Evaluating a land surface model at a water-limited site: implications for land surface contributions to droughts and heatwaves

Response to Reviewer 1

We thank the reviewer for the positive comments and constructive suggestions. We have addressed the various concerns below. Our responses to reviewer comments are highlighted in blue below each reviewer comment.

In this study the authors validate the CABLE land surface model against measurements representing multiple relevant state variables and fluxes from a site in Australia. Using different model configurations they test the relevance of a range of processes known to affect modelling performance and find that most of them are also important at the considered site. They conclude that land surface modelling and model development should focus on several variables and correspondingly multiple processes at the same time to ensure meaningful model performance is obtained, and for the right reasons.

Recommendation: I think the paper requires moderate revisions.

The topic of this study is timely, and relevant for the community and even beyond in the context of climate change projections. While there are many studies investigating particular known challenges in land surface modelling, I find it very insightful to see a joint consideration of these challenges, and of their interactions. But I also see some shortcomings in this paper which should be addressed before the paper is suitable for publication in HESS:

(1) The order of the changes applied to the model configuration is not motivated. I think it should at least be discussed why and how this order was chosen, as I believe that the different changes applied to the model interact with each other, thereby leading to over- or underestimation of the effect of individual changes.

> We chose to first resolve the soil evaporation bias because although this affects overall evapotranspiration partitioning, the impact of changes in soil evaporation should be constrained to the top-soil layers and would be of greatest importance following rain. We next tackled the initialisation of the water table as this fundamentally affects the root-zone moisture state. We then explored assumptions related to soil layer resolution/parameters, on the basis that the previous experiments would have (largely) resolved biases affecting overall soil moisture availability. Next, we explored optimising soil parameters on the basis that we had resolved substantive biases in the simulated hydrology. Finally, we explored assumptions related to the soil moisture stress function. We felt it was important to do this last as this water stress factor integrates, and arises from, the state of the soil moisture profile. As a result, exploring this water stress factor first and then fixing other biases (*e.g.* drainage) would affecting the overall soil moisture and would overstate the relative importance of the water stress factor. We note this is what is commonly done in studies that resolve a single process *e.g.* the water stress factor.
>
> We agree with the reviewer that the order of experiments may affect interpretation; however, we contend that there is no "perfect" experimental order and would argue there is merit to our chosen order. One could select an alternative order, but without any obvious reasoning, and the possible permutations are vast. We can assure the reviewer that we discussed the merits of these choices at length.

We have added to our justification of the experimental order on line 240-250:

"We choose to first resolve a soil evaporation bias as it affects ET partitioning; however, its importance is limited to the top soil layers and particularly to the period following rain. We then modified the initial water table depth as this fundamentally affects the root-zone soil moisture state. Next, we explored assumptions related to soil column discretisation and parameters, and further optimised key hydraulic parameters to improve overall soil moisture biases. We choose to explore parameter assumptions at this point in the experimental set up as the previous experiments aimed to resolve existing biases that affected the overall soil moisture availability. Finally, we explored alternative soil moisture stress functions as the last step because this factor integrates, and arises from, the soil moisture state. The experimental order allowed us to probe model biases in a systematic way but it is important to note that there is no perfect experimental order and alternative permutations would lead to subtly different interpretation of results. In fact, this is what commonly happens in model evaluation that explore a single factor (e.g. the soil moisture stress factor)."

(2) The observations with which the model simulations are compared are themselves subject to uncertainty. While I acknowledge that the authors are aware of this, and mention this here and there in section 4, I would like to see a more extensive discussion of this, particularly in the results section where model performance differences are assessed without discussing the significance of these changes in the light of observation uncertainties.

Thanks for the suggestion. We have now added a discussion of observation uncertainties in three locations:

[line 520-523]: *"Notably, soil evaporation was not directly measured at the site, but instead derived from the change in observed soil moisture over the top 5 cm, while ignoring days following rain (when the soil evaporative flux would likely be largest) (Gimeno et al., 2018a). As such, it contains soil moisture changes due to transpiration from a seasonal grass understorey but ignores evaporation below the top 5 cm, complicating model evaluation."*

[line 578-584]: *"On the other hand, the neutron probe measurements of soil moisture used for calibration also involve uncertainties (Gimeno et al. 2018a). The soil moisture estimates were derived by fitting two distinctive linear relationships between soil volumetric water content and raw neutron probe counts (see Figure S6) for clay (below 3m) and non-clay soil (above 3m). As a result, the observation error would be greatest in layers where the soil type differs from the assumed soil type at that depth. However, the fitted relationships were robust, since clay soils largely dominated the deeper profile (below 3 m depth) and sand soils mostly dominated shallow profile (above 3m depth)."*

