale model: Impact on surface fluxes in CABLE, *Journal of Advances in Modeling Earth Systems*, 9(1), 220–238, doi:10.1002/2016ms000832, 2017.

1295 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, doi:10.5194/hess-2-239-1998, 1998.

[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.](#)

1300 de Rosnay, P., Polcher, J., Bruen, M. and Laval, K.: Impact of a physically based soil water flow and soil-plant interaction representation for modeling large-scale land surface processes, *Journal of Geophysical Research: Atmospheres*, 107(D11), doi:10.1029/2001jd000634, 2002.

Diffenbaugh, N. S., Giorgi, F. and Pal, J. S.: Climate change hotspots in the United States, *Geophysical Research Letters*, 35(16), doi:10.1029/2008gl035075, 2008.

Dirmeyer, P. A.: A History and Review of the Global Soil Wetness Project (GSWP), *Journal of Hydrometeorology*, 110404091221083, doi:10.1175/jhm-d-10-05010, 2011.

1305 Donat, M. G., Lowry, A. L., Alexander, L. V., O’Gorman, P. A. and Maher, N.: More extreme precipitation in the world’s dry and wet regions, *Nature Climate Change*, 6(5), 508–513, doi:10.1038/nclimate2941, 2016.

Dore, S., Kolb, T. E., Montes-Helu, M., Eckert, S. E., Sullivan, B. W., Hungate, B. A., Kaye, J. P., Hart, S. C., Koch, G. W. and Finkral, A.: Carbon and water fluxes from ponderosa pine forests disturbed by wildfire and thinning, *Ecological Applications*, 20(3), 663–683, doi:10.1890/09-0934.1, 2010.

1310 Dore, S., Montes-Helu, M., Hart, S. C., Hungate, B. A., Koch, G. W., Moon, J. B., Finkral, A. J. and Kolb, T. E.: Recovery of ponderosa pine ecosystem carbon and water fluxes from thinning and stand-replacing fire, *Global Change Biology*, 18(10), 3171–3185, doi:10.1111/j.1365-2486.2012.02775.x, 2012.

d’Orgeval, T., Polcher, J., and de Rosnay, P.: Sensitivity of the West African hydrological cycle in ORCHIDEE to infiltration processes, *Hydrol. Earth Syst. Sci.*, 12, 1387–1401, 2008.

Deleted: D.


- Druel, A., Peylin, P., Krinner, G., Ciais, P., Viovy, N., Peregon, A., Bastrikov, V., Kosykh, N., and Mironycheva-Tokareva, N.: Towards a more detailed representation of high-latitude vegetation in the global land surface model ORCHIDEE (ORCHIDEE-VEGv1.0), *Geosci. Model Dev.*, 10, 4693–4722.
- Ducharne, A., Laval, K. and Polcher, J.: Sensitivity of the hydrological cycle to the parametrization of soil hydrology in a  
1320 GCM, *Climate Dynamics*, 14(5), 307–327, doi:10.1007/s003820050226, 1998.
- Ducharne, A., Ghattas, J., Maignan, F., Ottlé, C., Vuichard, N., Guimberteau, M., Krinner, G., Polcher, J., Tafasca, S., Bastrikov, V., Cheruy, F., Guénet, B., Mizuochi, H., Peylin, P., Tootchi, A. and Wang, F.: Soil water processes in the ORCHIDEE-2.0 land surface model: state of the art for CMIP6, In prep. for *Geosci. Model Dev.*
- Ducoudré, N. I., Laval, K. and Perrier, A.: SECHIBA, a New Set of Parameterizations of the Hydrologic Exchanges at the  
1325 Land-Atmosphere Interface within the LMD Atmospheric General Circulation Model, *Journal of Climate*, 6(2), 248–273, doi:10.1175/1520-0442(1993)006<0248:sansop>2.0.co;2, 1993.
- Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., Noblet, N. D., Duvel, J.-P., Ethé, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J.-Y., Guez, L.,  
1330 Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M.-P., Lefevre, F., Levy, C., Li, Z. X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N. and Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, *Climate Dynamics*, 40(9-10), 2123–2165, doi:10.1007/s00382-012-1636-1, 2013.
- 1335 Fang, H., Jiang, C., Li, W., Wei, S., Baret, F., Chen, J.M., Garcia-Haro, J., Liang, S., Liu, R., Myneni, R.B. and Pinty, B.: Characterization and intercomparison of global moderate resolution leaf area index (LAI) products: Analysis of climatologies and theoretical uncertainties. *Journal of Geophysical Research: Biogeosciences*, 118(2), 529-548, 2013.
- Ferrenberg, S. and Reed, S. C.: Biocrust ecology: unifying micro- and macro-scales to confront global change, *New Phytologist*, 216(3), 643–646, doi:10.1111/nph.14826, 2017.
- 1340 Garrigues, S., Lacaze, R., Baret, F.J.T.M., Morisette, J.T., Weiss, M., Nickeson, J.E., Fernandes, R., Plummer, S., Shabanov, N.V., Myneni, R.B. and Knyazikhin, Y.: Validation and intercomparison of global Leaf Area Index products derived from remote sensing data. *Journal of Geophysical Research: Biogeosciences*, 113(G2), 2008.
- Genuchten, M. T. V.: A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils<sup>1</sup>, *Soil Science Society of America Journal*, 44(5), 892, doi:10.2136/sssaj1980.03615995004400050002x, 1980.




- 1345 Gremer, J. R., Bradford, J. B., Munson, S. M. and Duniway, M. C.: Desert grassland responses to climate and soil moisture suggest divergent vulnerabilities across the southwestern United States, *Global Change Biology*, 21(11), 4049–4062, doi:10.1111/gcb.13043, 2015.
- Grippa, M., Kergoat, L., Frappart, F., Araud, Q., Boone, A., de Rosnay, P. D., Lemoine, J.-M., Gascoïn, S., Balsamo, G., Ottlé, C., Decharme, B., Saux-Picart, S. and Ramillien, G.: Land water storage variability over West Africa estimated by Gravity Recovery and Climate Experiment (GRACE) and land surface models, *Water Resources Research*, 47(5), doi:10.1029/2009wr008856, 2011.
- 1350 Grippa, M., Kergoat, L., Boone, A., Peugeot, C., Demarty, J., Cappelaere, B., Gal, L., Hiernaux, P., Mougin, E., Ducharne, A., Dutra, E., Anderson, M. and Hain, C.: Modeling Surface Runoff and Water Fluxes over Contrasted Soils in the Pastoral Sahel: Evaluation of the ALMIP2 Land Surface Models over the Gourma Region in Mali, *Journal of Hydrometeorology*, 18(7), 1847–1866, doi:10.1175/jhm-d-16-0170.1, 2017.
- 1355 Guimberteau, M., Drapeau, G., Ronchail, J., Sultan, B., Polcher, J., Martinez, J.-M., Prigent, C., Guyot, J.-L., Cochonneau, G., Espinoza, J. C., Filizola, N., Fraizy, P., Lavado, W., Oliveira, E. D., Pombosa, R., Noriega, L. and Vauchel, P.: Discharge simulation in the sub-basins of the Amazon using ORCHIDEE forced by new datasets, *Hydrology and Earth System Sciences*, 16(3), 911–935, doi:10.5194/hess-16-911-2012, 2012a.
- 1360 Guimberteau, M., Ronchail, J., Espinoza, J. C., Lengaigne, M., Sultan, B., Polcher, J., Drapeau, G., Guyot, J.-L., Ducharne, A. and Ciais, P.: Future changes in precipitation and impacts on extreme streamflow over Amazonian sub-basins, *Environmental Research Letters*, 8(1), 014035, doi:10.1088/1748-9326/8/1/014035, 2013.
- Guimberteau, M., Ducharne, A., Ciais, P., Boisier, J. P., Peng, S., Weirtdt, M. D. and Verbeeck, H.: Testing conceptual and physically based soil hydrology schemes against observations for the Amazon Basin, *Geoscientific Model Development*, 7(3), 1115–1136, doi:10.5194/gmd-7-1115-2014, 2014.
- 1365 Guimberteau, M., Perrier, A., Laval, K. and Polcher, J.: A comprehensive approach to analyze discrepancies between land surface models and in-situ measurements: a case study over the US and Illinois with SECHIBA forced by NLDAS, *Hydrology and Earth System Sciences*, 16(11), 3973–3988, doi:10.5194/hess-16-3973-2012, 2012b.
- 1370 [Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, \*Journal of Hydrology\*, 377\(1-2\), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.](#)
- Haverd, V., Ahlström, A., Smith, B. and Canadell, J. G.: Carbon cycle responses of semi-arid ecosystems to positive asymmetry in rainfall, *Global Change Biology*, 23(2), 793–800, doi:10.1111/gcb.13412, 2016.

- 1375 Hogue, T. S., Bastidas, L., Gupta, H., Sorooshian, S., Mitchell, K. and Emmerich, W.: Evaluation and Transferability of the Noah Land Surface Model in Semiarid Environments, *Journal of Hydrometeorology*, 6(1), 68–84, doi:10.1175/jhm-402.1, 2005.
- Huang, J., Yu, H., Dai, A., Wei, Y. and Kang, L.: Drylands face potential threat under 2 °C global warming target, *Nature Climate Change*, 7(6), 417–422, doi:10.1038/nclimate3275, 2017.
- 1380 IPCC, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp., 2013.
- [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.](#)
- 1385 Koster, R. D.: Regions of Strong Coupling Between Soil Moisture and Precipitation, *Science*, 305(5687), 1138–1140, doi:10.1126/science.1100217, 2004.
- Krinner, G., Viovy, N., Noblet-Ducoudré, N. D., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochemical Cycles*, 19(1), doi:10.1029/2003gb002199, 2005.
- 1390 Lian, X., Piao, S., Huntingford, C., Li, Y., Zeng, Z., Wang, X., Ciais, P., McVicar, T. R., Peng, S., Ottlé, C., Yang, H., Yang, Y., Zhang, Y. and Wang, T.: Partitioning global land evapotranspiration using CMIP5 models constrained by observations, *Nature Climate Change*, 8(7), 640–646, doi:10.1038/s41558-018-0207-9, 2018.
- Lohou, F., Kergoat, L., Guichard, F., Boone, A., Cappelaere, B., Cohard, J.-M., Demarty, J., Galle, S., Grippa, M., Peugeot, C., Ramier, D., Taylor, C. M. and Timouk, F.: Surface response to rain events throughout the West African monsoon, *Atmospheric Chemistry and Physics*, 14(8), 3883–3898, doi:10.5194/acp-14-3883-2014, 2014.
- 1395 MacBean, N., Maignan, F., Peylin, P., Bacour, C., Bréon, F.-M. and Ciais, P.: Using satellite data to improve the leaf phenology of a global terrestrial biosphere model, *Biogeosciences*, 12(23), 7185–7208, doi:10.5194/bg-12-7185-2015, 2015.
- Maestre, F. T., Salguero-Gomez, R. and Quero, J. L.: It is getting hotter in here: determining and projecting the impacts of global environmental change on drylands, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1606), 3062–3075, doi:10.1098/rstb.2011.0323, 2012.
- 1400 Manabe, S.: Climate And The Ocean Circulation1, *Monthly Weather Review*, 97(11), 739–774, doi:10.1175/1520-0493(1969)097<0739:catoc>2.3.co;2, 1969.

- 405 [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 priori predictions for the CO<sub>2</sub> response of a nutrient- and water-limited native Eucalypt woodland, \*Global Change Biology\*, 22\(8\), 2834–2851, doi:10.1111/gcb.13268, 2016.](#)
- [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.](#)
- 410 [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.](#)
- [Mualem, Y.: A new model for predicting the hydraulic conductivity of unsaturated porous media, \*Water Resources Research\*, 12\(3\), 513–522, doi:10.1029/wr012i003p00513, 1976.](#)
- 415 [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.](#)
- [Mueller, B. and Seneviratne, S. I.: Systematic land climate and evapotranspiration biases in CMIP5 simulations, \*Geophysical Research Letters\*, 41\(1\), 128–134, doi:10.1002/2013gl058055, 2014.](#)
- 420 [Niu, G. Y., and Yang, Z. L.: An observation-based formulation of snow cover fraction and its evaluation over large North American river basins, \*Journal of Geophysical Research: Atmospheres\*, 112\(D21\), 2007.](#)
- [Novick, K. A., Ficklin, D. L., Stoy, P. C., Williams, C. A., Bohrer, G., Oishi, A. C., Papuga, S. A., Blanken, P. D., Noormets, A., Sulman, B. N., Scott, R. L., Wang, L. and Phillips, R. P.: The increasing importance of atmospheric demand for ecosystem water and carbon fluxes, \*Nature Climate Change\*, 6\(11\), 1023–1027, doi:10.1038/nclimate3114, 2016.](#)
- 425 [Raoult, N., Delorme, B., Ottlé, C., Peylin, P., Bastrikov, V., Maugis, P. and Polcher, J.: Confronting Soil Moisture Dynamics from the ORCHIDEE Land Surface Model With the ESA-CCI Product: Perspectives for Data Assimilation, \*Remote Sensing\*, 10\(11\), 1786, doi:10.3390/rs10111786, 2018.](#)
- [Reynolds, C. A., Jackson, T. J. and Rawls, W. J.: Estimating soil water-holding capacities by linking the Food and Agriculture Organization Soil map of the world with global pedon databases and continuous pedotransfer functions, \*Water Resources Research\*, 36\(12\), 3653–3662, doi:10.1029/2000wr900130, 2000.](#)
- 430 [Richards, L. A.: Capillary Conduction Of Liquids Through Porous Mediums, \*Physics\*, 1\(5\), 318–333, doi:10.1063/1.1745010, 1931.](#)

**Deleted:** Peylin, P., et al.: The ORCHIDEE global land surface model version v2.0: description and evaluation. *Geoscientific Model Development*, in prep. 

Poulter, B., Frank, D., Ciais, P., Myneni, R. B., Andela, N., Bi, J., Broquet, G., Canadell, J. G., Chevallier, F., Liu, Y. Y., Running, S. W., Sitch, S. and Werf, G. R. V. D.: Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle, *Nature*, 509(7502), 600–603, doi:10.1038/nature13376, 2014. 

- Saux-Picart, S., Ottlé, C., Perrier, A., Decharme, B., Coudert, B., Zribi, M., ... & Ramier, D.: SEtHyS\_Savannah: A multiple source land surface model applied to Sahelian landscapes. *Agricultural and Forest Meteorology*, 149(9), 1421-1432, 2009.
- Scanlon, B. R., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoin, H., Lo, M. H., Müller-Schmied, H., Döll, P., Beek, R., Swenson, S., Lawrence, D., Croteau, M. and Reedy, R. C.: Tracking Seasonal Fluctuations in Land Water Storage Using Global Models and GRACE Satellites, *Geophysical Research Letters*, doi:10.1029/2018gl081836, 2019.
- 1445 Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Schmied, H. M., Beek, L. P. H. V., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P. and Bierkens, M. F. P.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, *Proceedings of the National Academy of Sciences*, 115(6), doi:10.1073/pnas.1704665115, 2018.
- 1450 Scott, R. L. and Biederman, J. A.: Critical Zone Water Balance Over 13 Years in a Semiarid Savanna, *Water Resources Research*, 55(1), 574–588, doi:10.1029/2018wr023477, 2019.
- Scott, R. L. and Biederman, J. A.: Partitioning evapotranspiration using long-term carbon dioxide and water vapor fluxes, *Geophysical Research Letters*, 44(13), 6833–6840, doi:10.1002/2017gl074324, 2017.
- Scott, R. L., Biederman, J. A., Hamerlynck, E. P. and Barron-Gafford, G. A.: The carbon balance pivot point of southwestern U.S. semiarid ecosystems: Insights from the 21st century drought, *Journal of Geophysical Research: Biogeosciences*, 120(12), 2612–2624, doi:10.1002/2015jg003181, 2015.
- 1455 Scott, R. L., Biederman, J. A., Hamerlynck, E. P. and Barron-Gafford, G. A.: The carbon balance pivot point of southwestern U.S. semiarid ecosystems: Insights from the 21st century drought, *Journal of Geophysical Research: Biogeosciences*, 120(12), 2612–2624, doi:10.1002/2015jg003181, 2015.
- 1460 Seager, R., Ting, M., Held, I., Kushnir, Y., Lu, J., Vecchi, G., Huang, H.-P., Harnik, N., Leetmaa, A., Lau, N.-C., Li, C., Velez, J. and Naik, N.: Model Projections of an Imminent Transition to a More Arid Climate in Southwestern North America, *Science*, 316(5828), 1181–1184, doi:10.1126/science.1139601, 2007.
- Sellers, P. J., Heiser, M. D. and Hall, F. G.: Relations between surface conductance and spectral vegetation indices at intermediate (100 m to 15 km) length scales, *Journal of Geophysical Research*, 97(D17), 19033, doi:10.1029/92jd01096, 1992.
- 1465 Seneviratne, S. I., Wilhelm, M., Stanelle, T., Hurk, B., Hagemann, S., Berg, A., Cheruy, F., Higgins, M. E., Meier, A., Brovkin, V., Claussen, M., Ducharne, A., Dufresne, J. L., Findell, K. L., Ghattas, J., Lawrence, D. M., Malyshev, S., Rummukainen, M. and Smith, B.: Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment, *Geophysical Research Letters*, 40(19), 5212–5217, doi:10.1002/grl.50956, 2013.

