Reviewer (R3) specific comments (C)

R3_C1:

Overall, Denager et al presents an interesting and impressive study, involving a complex model (CLM), top-level observations (especially regarding the water balance) and an advanced calibration scheme for optimizing the model parameters. Yet, the results are somewhat contradictory, and the manuscript should be improved before acceptance.

I also find it a somewhat remarkable result, that such elaborate setups and multitude of observations are needed to improve the model performance. To me, it points rather to issues in the model (unless all issues can be blamed on the lack of energy balance closure). Given the text in the introduction, and especially the highly relevant quote by Clark et al., I question whether the approach of keeping the highly complicated LSMs and needing to perform elaborate model calibrations (involving a large number of observations really) really is a good way forward for the community. It would be interesting if the authors could comment on this aspect in the study.

Reply: We will include a new paragraph in the discussion section; "Model physics in land surface models" to discuss these aspects. See paragraph below:

Model physics in land surface models: In this study we have shown that applying the model with observed parameter values were possible (Scenario Z) did not lead to an improved model performance (compared to Scenario X, which use look-up table parameter values), which can potentially be interpreted as deficiencies in the model physics. As stated by Clark et al. (2015) more and more advanced descriptions of the processes have been built into LSM codes. This induces increased model complexity and expands the associated number of parameters in the model equations (Mendoza et al., 2014). Parameter optimization in complex models is complicated, and there is a possibility that LSMs may not be parameterized appropriately. Several authors have contested the complexity of LSMs (Clark et al., 2015; Franks et al., 1999; McCabe et al., 2005; Williams et al., 2009) and suggested a reassessment of the structure and process representations. An overall simplification of the LSMs would enable a more profound parameter optimization and utilization of measured data. This would lead to more parsimonious LSMs and utilizing the well-establish model evaluation within hydrology considering uncertainties in data, model parameters and conceptual understanding (Refsgaard et al., 2021), would enhance the model evaluation of LSMs. In this way, the hydrology and LSM modelling communities could benefit even more from each other (Clark et al., 2015).

R3_C2:

Results and main conclusion: The many small tables of Figure 1 are hard to read and also hard to interpret.

Reply: See R2_C6

R3_C3:

I assume that it is the results in these summary tables that lead the authors to their main conclusion "that mathematical regularization is a compelling method to improve current practice of using look up tables to define parameter values in LSMs" (a similar claim is made in the first paragraph of the Discussion, page 19).

The authors should explain clearly how they reach this conclusion, since their approach also shows obvious weaknesses. Compared to the control, several error metrices increase when applying the optimization. A further example is that they can only demonstrate improvement by letting observed soil content properties drift away from their observations values (page 17, lines 9-23). Doesn't this rather point to a need for improving the model physics?

Reply: We do not agree that several error metrics increase when applying the optimization. Only 8 out of 80 error metrics included in the optimization in scenario A-H increase compared to the control run (Scenario X). However, we agree that there are outcomes of the study which point to a need for improving the model physics. We will elaborate on this is the discussion of the revised manuscript.

R3_C4:
Another conclusion (lines 13-14, page24) is that use of soil moisture data in the optimization improved soil water storage modeling. Isn't this a rather expected result? Maybe a quantification of this improvement would be more relevant, or a comment on how the model physics could potentially be improved for sites as well as a comment on what to do for the vast majority of sites where such elaborate measurements are not present.
Reply: Yes this is an expected result, and it shows that uncalibrated models are not able to match observed absolute soil moisture. The quantification of the improvement of soil moisture is shown in figure 1 and figure 3.

R3_C5: The authors considerable emphasis on the question whether LE or H is the main culprit when it comes to the lack of energy balance closure (Conclusions, lines 18-20). They highlight that their results point to H, but their site appears to have many more observations related to the water budget than for the heat balance. Neither air nor soil temperature is used in any of the calibration scenarios. Could this result not have been the complete opposite, if they had instead focused on the heat budget and neglected to include all the soil water and moisture parameters? My recommendation is to treat this result with more caution, and at least remove it from the abstract. Rather, the authors should highlight other advantages of their results, for example the relevant conclusion stated on lines 21-25 of the Conclusion.
Reply: The air temperature is an input to CLM5 and is therefore already included in the model setup. We agree that the conclusion on the bias on the simulated H should be taken with caution. As suggested, we have removed it from the abstract.