[line 115-118]: "*Etr estimates are derived from tree sapflow velocities (3-4 trees per experimental ring) using the heat pulse compensation technique (Gimeno et al., 2018a). Sapflow velocity is translated to Etr by multiplying the sapwood area estimated from basal area inside each ring and a correlation between sapwood and basal areas based on 35 trees adjacent to the experimental rings.*" And [line 120-121]: "*To represent variability in Etr and Es across rings, we show the mean and the uncertainty within ring estimates in all figures."*

(3) It would be nice to have some discussion on the representativeness of the obtained conclusions across spatial and temporal scales (actually I could not even find the temporal scale at which the model simulations were done). Are these model improvements expected to hold at larger spatial scales relevant

for climate (change) modelling? And more generally, to which extent can we possibly learn from such small scale analyses to improve large scale modelling?

Model performance in water limited conditions has been widely identified as a key source of weakness in model evaluations. Perhaps due to data limitations, past studies have focused on individual processes in isolation. By contrast, we have aimed to achieve a more holistic model evaluation of a range of processes, enabled by the exceptionally comprehensive observations. While we do not anticipate our findings to directly constrain any single model's global simulations, we hope that elements of our findings guide future model improvements at temporal scales including daily extremes, seasonal and annual scales.

The observational data covers a relatively short time period (2013-2019) but some of the biases identified here have been shown to re-occur annually in many LSMs (particularly at seasonally dry sites; Ukkola et al., 2016). Fixing these biases is therefore likely to be valuable in longer term simulations. More broadly, understanding gained from this study better informs how process assumptions feedback and affect coupled simulations of droughts and heatwaves. More specifically, we intend to extend these evaluations in future work (see future directions section) to resolve existing biases, working from the end point of these sensitivity experiments. While our study concentrates on the CABLE model, the process representation noted here are broadly shared across a number of other leading LSMs.

We now highlight the relevance and lessons learnt for other LSMs in multiple sections of the discussion:

[line 514-515]: "*LSMs commonly overestimate soil evaporation especially under a sparse canopy or over bare land (De Kauwe et al., 2017; Swenson and Lawrence, 2014), suggesting this is a key model weakness.*"

[line 538-540]: "*When we initialised from a drier starting position (Watr), the simulated soil moisture profile matched the observed better, with implications for other models using similar groundwater schemes (e.g. CLM4.5, Noah-MP, JULES and LEAFHYDRO).*"

[line 552-554]: "*LSMs typically define a fixed number of soil layers globally, anywhere up to 20 layers. Most LSMs assume constant parameters across the entire soil profile, either using an experimental look-up table based on soil classification or estimating parameters from empirical pedotransfer functions.*"

[line 588-590]: "*Studies commonly highlight the functions used to limit photosynthesis and stomatal conductance with water stress as a key weakness among models. The lack of theory in this space (Medlyn et al., 2016) has led to models employing a range of functions encompassing different shapes and sensitivities that are not constrained by data.*"

[line 598-599]: "*the linear θ-based function used in Ctl (common among models, e.g. SDGVM, Orchidee-CN and JULES)*"

We also mentioned the possibility to utilize the depth-varying soil parameters on a global scale with the novel dataset:

[line 562-564]: "*High-resolution global soil datasets (e.g. SoilGrids, Hengl et al., 2017) covering* *multiple soil layers up to 2m depth offer opportunities to improve LSM simulations of soil moisture by incorporating depth-varying soil parameters.*"

Additionally, we now explicitly state that CABLE was run at 30-min resolution as the reviewer correctly pointed out that this was not made clear in the original manuscript:

[line 105-107]: "*Following Yang et al. (2020), the meteorological data were gap-filled (0.8% of values, from 1 January 2013 to 31 December 2019) using linear interpolation, aggregated to 30-minute averages and subsequently used to force CABLE at the 30-min resolution.*"

(4) Similarly, I was missing some discussion on the potential applicability of the derived conclusions to other models and regions. How can modellers using different models and focusing on other sites/regions benefit from the results obtained in this study?

We wrote our discussion *deliberately* to be general in its findings such that the lessons learned extend beyond the CABLE model. We ordered our sub-headings accordingly to offer insight into: soil evaporation, aquifer initialization, pedotransfer functions, optimisation and water stress functions.

As for the point about regions, it is a little speculative to comment on. Clearly our analysis is site specific, but the processes we identify and discuss (see list above) are more general. Furthermore, the biases and processes explored here have been shown to lead to systematic biases in LSMs across multiple sites in previous studies (Ukkola et al., 2016; Trugman et al., 2018). We anticipate our findings would be applicable in many water-limited conditions, but equally, more mesic systems too.