- 1470 [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.](#)
- Sippel, S., Zscheischler, J., Heimann, M., Lange, H., Mahecha, M. D., Oldenborgh, G. J. V., Otto, F. E. L. and Reichstein, M.: Have precipitation extremes and annual totals been increasing in the world's dry regions over the last 60 years?, *Hydrology and Earth System Sciences*, 21(1), 441–458, doi:10.5194/hess-21-441-2017, 2017.
- 1475 Smith, S.D., Monson, R.K. and Anderson, J.E.: *Physiological Ecology of North American Desert Plants*, Springer-Verlag Berlin Heidelberg, 1997.
- Swenson, S. C. and Lawrence, D. M.: Assessing a dry surface layer-based soil resistance parameterization for the Community Land Model using GRACE and FLUXNET-MTE data, *Journal of Geophysical Research: Atmospheres*, 119(17), 1480 doi:10.1002/2014jd022314, 2014.
- Tietjen, B., Jeltsch, F., Zehe, E., Classen, N., Groengroeft, A., Schiffrers, K. and Oldeland, J.: Effects of climate change on the coupled dynamics of water and vegetation in drylands, *Ecohydrology*, doi:10.1002/eco.70, 2009.
- Ukkola, A. M., Kauwe, M. G. D., Pitman, A. J., Best, M. J., Abramowitz, G., Haverd, V., Decker, M. and Haughton, N.: Land surface models systematically overestimate the intensity, duration and magnitude of seasonal-scale evaporative droughts, 1485 *Environmental Research Letters*, 11(10), 104012, doi:10.1088/1748-9326/11/10/104012, 2016a.
- Ukkola, A. M., Pitman, A. J., Decker, M., Kauwe, M. G. D., Abramowitz, G., Kala, J. and Wang, Y.-P.: Modelling evapotranspiration during precipitation deficits: identifying critical processes in a land surface model, *Hydrology and Earth System Sciences*, 20(6), 2403–2419, doi:10.5194/hess-20-2403-2016, 2016b.
- Villarreal, S., Vargas, R., Yopez, E. A., Acosta, J. S., Castro, A., Escoto-Rodriguez, M., Lopez, E., Martinez-Osuna, J., 1490 Rodriguez, J. C., Smith, S. V., Vivoni, E. R. and Watts, C. J.: Contrasting precipitation seasonality influences evapotranspiration dynamics in water-limited shrublands, *Journal of Geophysical Research: Biogeosciences*, 121(2), 494–508, doi:10.1002/2015jg003169, 2016.
- Vuichard, N. and Papale, D.: Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis, *Earth System Science Data*, 7(2), 157–171, doi:10.5194/essd-7-157-2015, 2015. Wang, T., Ottlé, C., Boone, 1495 A., Ciais, P., Brun, E., Morin, S., Krinner, G., Piao, S. and Peng, S.: Evaluation of an improved intermediate complexity snow scheme in the ORCHIDEE land surface model, *Journal of Geophysical Research: Atmospheres*, 118(12), 6064–6079, doi:10.1002/jgrd.50395, 2013.
- 1500 [Wang, T., Ottlé, C., Boone, A., Ciais, P., Brun, E., Morin, S., Krinner, G., Piao, S., and Peng, S.: Evaluation of an improved intermediate complexity snow scheme in the ORCHIDEE land surface model, \*Journal of Geophysical Research: Atmospheres\*, 118\(12\), 6064–6079, 2013.](#)

Wang, F., Ducharme, A., Cheruy, F., Lo, M.-H. and Grandpeix, J.-Y.: [Impact of a shallow groundwater table on the global water cycle in the IPSL land-atmosphere coupled model](#), *Climate Dynamics*, 50(9-10), 3505–3522, doi:10.1007/s00382-017-3820-9, 2018.

Whitley, R., Beringer, J., Hutley, L. B., Abramowitz, G., Kauwe, M. G. D., Evans, B., Haverd, V., Li, L., Moore, C., Ryu, Y., Scheiter, S., Schymanski, S. J., Smith, B., Wang, Y.-P., Williams, M. and Yu, Q.: Challenges and opportunities in land surface modelling of savanna ecosystems, *Biogeosciences*, 14(20), 4711–4732, doi:10.5194/bg-14-4711-2017, 2017.

Whitley, R., Beringer, J., Hutley, L. B., Abramowitz, G., Kauwe, M. G. D., Duursma, R., Evans, B., Haverd, V., Li, L., Ryu, Y., Smith, B., Wang, Y.-P., Williams, M. and Yu, Q.: A model inter-comparison study to examine limiting factors in modelling Australian tropical savannas, *Biogeosciences*, 13(11), 3245–3265, doi:10.5194/bg-13-3245-2016, 2016.

Zhou, S., Yu, B., Zhang, Y., Huang, Y. and Wang, G.: [Partitioning evapotranspiration based on the concept of underlying water use efficiency](#), *Water Resources Research*, 52(2), 1160–1175, doi:10.1002/2015wr017766, 2016.

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.

1535

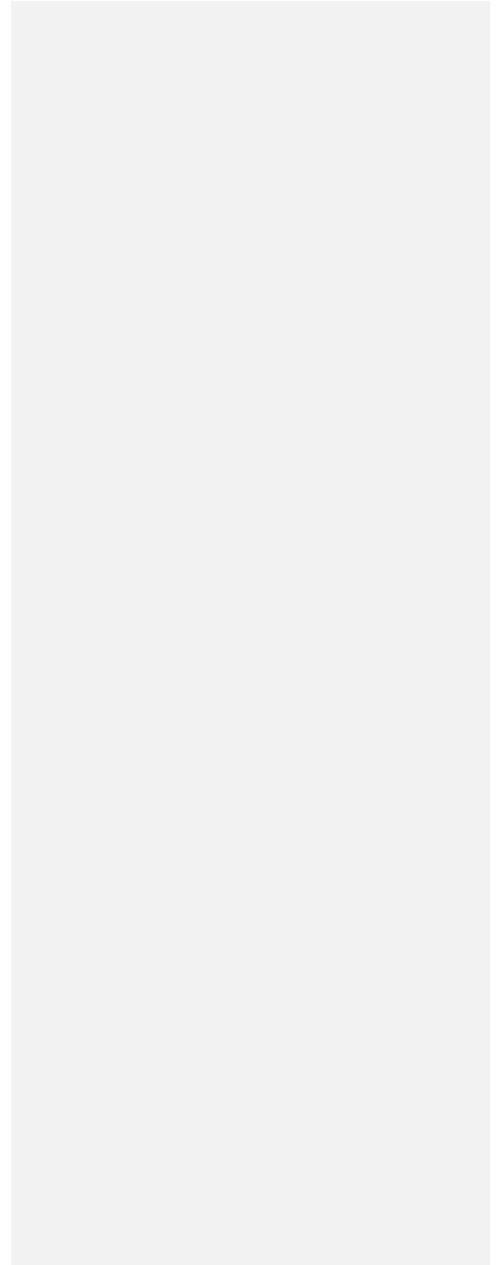


Table 1: Site descriptions, period of available site data, and associated ORCHIDEE model parameters, including vegetation plant functional type (PFT), soil texture fractions and maximum LAI used in ORCHIDEE model simulations (also see Table 1 for general site descriptions). Simulation period correspond to the period of available site data. PFT fractional cover and the maximum LAI for each soil texture class are defined in ORCHIDEE by the user. Note that ORCHIDEE does not contain an explicit representation of shrub PFTs; therefore, shrubs were included in the forest PFT. The maximum LAI has a default setting in ORCHIDEE that has not been used here; instead, values based on the site literature have been prescribed in the model. The USDA soil texture classification (12 classes – see section 2.3.3 for description) is used to define hydraulic parameters in the 2-layer mechanistic hydrology scheme (see Section 2.2.2 and 2.2.3 for a description). For some sites, soil texture fractions are taken from the ancillary Ameriflux BADM (Biological, Ancillary, Disturbance and Metadata) Data Product BIF (BADM Interchange Format) files (<https://ameriflux.lbl.gov/data/aboutdata/badm-data-product/>) that are downloaded with the site data. PFT acronyms: BS = Bare soil; TeNE = Temperate Needleleaved Evergreen forest; TeBE = Temperate Broadleaved Evergreen forest; TeBD = Temperate Broadleaved Deciduous forest; C3G = C3 grass; C4G = C4 grass.

Site ID	Description	Dominant Species	Soil texture	Period of site data	PFT fractions	Soil texture class fractions	Maximum LAI	Reference
US-SRM	Shrub encroached C4 grassland / savanna	<i>Prosopis velutina</i> , <i>Eragrostis lehmanniana</i>	Deep loamy sands	2004-2015	50% BS; 35% TeBD; 15% C4G	USDA: Loamy sand	0.85 (TeBD & C4G)	Scott et al. (2015); Ameriflux BADM.
US-SRG	C4 grassland	<i>Eragrostis lehmanniana</i>	Deep loamy sands	2008-2015	45% BS; 11% TeBD; 44% C4G	USDA: Loamy sand	1.0 (C4G)	Scott et al. (2015)
US-Whs	Shrub-dominated shrubland	<i>Larrea tridentata</i> , <i>Parthenium incanum</i> , <i>Acacia constricta</i> , <i>Rhus microphylla</i>	Gravelly sandy loams	2007-2015	57% BS; 40% TeBE; 3% C4G	USDA: Sandy loam	0.6 (TeBE & C4G)	Scott et al. (2015)
US-Wkg	C4 grassland	<i>Eragrostis lehmanniana</i> , <i>Bouteloua</i> spp. <i>Calliandra eriophylla</i>	Very gravelly, sandy to fine sandy, and clayey loams	2004-2015	60% BS; 3% TeBE; 37% C4G	USDA: Sandy loam	0.85 (C4G)	Scott et al. (2015); Ameriflux BADM.
US-Fuf	Unmanaged ponderosa pine forest	<i>Pinus ponderosa</i>	Clay loam	2005-2010	100% TeNE	USDA: Clay loam	2.4	Dore et al. (2012); Ameriflux BADM.



US-Vcp	Unmanaged ponderosa pine forest	<i>Pinus ponderosa</i>	Silt loam	2007-2014	100% TeNE	USDA: <del>Silt loam</del>	2.4	Anderson-Teixeira et al. (2011)
--------	---------------------------------	------------------------	-----------	-----------	-----------	----------------------------	-----	---------------------------------

Deleted: 100% s

**Table 2: Summary of differences between 2LAY and 1LAY model versions. All other parameters and processes in the model, including the PFT and soil texture fractions (Table 1), the vegetation and bare soil albedo coefficients (Section 2.2.1), and the multi-layer intermediate complexity snow scheme (Section 2.2.5) are the same in both versions.**

<b>Model Process</b>	<b>Model Version</b>	
	<b><u>2LAY</u></b>	<b><u>1LAY</u></b>
<b><u>Soil Moisture</u></b> (Section 2.2.2)	2-layer bucket scheme – upper layer variable to 10cm depth and can disappear	1D Richards equation describing moisture diffusion in unsaturated soils
<b><u>Maximum water holding (field) capacity</u></b> Section 2.2.2)	Constant (150kgm <sup>-2</sup> ) for all soil types	Derived using Van Genuchten (VG) relationships for characteristic matric potentials and vary with soil texture
<b><u>Runoff/Drainage</u></b> (Section 2.2.2)	When soil moisture exceeds field capacity 5% partitioned as surface runoff and 95% as groundwater drainage	Calculated soil hydraulic conductivity determines precipitation partitioning into infiltration and runoff. Drainage in form of free gravitational flow at bottom of soil.
<b><u>Bare soil evaporation resistance</u></b> (Section 2.2.3)	Based on depth of dry soil for bare soil PFT. Not optional – included by default	Empirical equation based on relative water content of the 1 <sup>st</sup> four layers. Optional – not included by default
<b><u>Empirical plant water stress function, <math>\beta</math></u></b> (Section 2.2.4)	Based on dry soil depth of upper layer	Based on plant water availability for root water uptake throughout soil column
<b><u>E and T over vegetated grid cell fraction</u></b> (Section 2.2.1)	Only T occurs	Both T and E occur over effective vegetated and effective bare soil fraction, respectively. Calculation of effective fractions based on LAI (Beer-Lambert approach)

5

10

15

**Table 3: Soil moisture measurement depths (and corresponding model layer in brackets – see Table S1).**

Deleted: Table 2

	US-SRM	US-SRG	US-Whs	US-Wkg	US-Fuf	US-Vcp
Soil moisture depths	2.5-5cm (5)	2.5-5cm (5)	5cm (6)	5cm (6)	2cm (4)	5cm (6)
	15-20cm (7)	15-20cm (7)	15cm (7)	15cm (7)	20cm (8)	20cm (7)
	60-70cm (9)	75cm (9)	30cm (8)	30cm (8)	50cm (9)	50cm (9)

5 **Table 4: Model evaluation metrics comparing the 2LAY and 11LAY daily upper layer soil moisture (re-scaled via linear CDF matching) and daily ET simulations to observations across the whole timeseries (where data present – see Fig. S2). Metrics include: correlation coefficient (R), root mean squared error (RMSE), mean absolute bias, and a measure of the relative variability,  $\alpha$ , between the model and the observations. The mean absolute bias = model – observations; therefore, a negative value represents a mean model underestimation of observed ET.  $\alpha = \frac{\sigma_m}{\sigma_o}$  (see section 2.4, with ‘ideal’ values approaching 1).**

Deleted: Table 3

Deleted: Comparison

Deleted: of

Deleted: model-data evaluation metrics

Deleted: 3

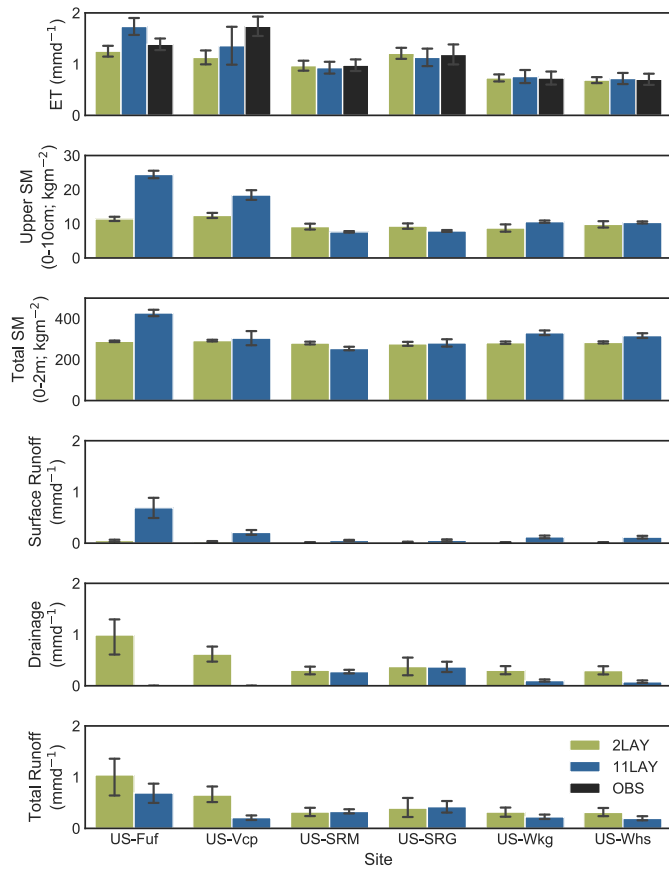
Deleted: . M

Formatted Table

Site	Model Version	Upper Layer (0-10cm) Soil Moisture R	ET R	ET RMSE (mmd <sup>-1</sup> )	ET Mean Bias (mmd <sup>-1</sup> )	ET relative variability, $\alpha$
US-Fuf	2LAY	0.30	0.36	1.04	-0.08	1.08
	11LAY	0.78	0.76	0.86	0.38	1.33
US-Vcp	2LAY	0.27	0.26	1.39	-0.54	0.79
	11LAY	0.37	0.59	1.02	-0.27	0.82
US-SRM	2LAY	0.52	0.53	0.84	-0.03	0.70
	11LAY	0.85	0.84	0.53	-0.07	0.87
US-Whs	2LAY	0.56	0.54	0.68	-0.03	0.67
	11LAY	0.90	0.85	0.43	-0.02	0.89
US-SRG	2LAY	0.48	0.52	1.02	0.01	0.70
	11LAY	0.67	0.88	0.57	-0.11	0.90
US-Wkg	2LAY	0.46	0.62	0.63	0	0.71
	11LAY	0.76	0.9	0.37	-0.01	1.07

