Response to Reviewer 1

General comments

This paper demonstrates the effect that modelling groundwater in the CABLE land-surface model has on droughts and heatwaves, using two droughts and multiple heatwaves in South East Australia as case studies. This is a very important and topical issue, and well within the ESD remit. The analysis is thorough and well-designed, and described in sufficient detail to allow reproduction. Particular attention is paid to understanding the mechanisms behind the results, which considerably strengthens the conclusions. The relevance to climate model projections is particularly well put, clearly stating the important implications of this study whilst carefully outlining uncertainties and avoiding over-generalisation.

All of the plots in both the main manuscript and the supplementary material are important for the arguments presented, and of a high production standard. The prose is well written, the structure is good, and there is a high attention to detail.

Overall, this is a very strong manuscript, which will make an important contribution to the field.

> We thank the reviewer for the positive summary of our work and we address the reviewer's concerns below. Comments are shown in black, with our response below in blue in each case.

Specific comments

Section 3.1: This section shows that CABLE-GW has a very good agreement with GRACE total water storage, and that the GW run has a better agreement with GRACE than the FD run. However, I'm not completely convinced by the conclusion that the underestimation of TWSA in the FD run is because of the lack of groundwater representation. (i.e. I don't think that other model deficiencies with the potential to reduce E have been ruled out). This is particularly the case as the accumulated P-E in GW run is still substantially different to GLEAM, showing that there is still an issue with the model even with groundwater included. To address this I would suggest (a) being a bit more cautious in the phrasing so that the text doesn't imply that including ground water is the only way to make the model match more closely to the observations and (b) including some text discussing possible reasons why the GW and GLEAM lines do not agree in figure 1b.

> We thank the reviewer for this comment, we fully agree that underestimation of TWSA is likely to also relate to other model biases. We will clarify in the manuscript that the underestimation of TWSA was in reference to FD compared to GW; these runs are identical for all parameterisations except groundwater, allowing us to attribute the bias to the groundwater processes:

“This underestimation in FD compared to GW is linked with the lack of aquifer water storage in the FD simulations which provides a reservoir of water that changes slowly and has a memory of previous wet/dry climate conditions (Figure 1a).”

With respect to the comment about GW and GLEAM lines not agreeing, we first note that GLEAM is itself a model (which ingested observations), so a mis-match does not necessarily point to a CABLE issue. To better illustrate this point we will add a second ET product, the Derived Optimal Linear Combination Evapotranspiration (DOLCE) (Hobeichi et al., 2021) estimate to Figure 1b (see blue line in the figure below). DOLCE is an observationally constrained product, similar to GLEAM. As can be seen, the GW simulation sits between the DOLCE and GLEAM estimates, whereas FD is outside the envelope of these observationally-constrained products.

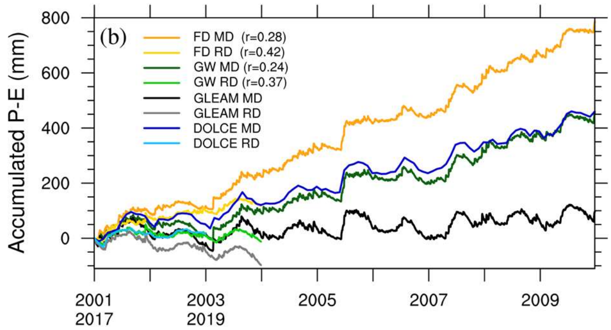

Figure 1 panel b: accumulated P–E for the two droughts over S.E. Australia. It shows the accumulated P–E for two periods; the dark lines show the 2001–2009 Millennium drought (MD) and the light lines show the 2017-2019 recent drought (RD). The correlation (r) between the P and E is shown in the legend of panel (b).

We will add a sentence to address the reviewer’s point about attributing all of the improvement in TWSA to GW alone, linking to both the GLEAM and DOLCE estimates, to highlight that we fully expect CABLE to still have other evaporative biases:

“Although the total land evaporation products display some differences, the GW simulations are closer overall to the DOLCE and GLEAM estimates. Whilst the biases in evapotranspiration can arise from multiple sources, the better match of GW than FD to the two observationally-constrained products implies that overall adding groundwater improves the simulations during droughts.”