To make these points clearer we now add two sentences (line 495-496, "*Whilst our analysis is site specific, the issues indicated here have been reported to lead to systematic biases in LSMs across multiple sites (Ukkola et al., 2016a; Trugman et al., 2018)*", and line 503-505, "*Since our study attempts to articulate the common issues in the simulated dry conditions in LSMs, we anticipate our findings would be applicable in many water-limited conditions, but equally, more mesic systems too*") in the general discussion and a few sentences in these sub-headings to explain to the reader that our discussion is intended to be generally applicable to models and our thoughts on regional transferability. We have also added examples where other LSMs share similar parameterisations to CABLE (see previous comment).

Specifically, we illustrated and rephrased our suggestions to other LSMs in every sub-section of discussion:

[line 526-533]: "*However, a number of studies using alternative process-based schemes have been shown to improve individual model simulations (Haverd and Cuntz, 2010; Lehmann et al., 2018; Or and Lehmann, 2019). For example, Swenson and Lawrence (2014) introduced a dry surface layer-based soil evaporation resistance into CLM to depict water diffusion from dry soil, reducing biases in evapotranspiration and total water storage relative to FLUXNET-MTE and GRACE datasets. Based on a pore-scale model (Haghighi and Or, 2015), Decker et al. (2017) added the resistances of capillary-viscous and boundary layer to CABLE soil evaporation scheme and lowered the positive Es bias in springtime and improved seasonality of evapotranspiration. Hence, a focussed intercomparison of competing approaches against data originating from different ecosystems would be a valuable area of future work.*"

[line 540-550]: "*First, our results imply that LSMs that incorporate groundwater schemes need to be careful about aquifer initialisation because this strongly affects soil moisture dynamics. Second, there is no obvious solution to this initialisation and spin-up problem because drainage into the aquifer is a slow process, and it may take hundreds of years to reach a realistic equilibrium state. For global simulations, this suggests the need to a priori initialise the starting aquifer state and to assess against satellite-based products like GRACE (Döll et al., 2014; Niu et al., 2007) or implement off-line spin-up using meteorological forcing consistent with the subsequent simulations. However, while spin-up with observations is attractive, when the resulting states are incorporated into a coupled global model, inconsistencies are inevitable. Third, CABLE currently assumes an identical spin-up approach for the aquifer as the soil moisture, iterating until state changes between sequences of years are smaller than some threshold. LSMs that employ similar iteration approaches (Gilbert et al., 2017) are likely to encounter similar problems as CABLE because the rate of drainage into the aquifer is very slow, leading to negligible changes between iterations and thus satisfying the criteria for equilibrium.*"

[line 561-567]: "*The development in pedotransfer functions via machine learning or multi-model ensemble provides new avenues to reduce errors from parameters (Zhang and Schaap, 2017; Dai et al., 2019). High-resolution global soil datasets (e.g. SoilGrids, Hengl et al., 2017) covering multiple soil layers up to 2m depth offer opportunities to improve LSM simulations of soil moisture by incorporating depth-varying soil parameters. It is noteworthy that these global datasets of soil hydraulic parameters (Montzka et al, 2017; Zhang et al., 2019) have existed for several years but have not been widely used. Furthermore, at the EucFACE site, the observed soil texture information enabled the separation of parameter uncertainties from biases in process representations and model structural errors, a valuable step in better constraining LSM simulations.*"

[line 611-617]: "*Alternatives to the β functions have emerged to fill the theoretical gap, including plant hydraulic (Christoffersen et al., 2016; Xu et al., 2016) and stomatal optimality approaches (Sperry et al., 2017) but are yet to be widely adopted in LSMs (but see Eller et al., 2020; De Kauwe et al., 2020; Kennedy et al., 2019; Sabot et al., 2020). Replacing the empirical soil water stress factor by these plant physiology schemes reduces model arbitrariness associated with the representation of soil water stress and reduces the simulated biases in transpiration either over water deficit regions or areas with obvious dry seasons (Bonan et al., 2014; De Kauwe et al., 2020; Sabot et al., 2020). We can envision a wider application of these processes-based models will offer a chance to improve water stress representation in more LSMs.*"

Specific comments:

lines 23-24: not clear at this point what 'median level of water stress' is

We have clarified the text to now read, *"reduced the soil water stress on plants by 36 % during drought and 23 % at other times"*, which we think is enough detail in the abstract (line 23-24).

line 25: 'Alternative' could be replaced by 'The range of tested' for improved clarity

Agreed – we have made this change (line 26).

lines 42-43: you could cite here Orth and Destouni 2018

Thanks, and we have added the reference (line 44).

line 79: ‘soil moisture extending root zone’, please improve phrasing

Agreed – this is clumsy English. We have modified the text to “*to utilise observations of soil moisture extending through the root zone with concurrent measurements of water fluxes at high temporal frequency*” (line 79-80).

line 95: what is meant with ‘Sm’ here?