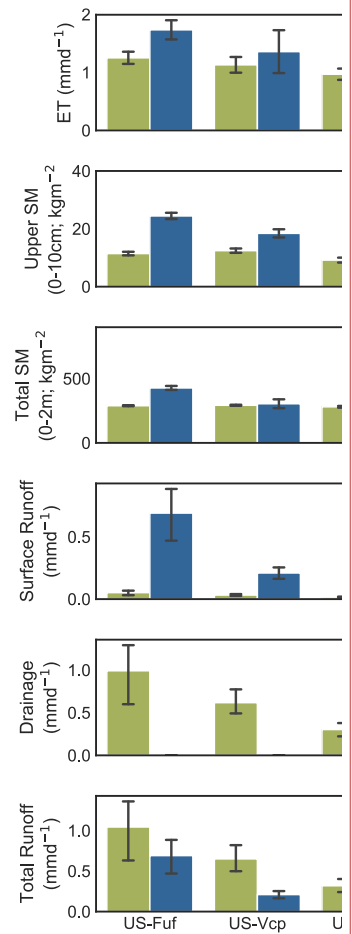
10

Figure 1: Comparison of the 2LAY versus 11LAY mean daily hydrological stores and fluxes: i) evapotranspiration (ET,  $\text{mmd}^{-1}$  – top panel); ii) **total** soil moisture (SM,  $\text{kgm}^{-2}$ ) in the upper 10cm of the soil (2<sup>nd</sup> panel); iii) total column (0-2m) SM (3<sup>rd</sup> panel); iv) surface runoff ( $\text{mmd}^{-1}$ , 4<sup>th</sup> panel); v) drainage ( $\text{mmd}^{-1}$ , 5<sup>th</sup> panel); and vi) total runoff (surface runoff plus drainage – bottom panel). Error bars show the standard deviation for ET and SM, and 95% confidence interval for runoff and drainage. **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.**



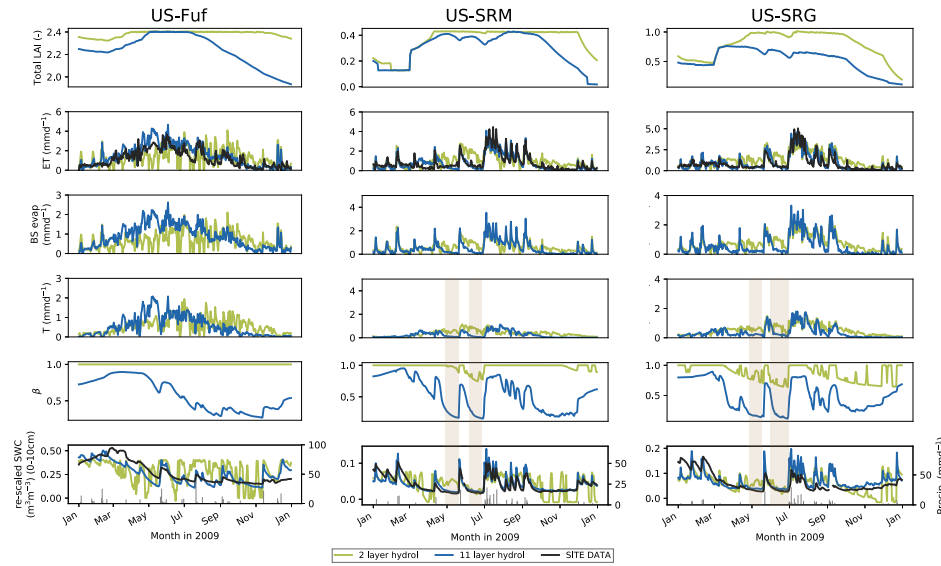
Formatted: Superscript

Formatted: Caption, Centered

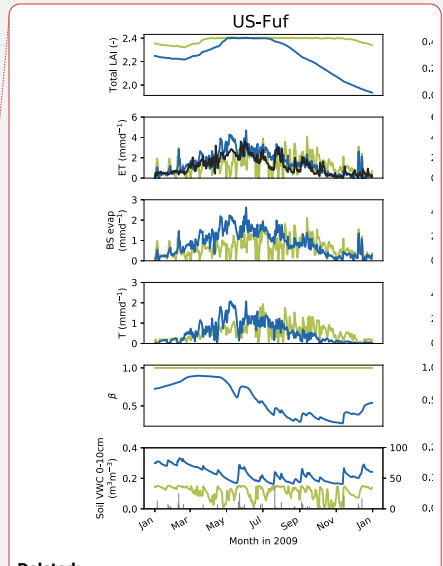


Deleted:

Figure 2: Comparison of daily time series (for 2009) of upper layer soil moisture, surface water fluxes and related variables between the 2LAY (green curve) and 11LAY (blue curve) simulations. Changes between the two versions are shown for three sites representing the main vegetation types: left column = high-elevation tree-dominated site (US-Fuf); middle column = low-elevation mesquite shrub-dominated site (US-SRM); right column = low-elevation C4 grass site (US-SRG). At each site, top panel: LAI; 2<sup>nd</sup> panel: ET compared to observations (black curve); 3<sup>rd</sup> panel: bare soil evaporation; 4<sup>th</sup> panel: transpiration; 5<sup>th</sup> panel: empirical water limitation function ( $\beta$ ) that scales photosynthesis and stomatal conductance; bottom panel: model soil moisture (re-scaled via linear CDF matching) expressed as volumetric water content (VWC) in the uppermost 10cm of the soil compared to observations (black curve). Precipitation is shown in the grey lines in the bottom panel for each site. (Note: full time series across all years are shown for all site in Figs. S2a-f).



Deleted: D  
 Deleted: of variables influencing changes in  
 Deleted: ET  
 Deleted: transpiration;  
 Deleted: bare soil evaporation



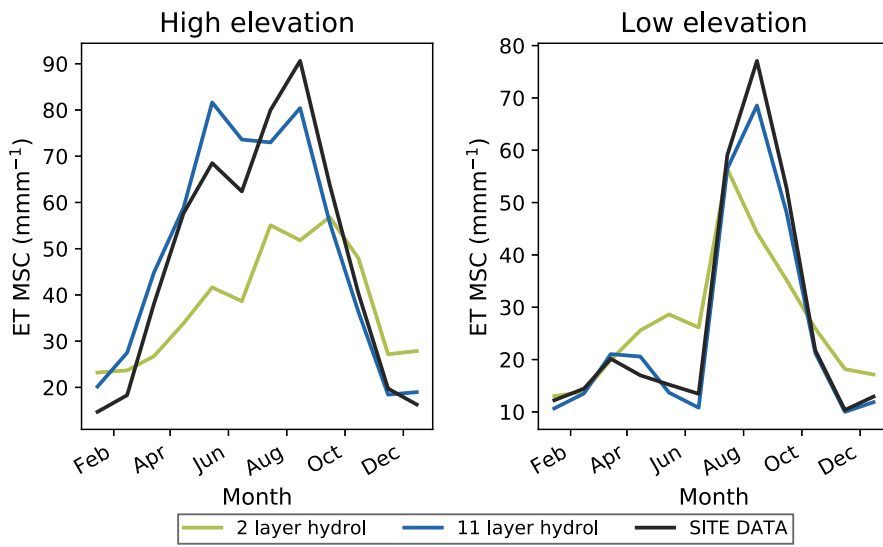
5

10

15

20

Figure 3: Evapotranspiration (ET) monthly mean seasonal cycle comparing the 2LAY (green curve) and 11LAY (blue curve) simulations with observations (black curve). Individual site simulations have been averaged over the high-elevation tree dominated sites (left panel) and across all the low-elevation grass- and shrub-dominated sites (right panel). Units are mm per month (mmm<sup>-1</sup>).

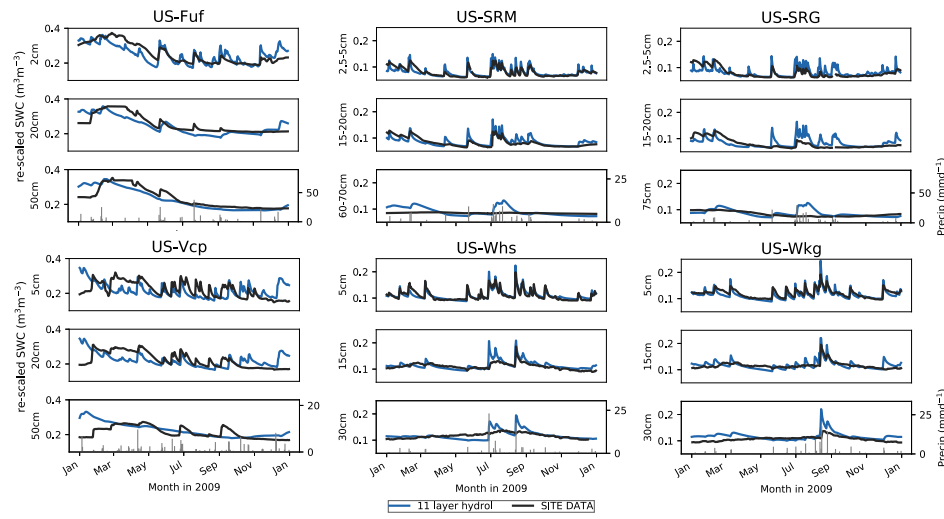


5

10

15

Figure 4: Daily simulated volumetric soil water content (SWC -  $m^3m^{-3}$ ) in 2009 (re-scaled via linear CDF matching) compared to observations at each site for three depths (upper, middle, lower) in the soil profile. The soil depths and their corresponding model layers are given in Table 3. Precipitation is shown in the grey lines in the bottom panel for each site.



Deleted: VWC

Deleted: compared to

Deleted: (

Deleted: Table 2

Deleted:

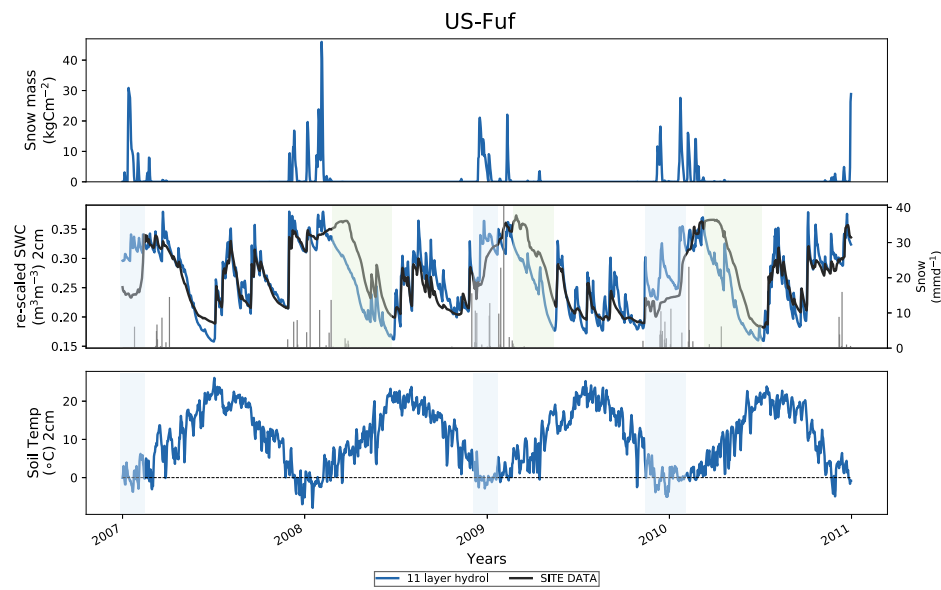
5

10

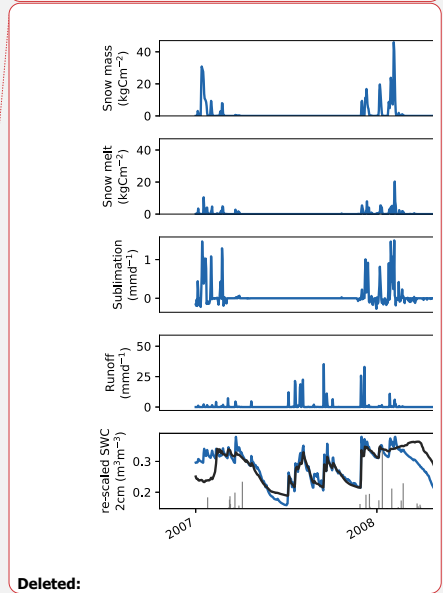
15

Figure 5: a) US-Fuf and b) US-Vcp 11LAY (blue curve) daily time series (2007-2010) of model (re-scaled via linear CDF matching) versus observed volumetric soil water content (middle panel SWC -  $m^3m^{-3}$ ) (black curve), compared to simulated snow mass (top panel) and soil temperature from the corresponding 2cm soil thermal layer (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.

a) US-Fuf



Deleted: versus  
 Deleted: (  
 Deleted: runoff, sublimation, snow melt and  
 Deleted: bottom  
 Deleted: panel



Deleted:

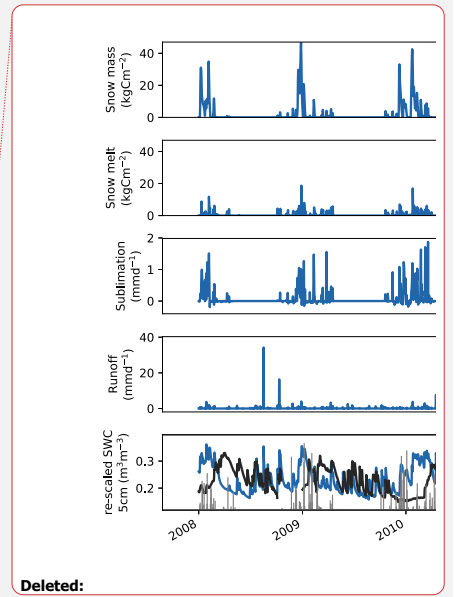
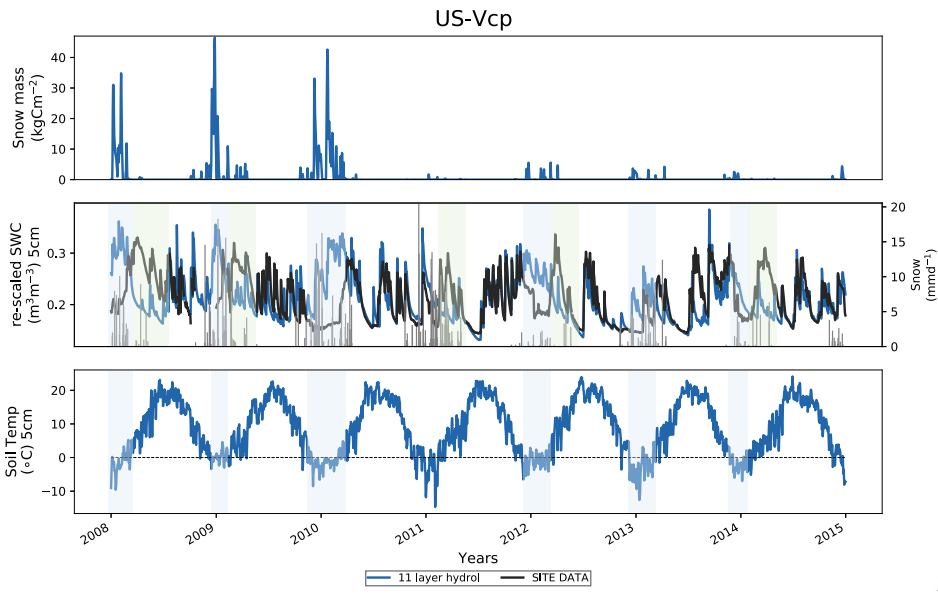
5