In response to the concern raised by the reviewer that the heat budget is less constrained than the water budget we have expanded the study by two additional scenarios that include soil temperature as a target variable. The results of scenario E2 and H2 (including Tsoil) are only slightly different from scenario E and H (not including Tsoil). In the revised manuscript scenario E2 and H2 will get other names, so that all the scenario names are in alphabetic order.

A slight improvement in ME_H from -11 Wm^{-2} to -7.9 Wm^{-2} is obtained from scenario E to E2, but ME_H is still highly biased, and the remaining metrics for H are not improved when including Tsoil as target variable.

Scenario E2 (H and Tsoil)

	??	ME	MAE	RMSE	NSE	r
LE	81	-2.6	18	34	0.65	0.88
H	84	-7.9	22	33	0.55	0.77
q	107	0.006	0.050	0.10	-0.00	0.44
SWC	489	-0.059	0.064	0.07	-2.19	0.84
Ts	72	-7.9	15	47	0.01	0.20
Sout	93	-0.6	8	16	0.77	0.88
Rn	98	0.4	15	23	0.97	0.99

Scenario H2 (H, q, SWC and Tsoil)

	??	ME	MAE	RMSE	NSE	r
LE	75	0.1	17	32	0.68	0.87
H	85	-10.1	22	33	0.54	0.78
q	88	0.002	0.046	0.09	0.18	0.52
SWC	50	-0.000	0.016	0.02	0.67	0.83
Ts	71	-8.0	15	47	0.01	0.20
Sout	93	-0.5	8	16	0.77	0.88
Rn	101	1.0	15	23	0.97	0.99

Figure 1: Summary statistics for additional scenarios E2 and H2. Blue color indicate that the variable were included in the calibration for the given scenario.

R3_C6: In agreement with another reviewer, I think that the manuscript would gain in value if the authors focused more on how to choose target calibration variables, and what the presented results tell could us in terms of method applicability and generality.

Reply:
The abstract will be rewritten (see R2_C3) and a paragraph on "Model physics in land surface models" included in the discussion (See R3_C1).

MINOR

R3_C7: Regarding conservation of energy (Eq 1). The equation should at least include the heating of the surface, the top soil and the air, which must be included in any land surface model.
Reply: As in standard eddy covariance studies, Eq. 1 neglect minor fluxes and storage terms. However, the heating of surface, the top soil and the air is included in CLM5.

R3_C8: Numerous places in the paper the term "physical laws" are mentioned and these should be replaced with the precise terms. Which physical laws are, for example, used to simulate H and LE (page 7, row 3)? LSMs typically apply parameterizations including many parameters.
Reply: CLM5 use the Monin-Obukhov similarity theory to simulate H and LE. This is stated the manuscript. We agree that the term "physical laws" should be changed to Monin-Obukhov similarity theory where relevant.

R3_C9: Table 1: In two places, the percentage value of sand content exceeds 100% indicating that the parameter values have not been properly bounded.
Reply: We are certainly aware of this problem and have already commented on this in the manuscript and in R2_C8.

R3_C10: Appendix A: The columns for X and Z appear to be swapped.
Reply: We have double-checked, and the columns are not swapped.

Reviewer: Appendix A: The authors claim that the site is homogeneous, which means that the measured LAI on the site is valid for the whole footprint of the EC flux observations. The scenarios based including LE tends to yield larger values for LAI compared to the observed values, which could indicate the presence of photosynthesizing plants in the footprint of the EC observations. There are very few sites that can be characterized as being completely homogeneous, and all inhomogeneities add to the mismatch between the model world and the real-world situation.

Reply: We agree, no sites can be claimed as homogeneous and we will thus remove this statement.