It is common practice for some species to also include an abbreviation after the genus and species name, this denoted the person or persons who first formally described/discovered the species. In this case Sm. refers to Smith.

line 105: You talk about gap filling here. How many gaps were filled this way?

Only 0.8% of the meteorological forcing was gap filled and we have added this detail in the paper (line 106).

lines 120-123: Could you give some details on how the neutron probe measurement works and is done at 12 different depths?

We have provided additional detail (line 127-129, “*The neutron probe counts are converted to θ via the site specific linear correlation between the raw reading of neutron probe and the lab measured soil θ sampled at the same depth as probes (Gimeno et al. 2018a)*”), but more critically added the reference to the details of how this was done.

line 286: why 31 layers?

This was done to match the resolution of observed soil texture which were mostly sampled at 15 cm intervals (leading to 31 layers in total over the total soil depth). We now make this clear in the text (line 301-302, “*the number of vertical soil layers was increased from 6 to 31 (to match the resolution of observed soil texture which was sampled at 15-cm intervals)*”).

line 309: ’Due to muted variability’, can you please give more details here?

We have reworded the sentence for clarity (line 323-325, “*θ_{sat} was not adjusted below 30 cm as the observed maximum θ_{sat} is unlikely to represent saturated conditions due to lower soil moisture variability at depth.*”).

line 339: Why not stating the applied exponent 0.425 here?

Agreed, and we have added the value into the text (line 354).

line 354: Why would accounting for defoliation by decreased LAI be insufficient?

Insect attack can also damage the phloem hence the full impact may not be captured by a reduction in LAI. We have modified the text to clarify this (line 369-370, “*CABLE only accounts for canopy* *defoliation via a decline in LAI but not other damage e.g. to the phloem*").

line 379: it is not mentioned in the respective section 2.4.3 that the aquifer is 'initialised' drier

We have added additional text to clarify this in Section 2.4.3 (line 289-290, "*which reduced the initial saturation of aquifer from 100% to 52%*").

lines 407-411: Figure 6 should be mentioned earlier in this paragraph

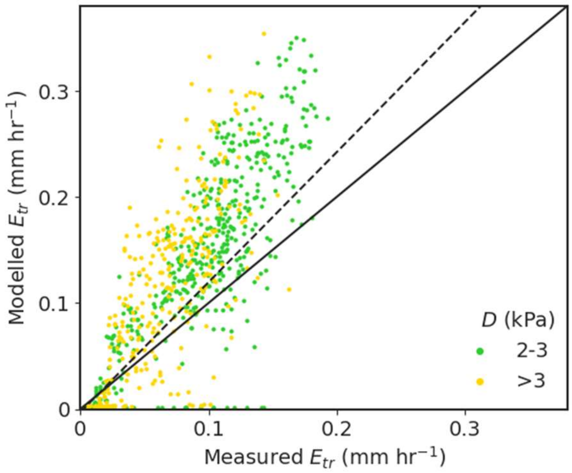
Agreed and corrected (line 424).

line 439: 98% is relative to the maximum I guess?

We have modified the text to make our statement clear (line 455-456, "*a difference of 98 % relative to the averaged median of all the simulations for β simulated during drought*").

line 467-469: Shouldn't this be the other way round?

We have clarified our text to make our meaning clear as:"*during heatwaves **when D is higher** the model would overestimate Etr"* in line 484-485, and the figure below demonstrates this.



line 598: typo in 'transpiration'

Thanks, and corrected (line 637).

line 615: you could cite here Orth et al. 2017

Agreed, and we have added the reference at line 655.

Figures 2-7: Please point the reader to the different time axes used in this plot, and/or use a regular time step spacing in plots c,d,e while showing data gaps e.g. in gray. This can improve readability and comparability across plots I think.

We have modified the figure legend to make this much clearer and explicit. We now state "*Note the different time axis for (c-e) relative to (a-b) due to different sampling intervals for soil moisture and fluxes*."

Figure 10: The different timing of the peaks which you repeatedly refer to in the text could be illustrated by vertical thin lines with respective colors highlighting these peaks.

Thanks for the suggestion. We tried to plot these vertical lines but they are too crowded and hard to be read. However, we have added vertical reference lines at a 6-hour interval to assist reading Figure 10.

References:

Trugman, A. T., Medvigy, D., Mankin, J. S. and Anderegg, W. R. L.: Soil moisture stress as a major driver of carbon cycle uncertainty, Geophys. Res. Lett., 45(13), 6495–6503, doi:10.1029/2018GL078131, 2018.

Ukkola, A. M., De Kauwe, M. G., Pitman, A. J., Best, M. J., Haverd, V., M., D., G., A. and Haughton, N.: Land surface models systematically overestimate the intensity, duration and magnitude of seasonal-scale evaporative droughts. In review., Environ. Res. Lett., 11, 104012, 2016.