10

15



b) US-Vcp

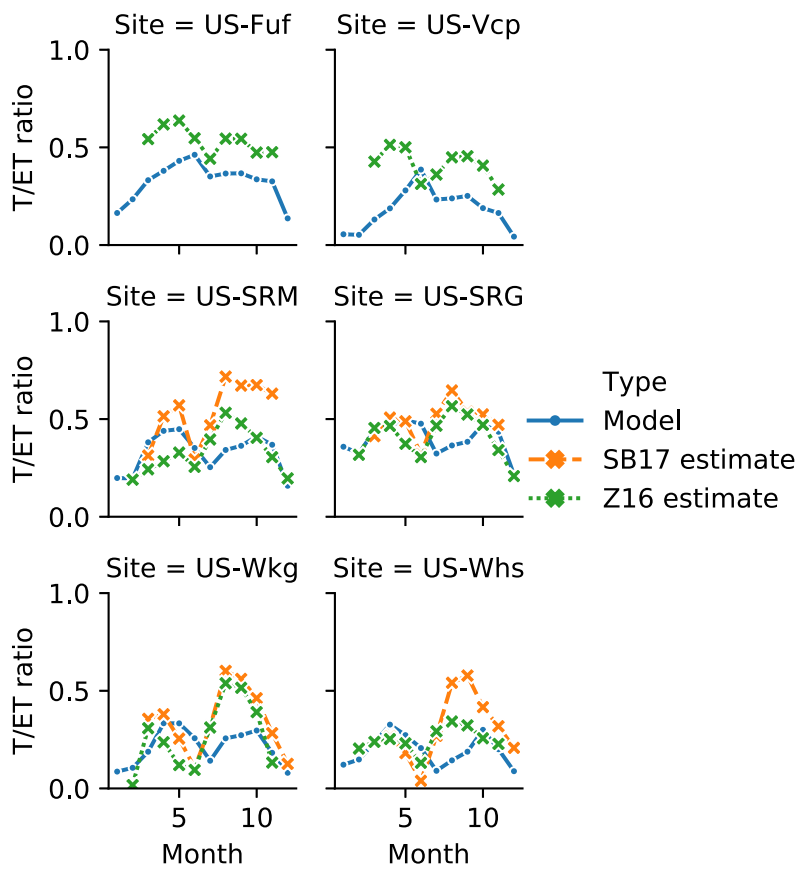


5

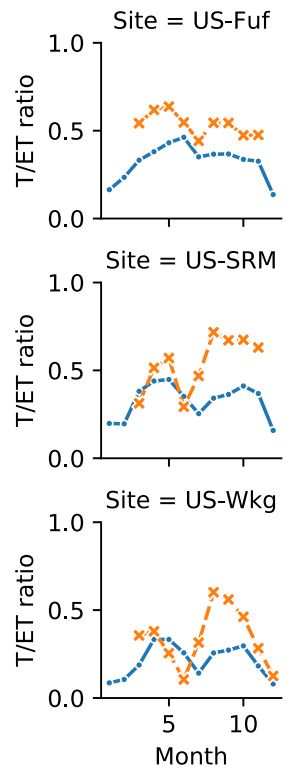
10

15

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.

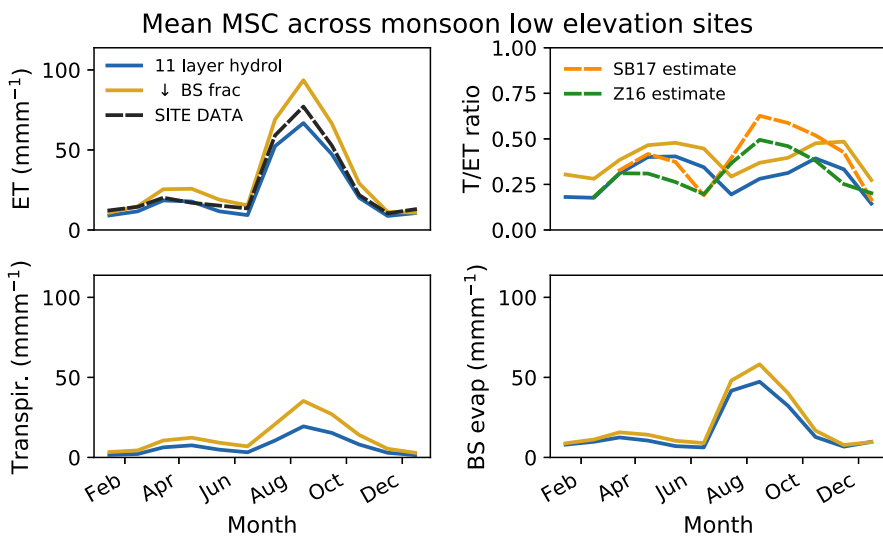


Deleted: ¶  
 Deleted: M  
 Deleted: Blue curves show the model estimates at each site; orange curves show T/ET ratios at sites/months for which data-driven estimates are available.  
 Deleted: ratios  
 Deleted: ratios  
 Deleted: derived following  
 Deleted: et al.

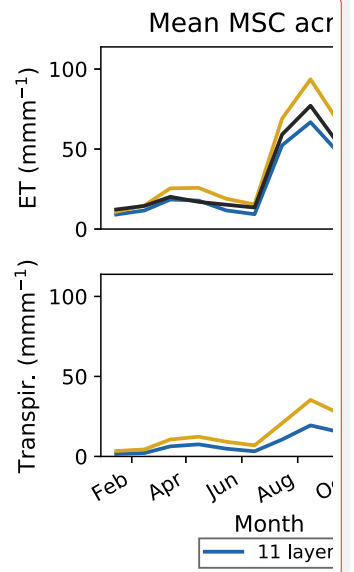


Deleted:

Figure 7: Monthly mean seasonal cycle for ET, T/ET ratios, T and E averaged across all low-elevation grass- and shrub-dominated sites comparing the default 11LAY simulations (blue curve) with a simulation in which bare soil fraction is decreased (C4 grass cover increased) (yellow curve). ET is compared to observations (black dashed curve) and T/ET ratios are compared to the data-derived estimates from Scott and Biederman (2017 – orange dashed curve) and Zhou et al. (2016 – green dashed curve). Units are mm per month ( $\text{mm}^{-1}$ ).



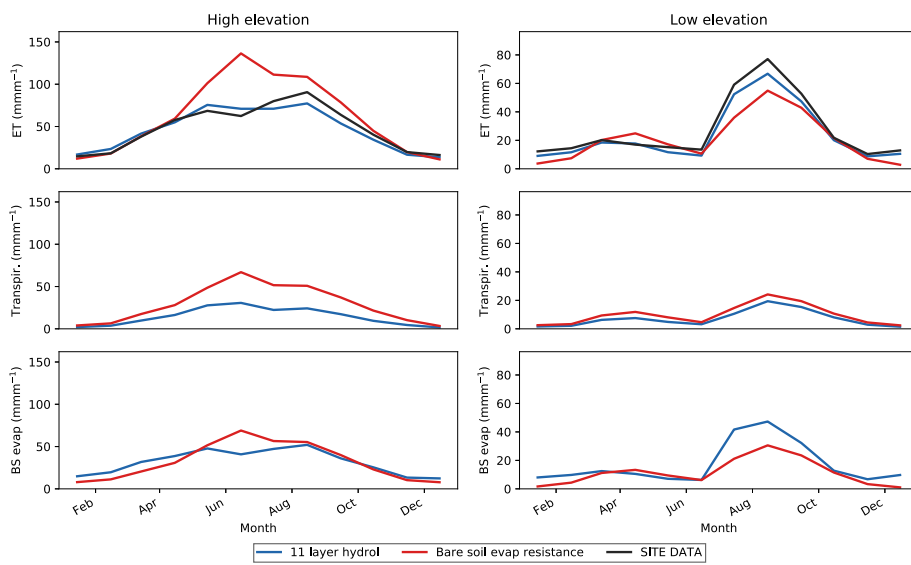
Deleted: Units in  $\text{mm}^{-1}$ .



Deleted:

Figure 8: Monthly mean seasonal cycle for evapotranspiration (ET), transpiration and bare soil evaporation averaged across all high-elevation forest sites (left column) and low-elevation monsoon grass- and shrub-dominated sites (right column) comparing the default ILLAY simulations (blue curve) with a simulation that included an additional bare soil evaporation resistance term (red curve). ET is compared to observations (black curve). Units are mm per month ( $\text{mmm}^{-1}$ ).

Deleted: Units in  $\text{mmmonth}^{-1}$ .



Supplementary Information for

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

**Deleted:** Multi-variable, multi-configuration testing of ORCHIDEE land surface model water flux and storage estimates across semi-arid sites in the southwestern US\*

5 Natasha MacBean<sup>1\*</sup>, Russell L. Scott<sup>2</sup>, Joel A. Biederman<sup>2</sup>, Catherine Ottlé<sup>3</sup>, Nicolas Vuichard<sup>3</sup>, Agnès Ducharne<sup>4</sup>, Thomas Kolb<sup>5</sup>, Sabina Dore<sup>6</sup>, Marcy Litvak<sup>7</sup>, David J.P. Moore<sup>8</sup>.

<sup>1</sup>Department of Geography, Indiana University, Bloomington, IN 47405, USA.

<sup>2</sup>Southwest Watershed Research Center, United States Agricultural Department, Agricultural Research Service, Tucson, AZ 85719, USA.

10 <sup>3</sup>Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, F-91191, France.

<sup>4</sup>UMR METIS, Sorbonne Université, CNRS, EPHE, Paris, F-75005, France

<sup>5</sup>School of Forestry, Northern Arizona University, Flagstaff, AZ, 86011, USA.

<sup>6</sup>Hydrofocus, Inc., Davis, CA, 95618, USA.

15 <sup>7</sup>Department of Biology, University of New Mexico, Albuquerque, NM, 87131, USA.

<sup>8</sup>School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, 85721, USA.

\*Correspondence to: Natasha MacBean (nlmacbean@gmail.com)

Table S1: Depths of the ORCHIDEE 11-layer discretized hydrology model

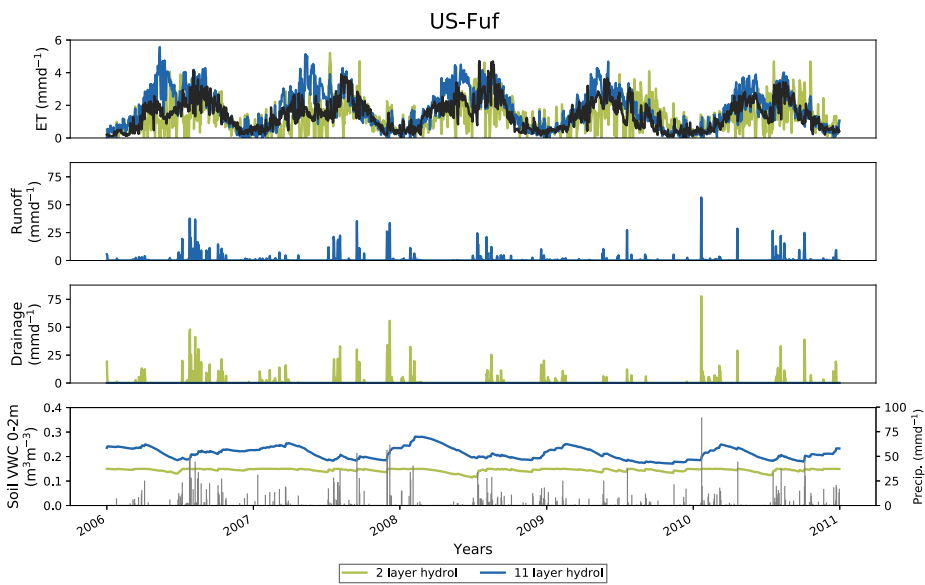
ORCHIDEE Layer	Layer thickness (m)	Cumulative Depth (m)
1	0.001	0.001
2	0.003	0.004
3	0.006	0.01
4	0.012	0.022
5	0.023	0.045
6	0.047	0.092
7	0.092	0.186
8	0.188	0.374
9	0.375	0.750
10	0.750	1.5
11	0.5	2.0

20

25

30 **Figure S1: Complete daily time series comparing the 2LAY (green curve) and 11LAY (blue curve) simulations for the following hydrological variables: i) ET (top panel for each site) compared to observations (black curve); ii) surface runoff (2<sup>nd</sup> panel for each site); iii) drainage (3<sup>rd</sup> panel for each site); and iv) total 2m column volumetric water content (VWC) soil moisture (bottom panel for each site). Precipitation is shown in the grey bars in the bottom panel for each site. Sites in following order: a) US-Fuf; b) US-Vcp; c) US-SRM; d) US-Whs; e) US-SRG; f) US-Wkg. Precipitation is shown in the grey lines in the bottom panel for each site.**