Line 229: “FD underestimates the magnitude of monthly TWSA variance (standard deviation, SD = 37.18 mm) compared to GRACE (47.74 mm) or GW (47.67 mm)”: consider showing this explicitly in a plot, as it’s an important result, which is difficult to read off Figure 1a.

We thank the reviewer for this point. We note that we did add these metrics to the top left corner of Figure 1a. To make this clearer to our readers, we will revise the colour of these metrics from black to blue in the figure panel.

Line 351: elaborate on why MODIS LST is lower than all the model lines in Figure 6h, even DR.

We checked the LAI in the underlying grid box in Figure 6h and it suggests a high LAI coverage, which would tend to imply that the MODIS LST is representing a good approximation of the canopy temperature. As a result, the lower MODIS LST$-T_{air}$ would tend to imply that CABLE is underestimating transpiration, leading to a greater $T_{canopy} - T_{air}$. We will make this clearer in the results:

"The shallower WTD region tends to have a high LAI coverage, implying that the MODIS LST likely represents a good approximation of the canopy temperature over this region. Consequently, the lower MODIS ΔT implies that CABLE is likely underestimating transpiration, leading to an overestimation of ΔT in all three simulations."

Line 369: "first ~two years of a multi-year drought": link this explicitly to plots (as far as I could see, this is the first time this was mentioned, but it's picked out as one of the main points of the study in both the abstract and the conclusion).

We will add shading similar to Figure 1 to Figure 3 to make it clearer where the drought periods are and hope this will make the link clearer. To clarify the statement of "first ~ two years of a multi-year drought", we will make this result clearer in the results and in the discussion:

"FD underestimates the magnitude of monthly TWSA variance (standard deviation, SD = 37.18 mm) compared to GRACE (47.74 mm) or GW (47.67 mm), particularly during the wetter periods (2000, 2011-2016) and the first ~2 years of the droughts (2001-2, 2017-8) (Figure 1a)",

"For all variables (ΔT, EF, Et and β), the difference between GW and FD is greatest during the wetter periods (e.g. 2013) and in the first 1-2 years of the multi-year drought (2001-2002 for the Millennium Drought or 2017-2018 for the recent drought). After drought onset, the FD and GW simulations converge as depleting soil moisture reservoirs reduce the impact of groundwater on canopy cooling and evaporative fluxes",

and "We found that the influence of groundwater was the most important during the wetter periods and the first ~ two years of a multi-year drought (~2001-2002 and 2017-2018; Figure 1 and 3)".

Line 398-9 "Our regional based results support this hypothesis and in particular highlight the importance of groundwater for explaining the amplitude of fluxes in wet regions (Figure 1)" Elaborate on how Figure 1 shows this.

We thank the reviewer for spotting this mistake. We will change "wet regions" to "wetter periods" in the text:

"Our regional based results support this hypothesis and in particular highlight the importance of groundwater for explaining the amplitude of fluxes in wet periods, as well as sustaining evaporation during drought (Figure 1)".

We will also better highlight this behaviour during the wetter periods in the results:

"FD underestimates the magnitude of monthly TWSA variance (standard deviation, SD = 37.18 mm) compared to GRACE (47.74 mm) or GW (47.67 mm), particularly during the wetter periods (2000, 2011-2016) and the first ~2 years of the droughts (2001-2, 2017-8) (Figure 1a)",

and "For all variables (ΔT, EF, Et and β), the difference between GW and FD is greatest during the wetter periods (e.g. 2013) and in the first 1-2 years of the multi-year drought (2001-2002 for the Millennium Drought or 2017-2018 for the recent drought)".

Technical corrections

Line 434-436: This sentence doesn't read well. Is it missing a "that" or an "and"?

To solve this comment, we will change the sentence to:

"Figure 1 gives us confidence that CABLE-GW is performing well, based on the evaluation against the GRACE, DOLCE and GLEAM products, and as well as previous work that showed the capacity of CABLE-GW to simulate E well (Decker et al., 2017). However, we also note that key model parameterisations that may influence the role of groundwater are particularly uncertain".

References:

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, J. Adv. Model. Earth Syst., 9(1), 220–238, doi:10.1002/2016MS000832, 2017.

Hobeichi, S., Abramowitz, G. and Evans, J. P.: Robust historical evapotranspiration trends across climate regimes, Hydrol. Earth Syst. Sci., 25(7), 3855–3874, doi:10.5194/hess-25-3855-2021, 2021.