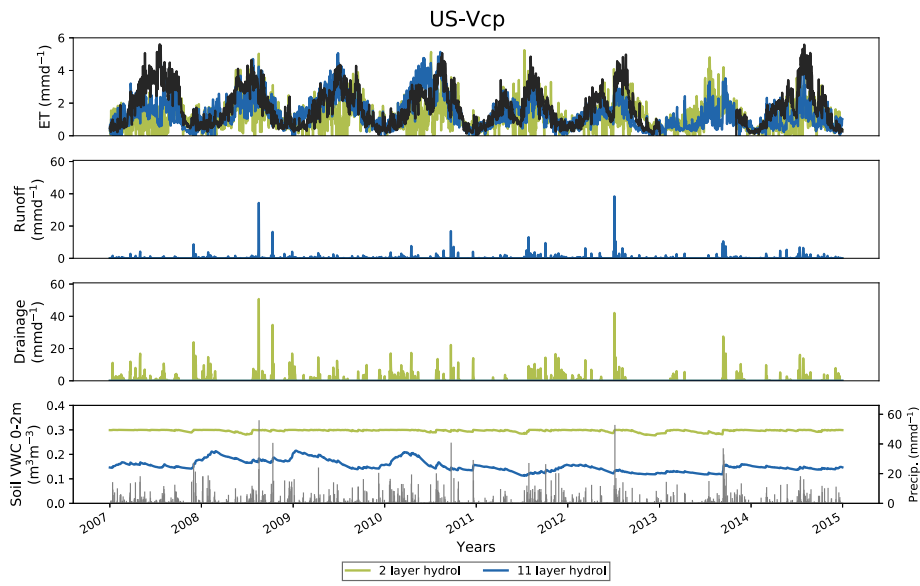
35 **a) US-Fuf**



40

45

b) US-Vcp

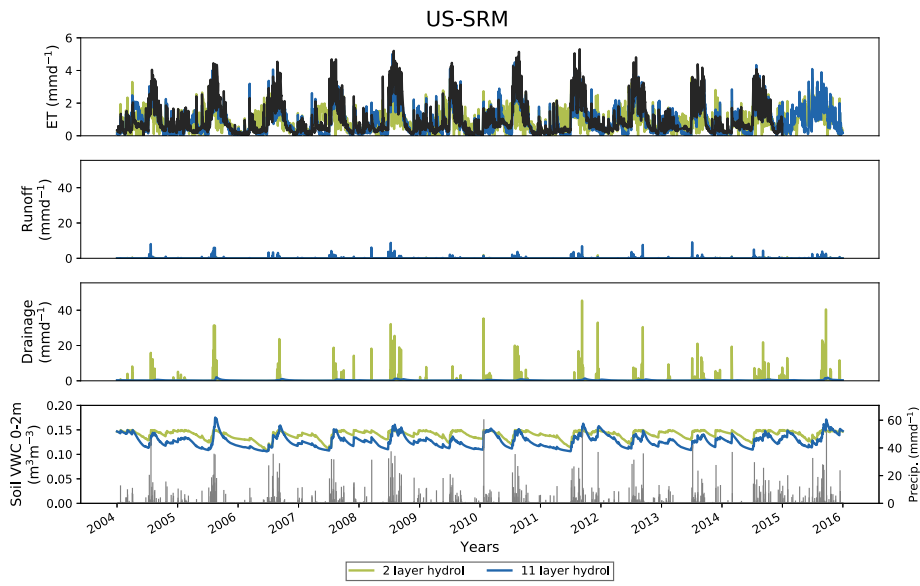


50

55

60

c) US-SRM



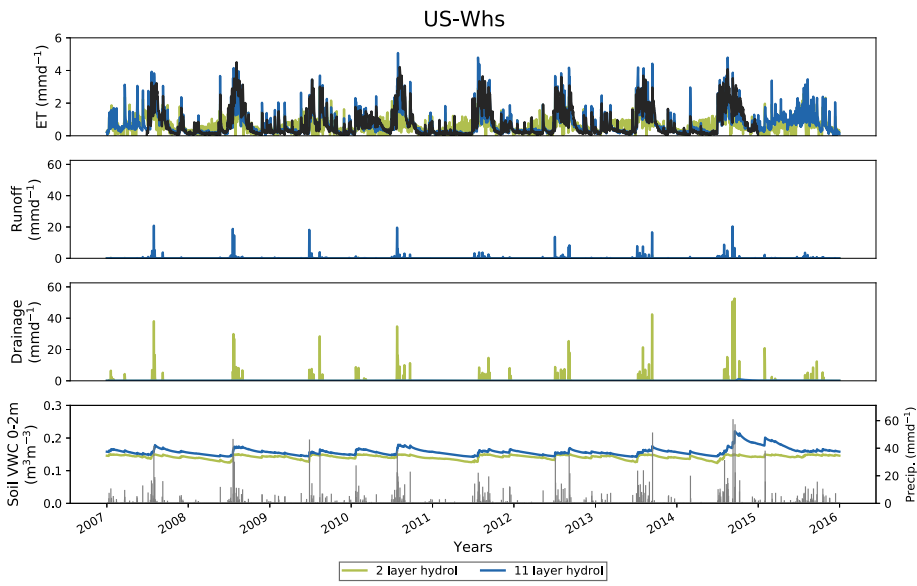
65

70

75



d) US-Whs

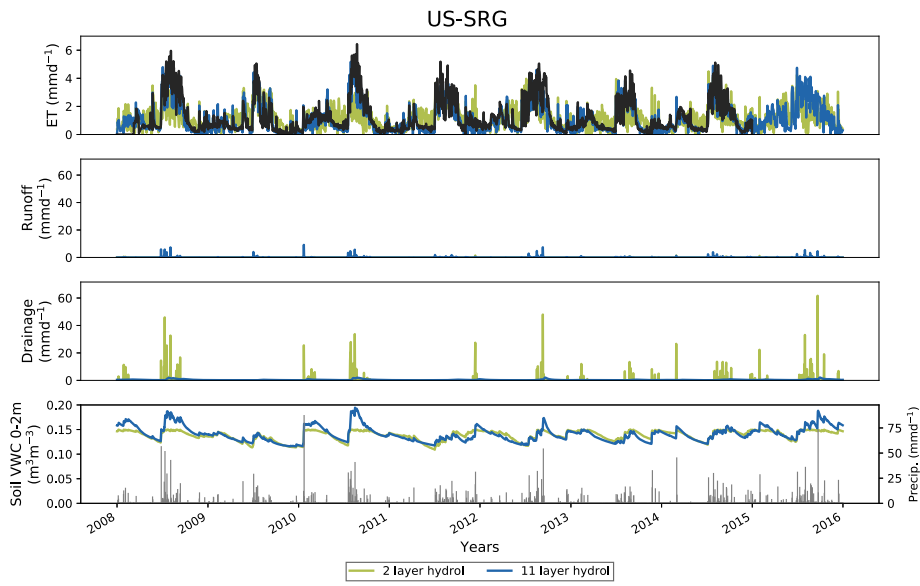


80

85

90

95 e) US-SRG

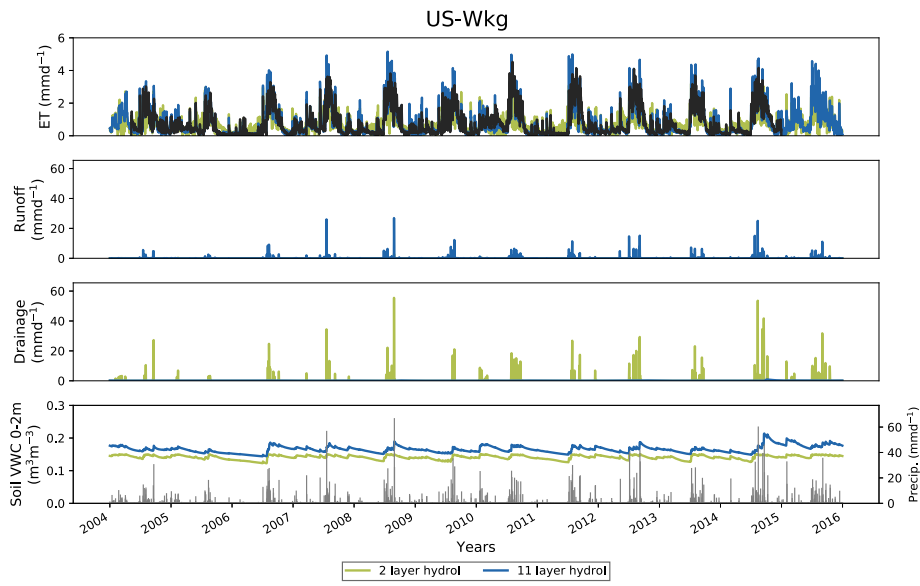


100

105

110

f) US-Wkg



115

120

125

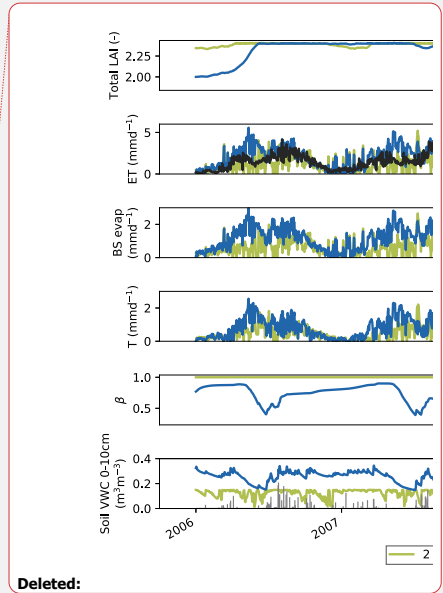
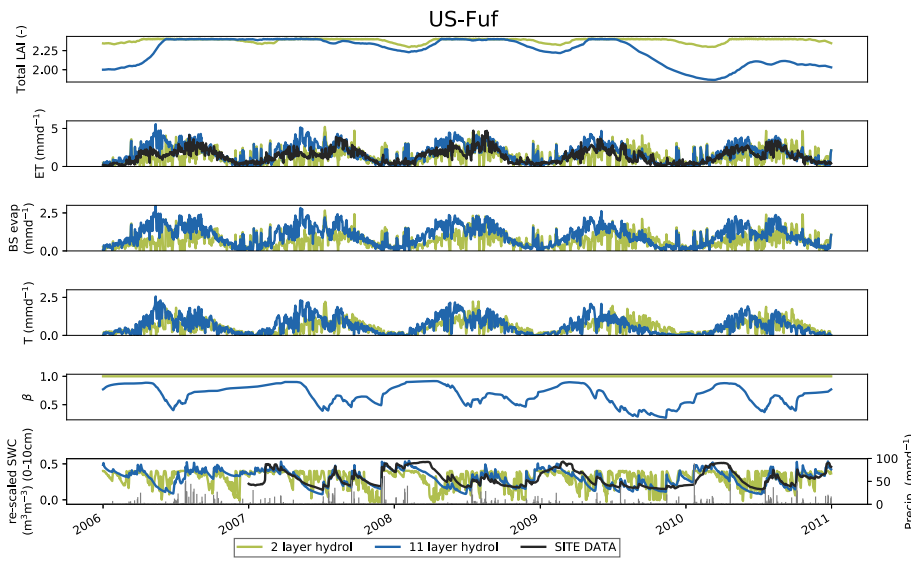
130

Figure S2: Complete daily time series of upper layer soil moisture, surface water fluxes and related variables between the 2LAY (green curve) and 11LAY (blue curve) simulations for all sites – equivalent to Fig. 2. At each site, top panel: LAI; 2<sup>nd</sup> panel: ET compared to observations (black curve); 3<sup>rd</sup> panel: bare soil evaporation; 4<sup>th</sup> panel: transpiration; 5<sup>th</sup> panel: empirical water limitation function ( $\beta$ ) that scales photosynthesis and stomatal conductance; bottom panel: model soil moisture (re-scaled via linear CDF matching) expressed as volumetric water content (VWC) in the uppermost 10cm of the soil compared to observations (black curve). Precipitation is shown in the grey bars in the bottom panel for each site. Sites in following order: a) US-Fuf; b) US-Vcp; c) US-SRM; d) US-Whs; e) US-SRG; f) US-Wkg. Precipitation is shown in the grey lines in the bottom panel for each site.

**Deleted:** variables influencing changes in ET  
**Deleted:** at three sites: left column = high elevation tree-dominated site (US-Fuf); middle column = low elevation mesquite shrub-dominated site (US-SRM); right column = low elevation C4 grass site (US-SRG)  
**Deleted:** transpiration  
**Deleted:** bare soil evaporation

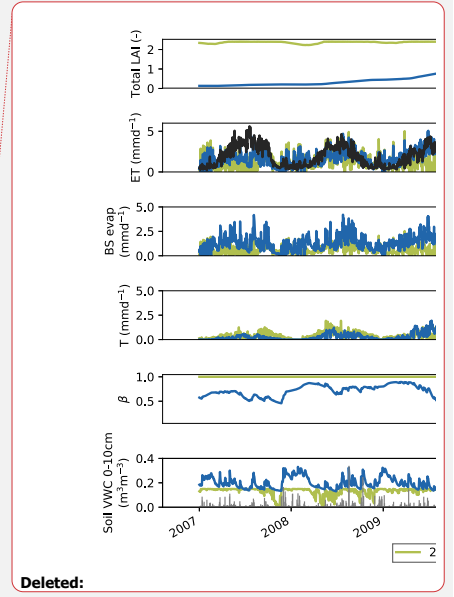
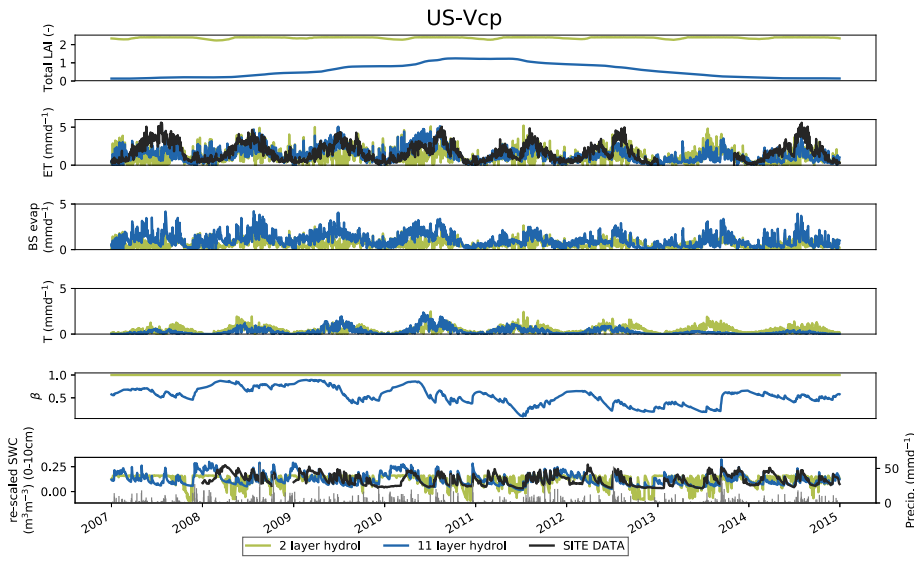
135

a) US-Fuf



140

b) US-Vcp

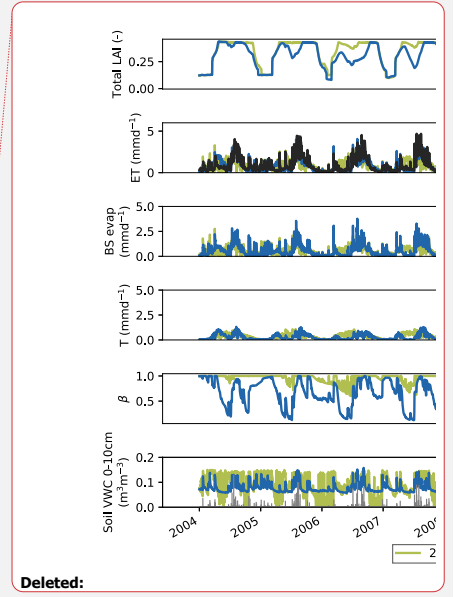
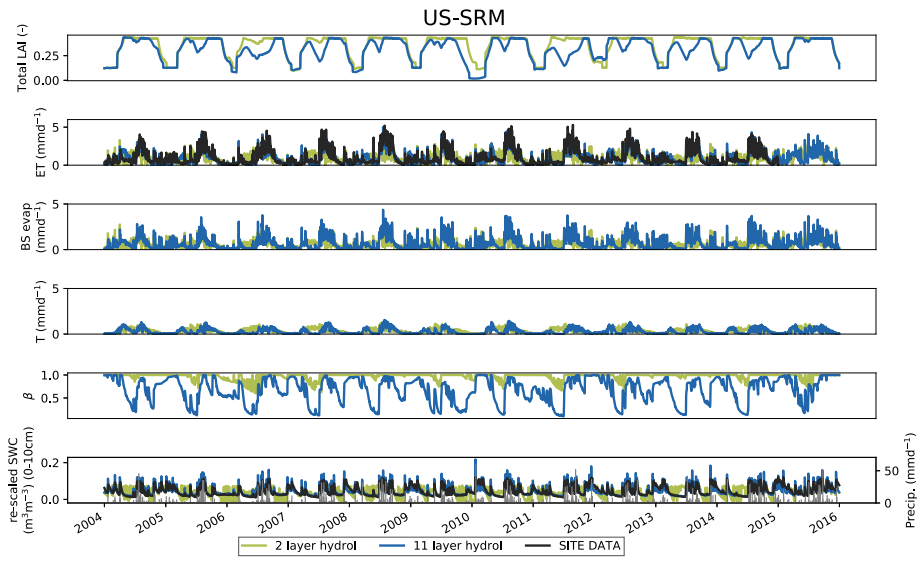


155

160

165

c) US-SRM

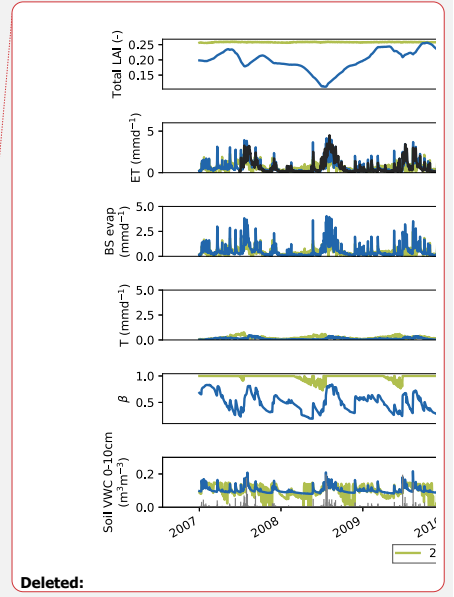
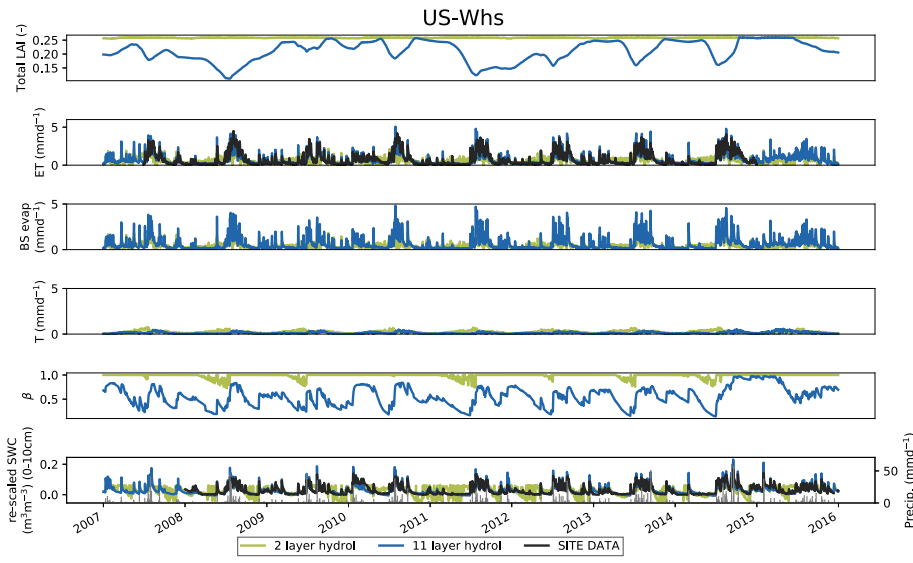


170

175

180

d) US-Whs

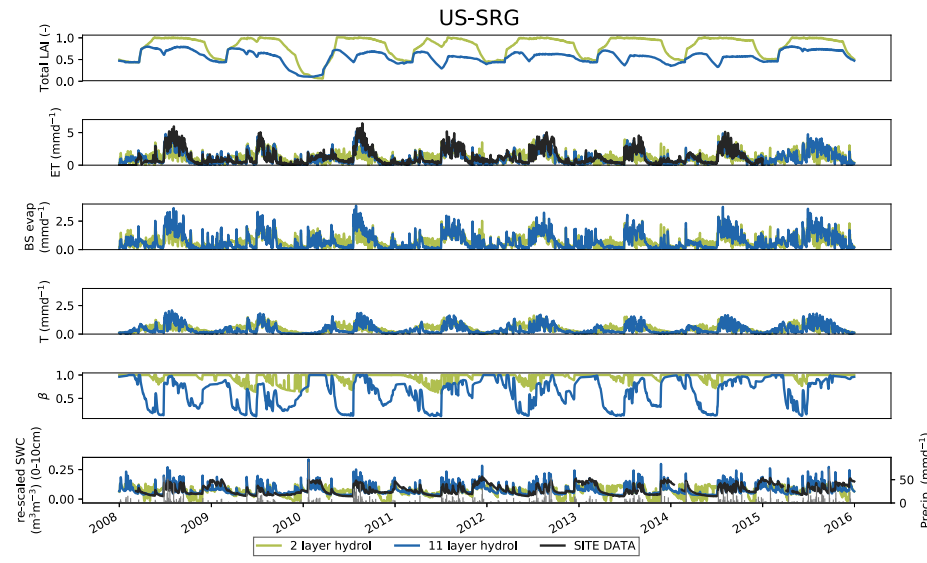


185

190

195

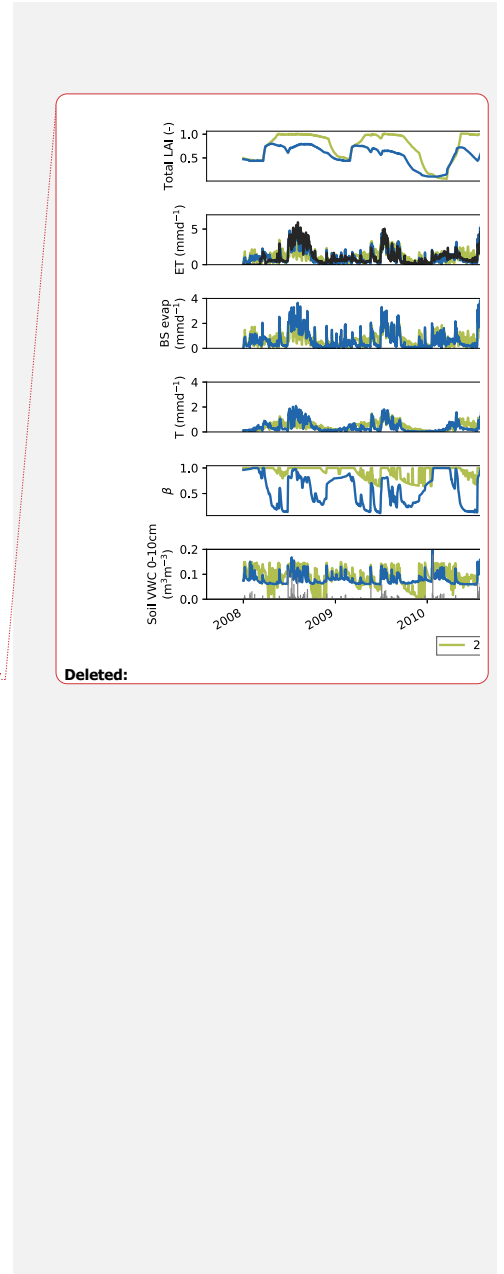
e) US-SRG



200

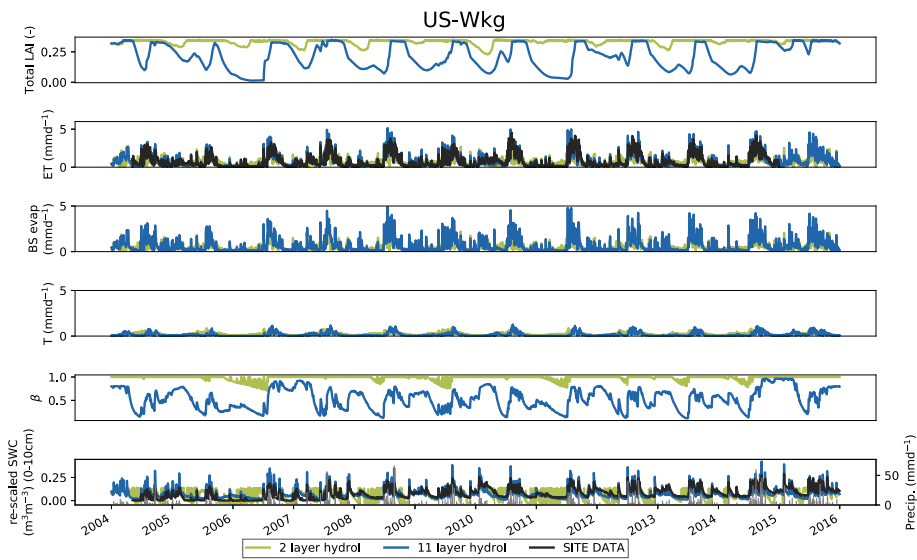
205

210





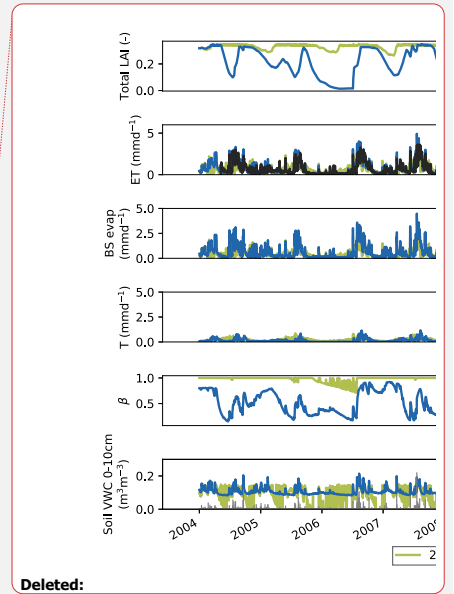
f) US-Wkg



215

220

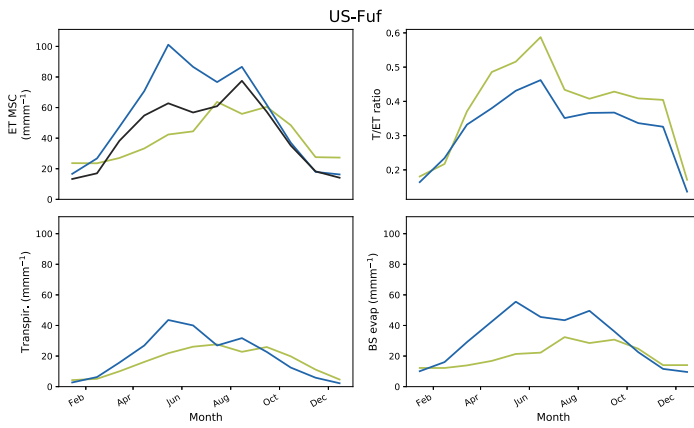
225



230

Figure S3: Monthly mean seasonal cycle for each site comparing the 2LAY (green curve) and 11LAY simulations (blue curve) with observations (black curve). Top left: ET; top right: T/ET ratios; bottom left: transpiration; bottom right: bare soil evaporation. Units in  $\text{mm}^{-1}$ . Sites in following order: a) US-Fuf; b) US-Vep; c) US-SRM; d) US-Whs; e) US-SRG; f) US-Wkg. Units are mm per month ( $\text{mm}^{-1}$ ).

a) US-Fuf

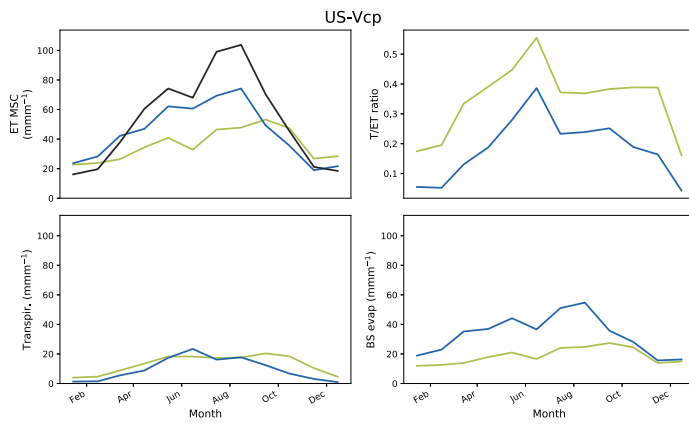


235

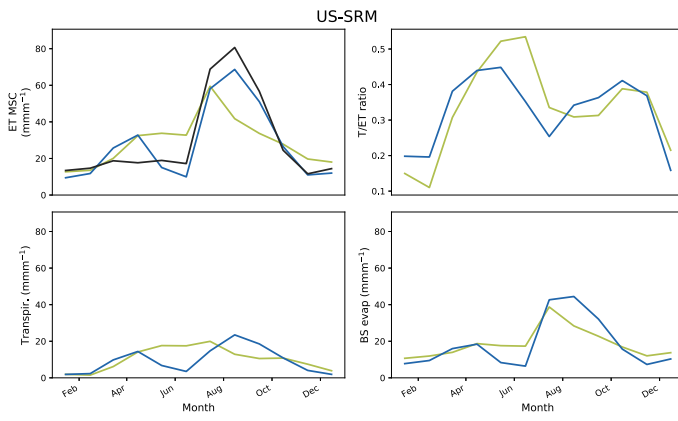
240

245

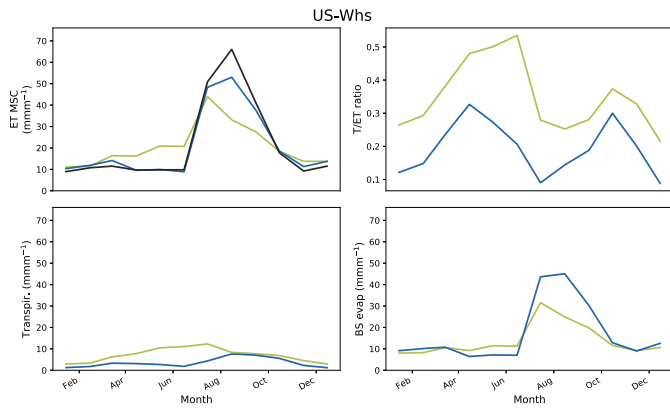
b) US-Vcp



c) US-SRM

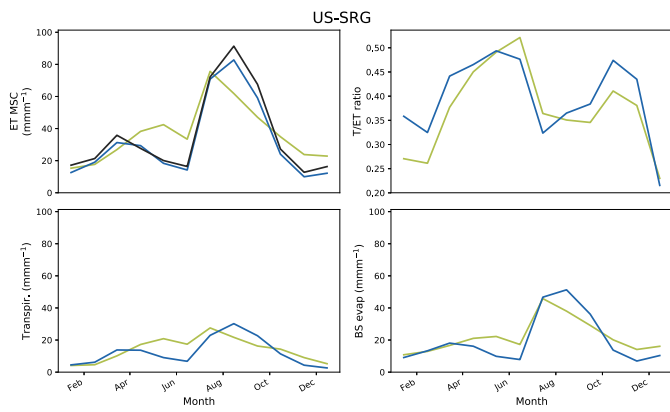


d) US-Whs



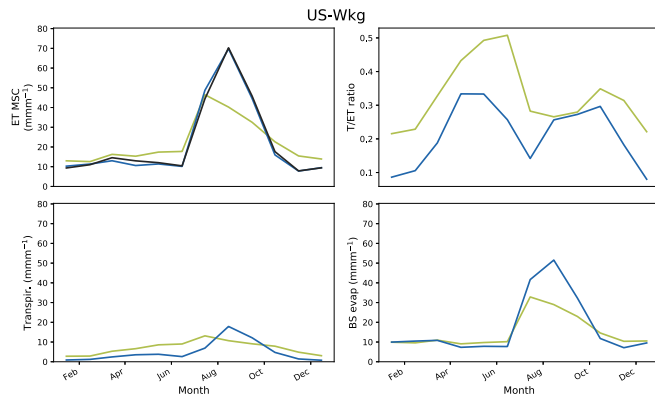
260

e) US-SRG



265

f) US-Wkg



270

275

280

285

290

Figure S4: Daily simulated volumetric soil water content (VWC –  $m^3m^{-3}$ ) across all site years (re-scaled via linear CDF matching) compared to observations at each site for three depths (upper, middle, lower) in the soil profile – equivalent to Fig. 4. The soil depths and their corresponding model layers are given in Table 3. Precipitation is shown in the grey lines in the bottom panel for each site.

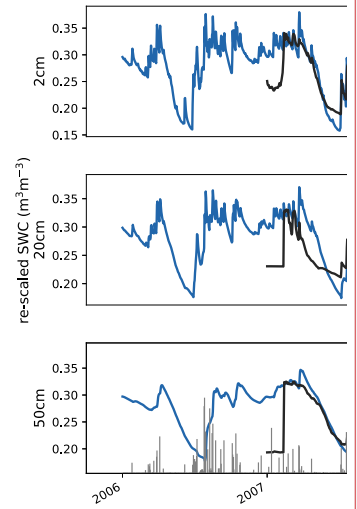
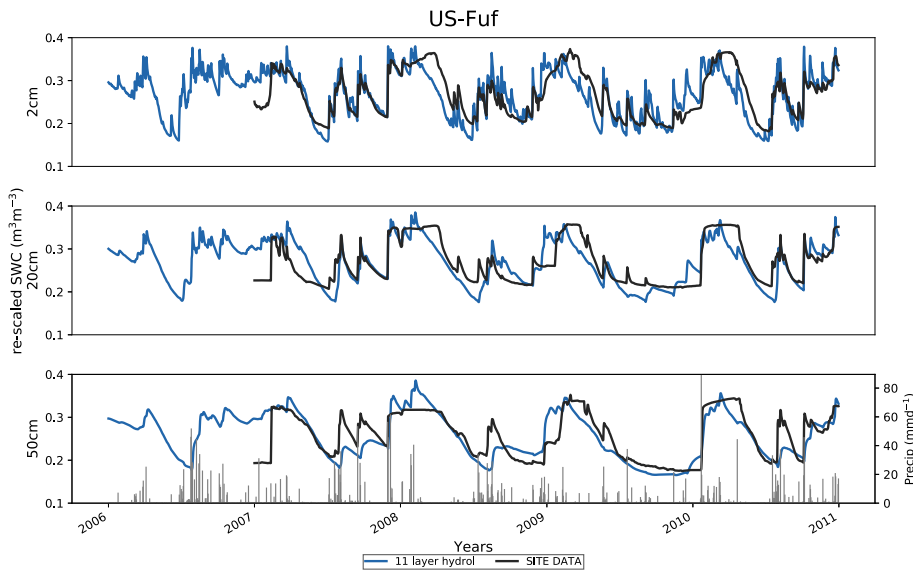
Deleted: compared to

Deleted:

Deleted: (

Deleted: Table 2

a) US-Fuf

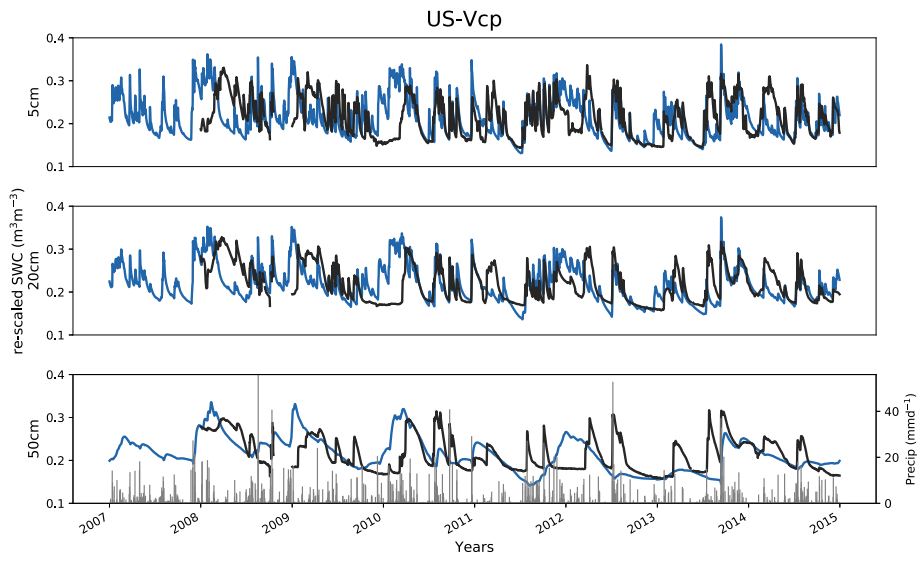


Deleted:

295

300

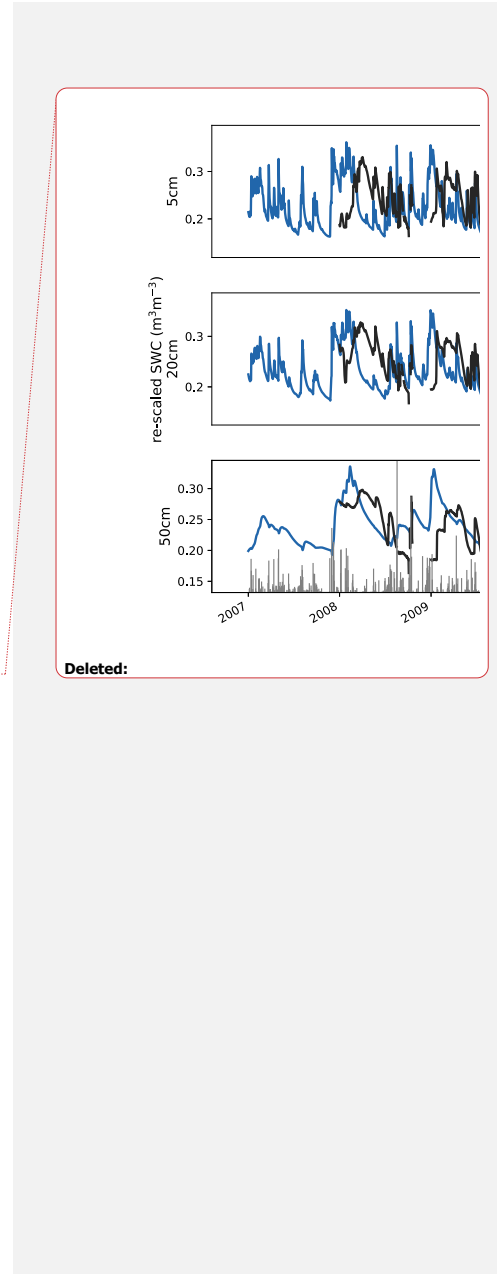
b) US-Vcp



310

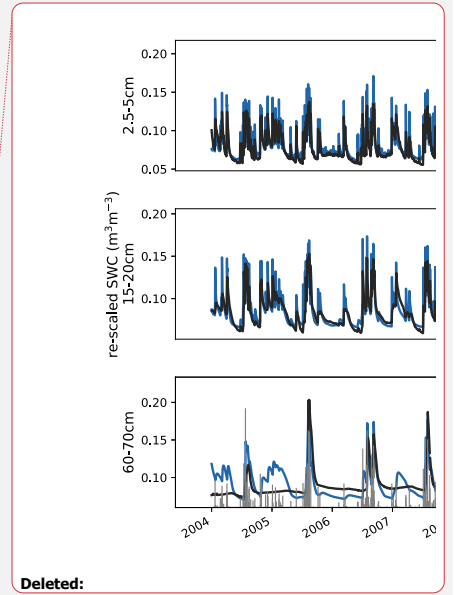
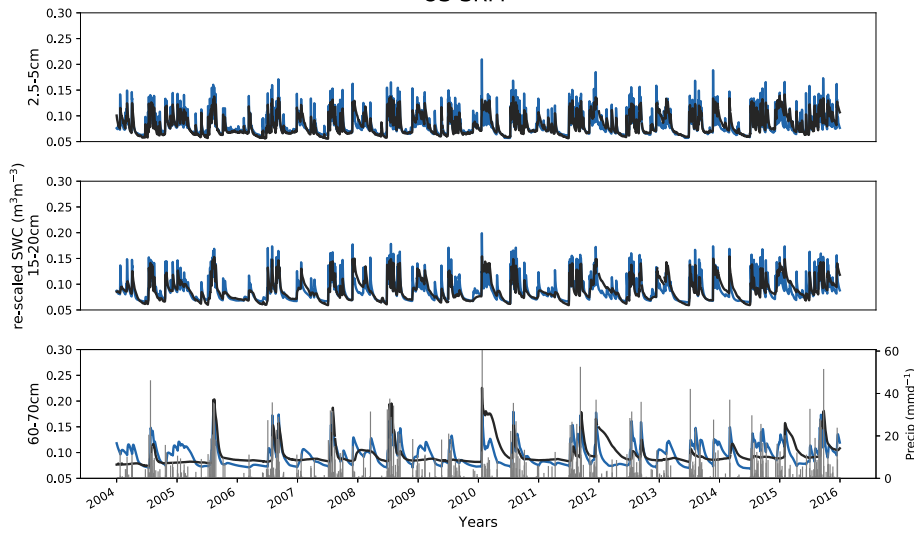
315

320



c) US-SRM

US-SRM



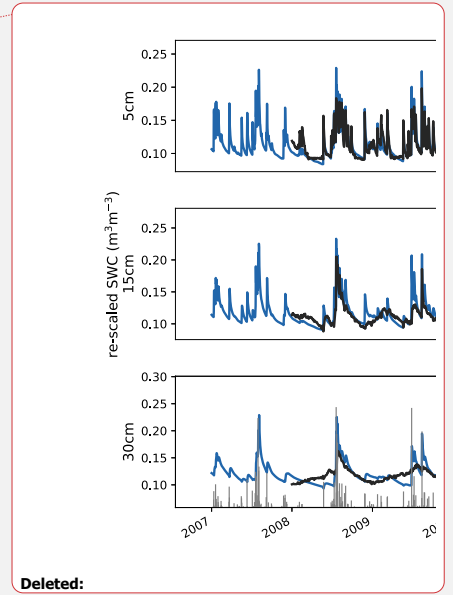
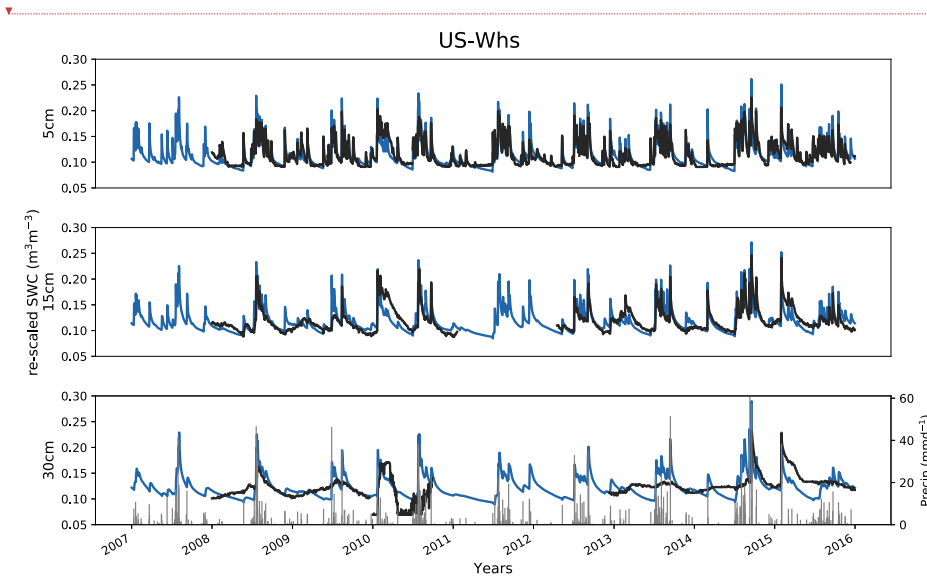
330

335

340



d) US-Whs



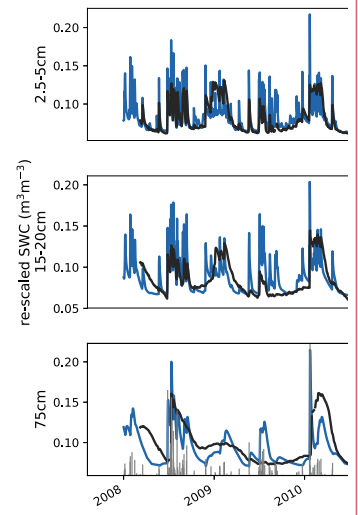
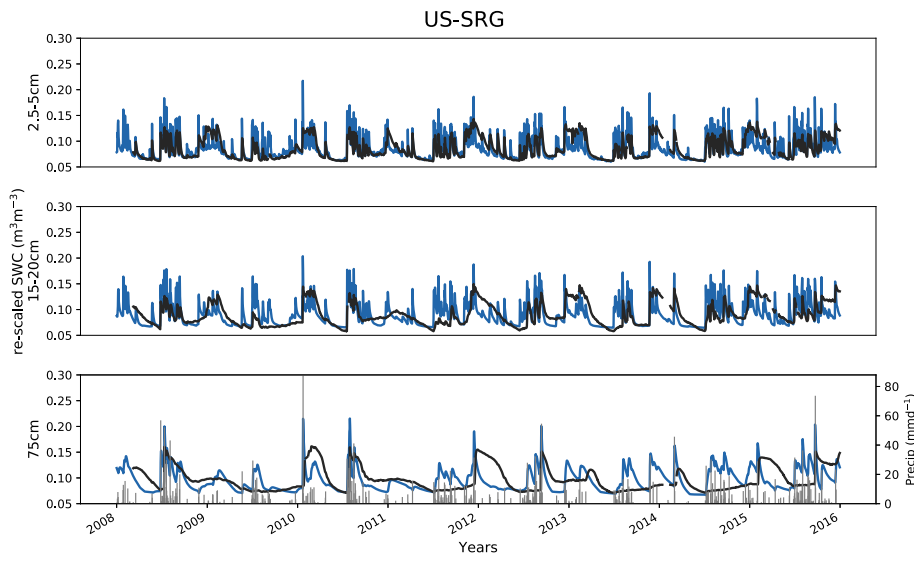
345

350

355

360

e) US-SRG



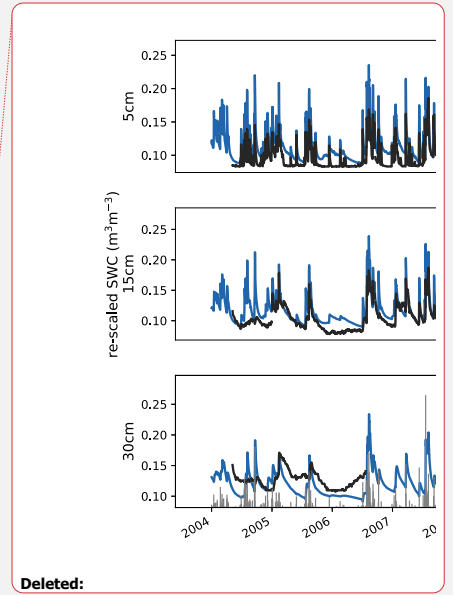
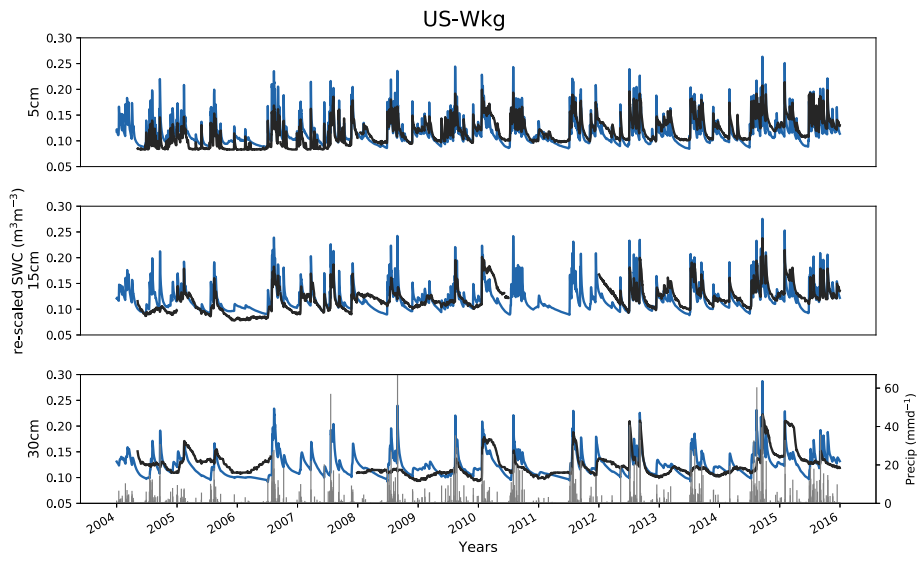
Deleted:

365

370

375

f) US-Wkg



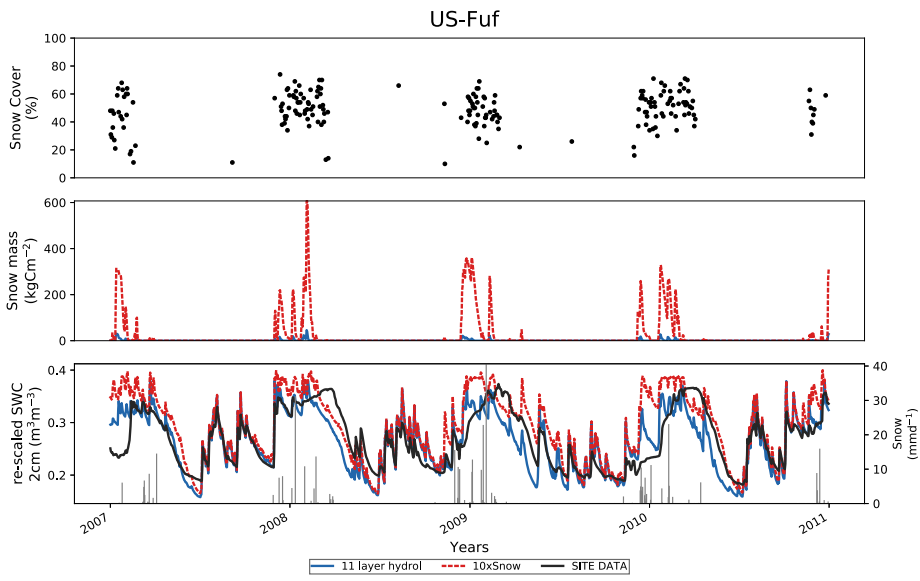
380

385

390

395

**Figure S5: Linear regressions between spring (March-April) mean monthly LAI ( $m^2m^{-2}$ ) and spring mean monthly ET (mmmonth<sup>-1</sup>) model-data misfits for each site. The dominant PFT is given in brackets for each site. See Table 1 for PFT acronyms.**

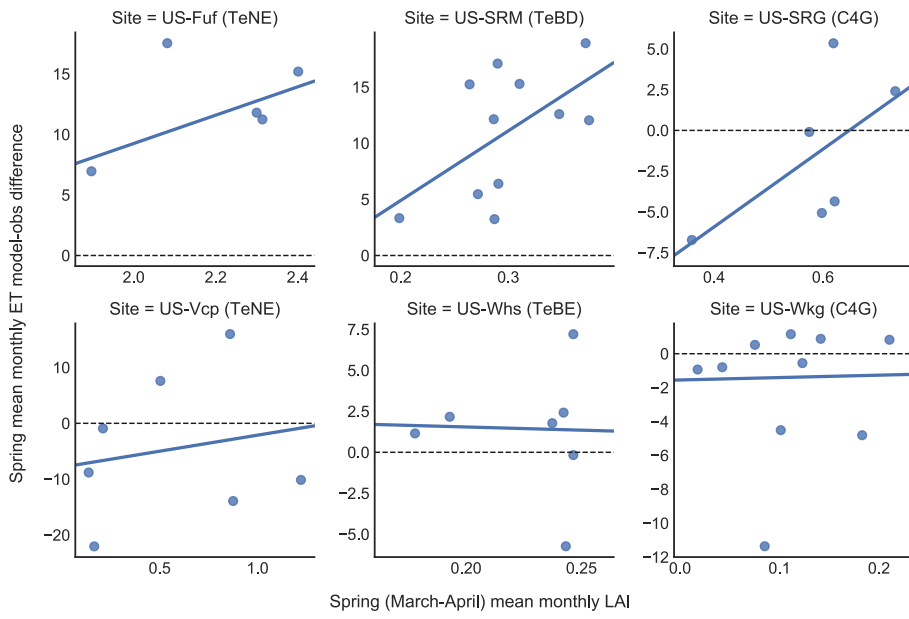


400

405

Deleted: 5

Figure S6: Linear regressions between spring (March-April) mean monthly LAI ( $m^2m^{-2}$ ) and spring mean monthly ET (mmmonth<sup>-1</sup>) model-data misfits for each site. The dominant PFT is given in brackets for each site. See Table 1 for PFT acronyms.



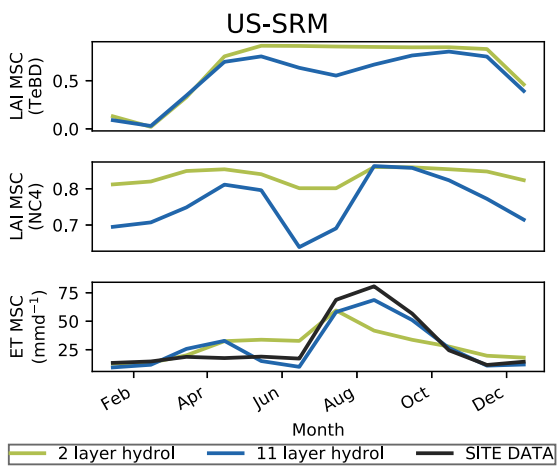
410

415

420

425 **Figure S2:** Plots comparing ET and LAI for C4 grasses (C4G) and mesquite shrubs (Temperate Broadleaved Deciduous – TeBD – PFT in ORCHIDEE) monthly mean seasonal cycles at US-SRM for the 2LAY (green curve) and 11LAY (blue curve) model versions in comparison to observations (black curve).

Deleted: 6

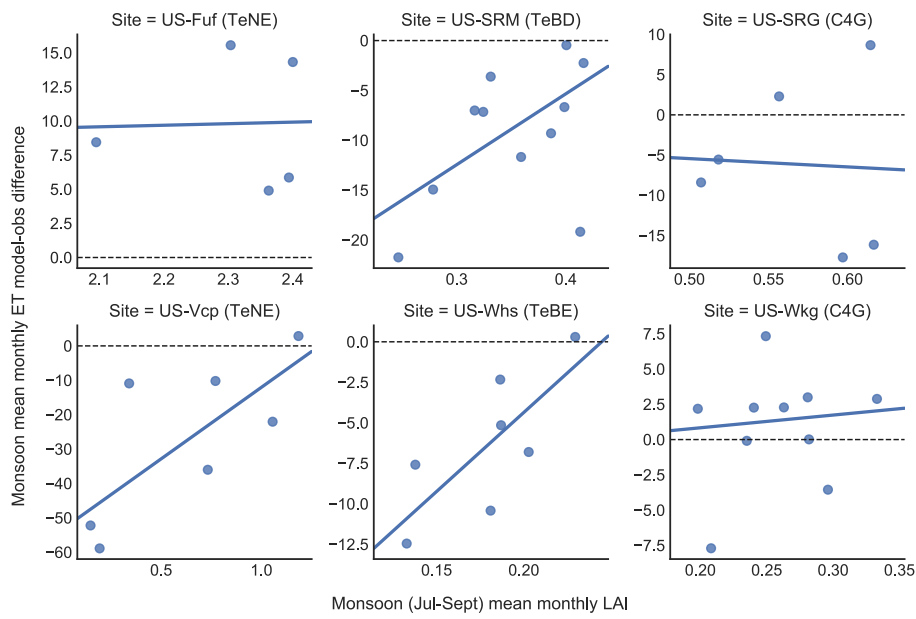


430

435

440

**Figure S8:** Linear regressions between monsoon (July-September) mean monthly LAI ( $m^2m^{-2}$ ) and monsoon 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.



Deleted: 7  
 Formatted: Superscript  
 Formatted: Superscript  
 Formatted: Superscript

450

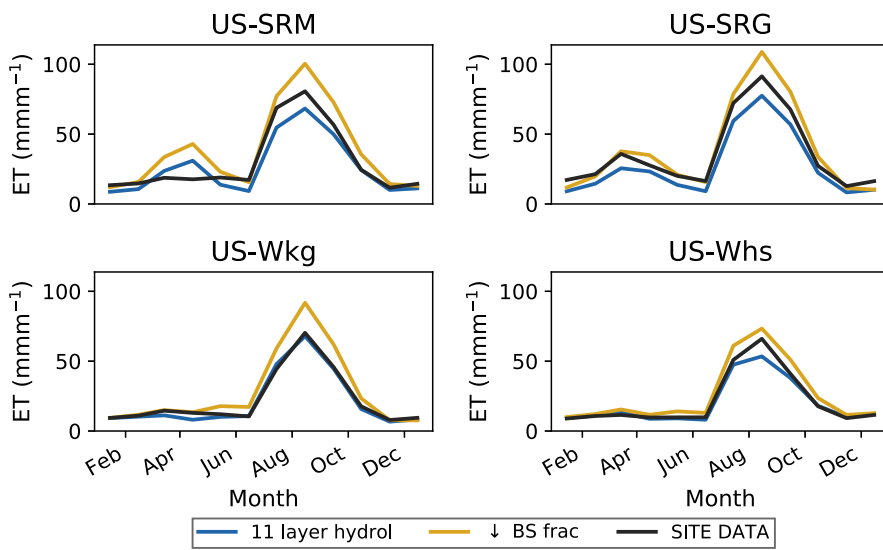
455

460

Figure S2: ET monthly mean seasonal cycle for all low elevation sites comparing the default 11LAY simulations (blue curve) with a simulation that increased the C4 grass fraction at the expense of the bare soil fraction (yellow curve). ET is compared to observations (black curve). Units are mm per month ( $\text{mm}^{-1}$ ).

Deleted: 8

Deleted: Units in  $\text{mmmonth}^{-1}$ .



465

470

475



480

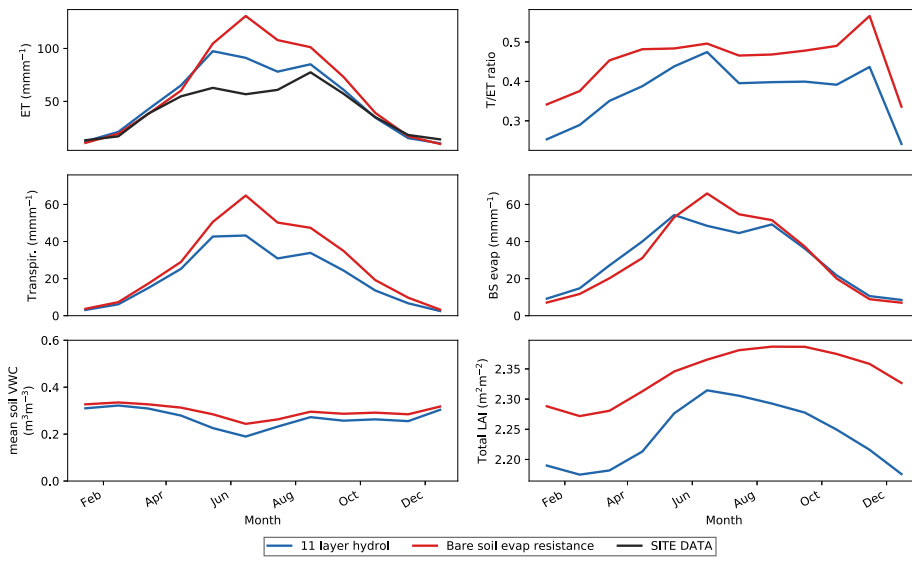
Figure S10: Monthly mean seasonal cycle for all sites comparing the default 11LAY simulations (blue curve) with a simulation that included an additional bare soil evaporation resistance term (red curve). ET is compared to observations (black curve). In all subfigures – top right: T/ET ratios; bottom left: transpiration; bottom right: bare soil evaporation. Units are mm per month (mmm)

Deleted: 9

Deleted: Units in mmmonth<sup>-1</sup>.

a) US-Fuf

US-Fuf

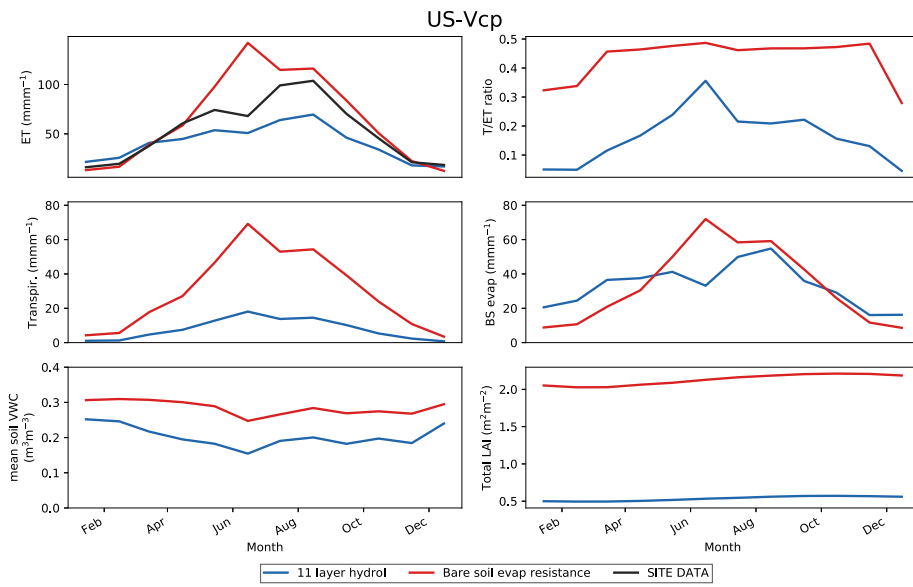


485

490

495

b) US-Vcp

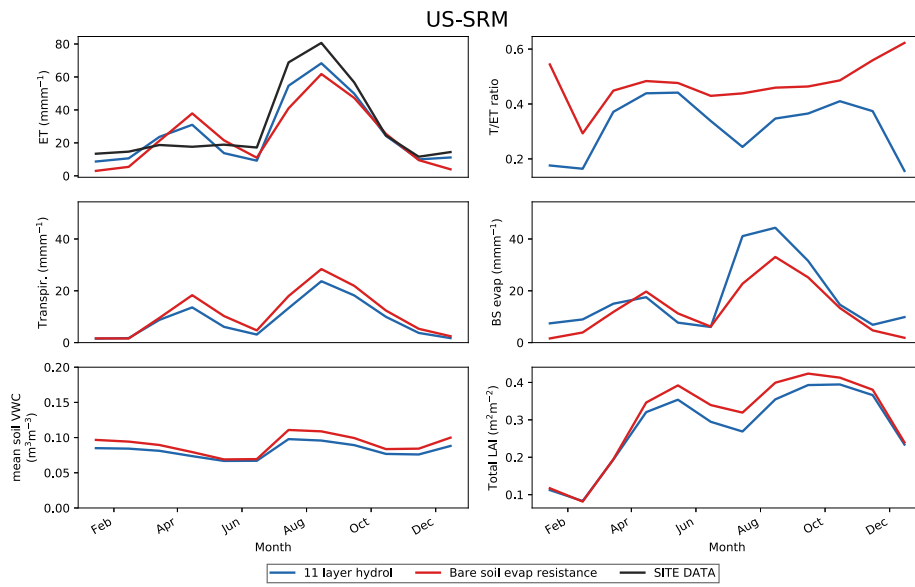


500

505

510

515 c) US-SRM



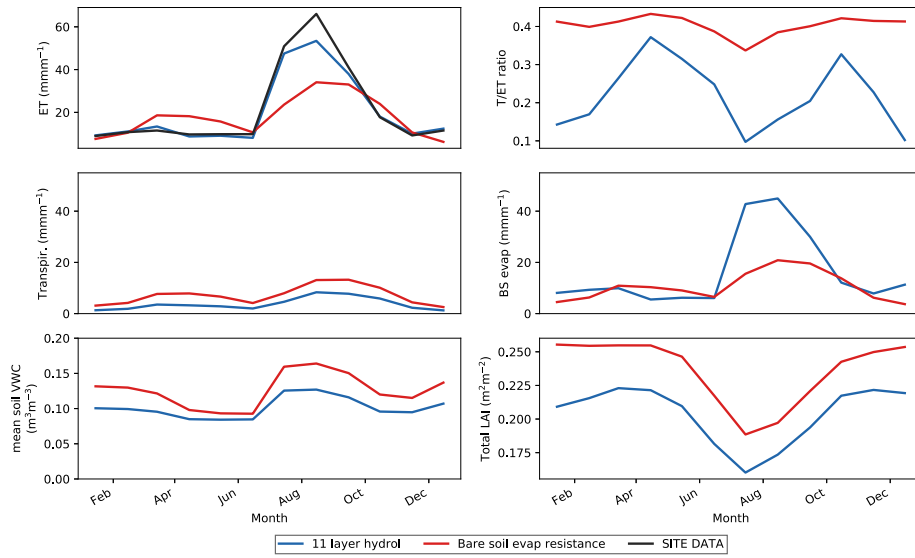
520

525

530

d) US-Whs

US-Whs



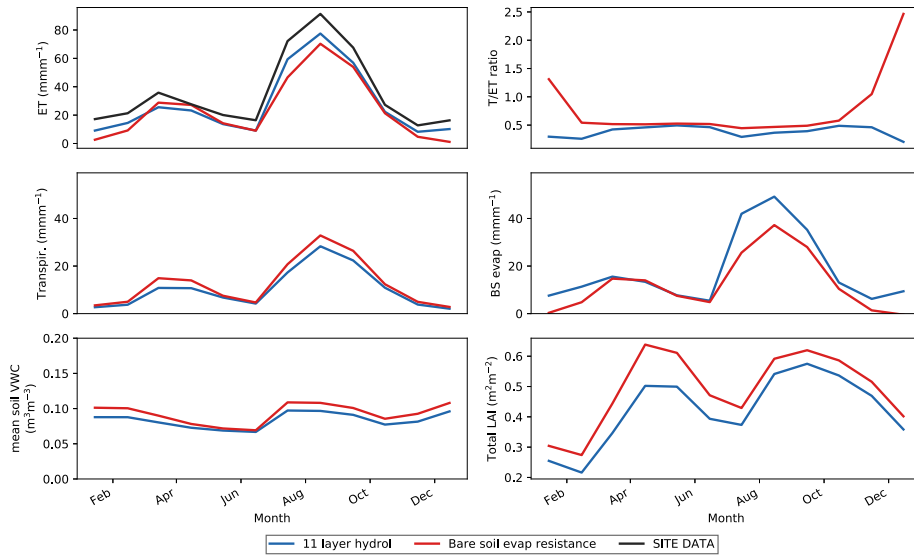
535

540

545

e) US-SRG

US-SRG



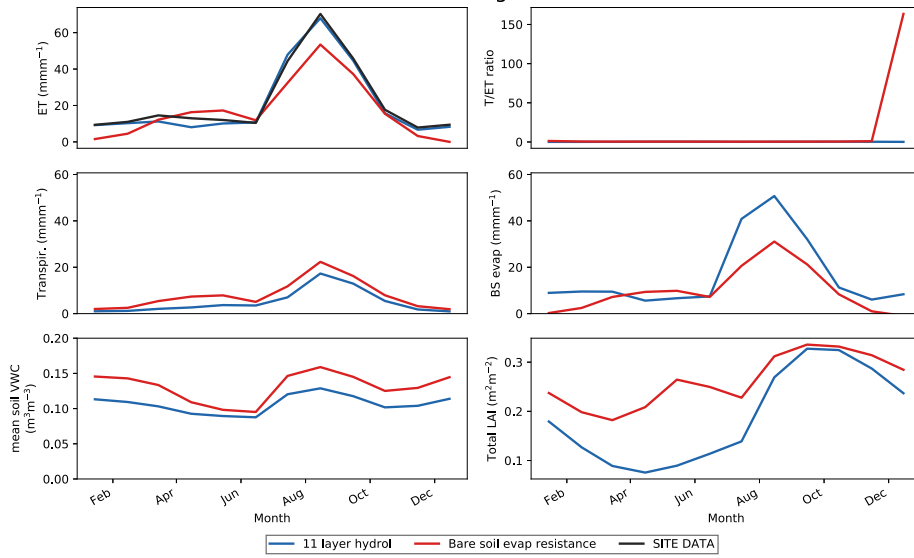
550

555

560

f) US-Wkg

US-Wkg



565

570

575

Figure S1: Monthly mean seasonal cycle for all sites comparing the default 11LAYER simulations (blue curve) with a) with a simulation that increased the C4 grass fraction at the expense of the bare soil fraction (yellow curve); and b) a simulation that included an additional bare soil evaporation resistance term (red curve). In all subfigures – top left: mean soil moisture; top right: ET compared to observation (black curve); bottom left: transpiration; bottom right: bare soil evaporation. Units are mm per month (mmm<sup>-1</sup>).

Deleted: 0

Deleted: Units in mmm<sup>-1</sup>.

