Response to Report #2

We thank both reviewers for revising our manuscript. In the following we address each of the comments of referee #3 point-by-point. We paste the original comments in red and our response in black. Our corresponding changes to the manuscript are set in italic.

Report #2

I was not one of the original reviewers of the MS but I have reviewed their comments, the authors' replies and the amended MS.

I think the MS would benefit from some English copy editing, which should be provided by the journal. It is a hard slog to get through the paper. Part of it is due to writing style.
We thank the reviewer for taking the burden! As non-native English speakers we cannot judge how much of the 'slog' is caused by unlucky formulations but would definitely appreciate suggestions for improvement, maybe this is a task for the journal's production team during the last stage of the publication process.

Generally the paper could be shortened as it contains a fair amount of discussion in the results section and thereby some overlap/repetition in the actual discussion section. Ideally a Results and Discussion section would be created along with a separate Conclusions. Although I would understand if appetite was low for such a large rewrite.
Assessment papers, like ours, are generally not very sexy and reading is typically over large parts quite boring (which may be another reason that working through it is a real 'slog'). Nevertheless, such papers are a necessity. To make reading a bit more interesting we partly added to the results section also some immediate discussion. As a consequence there is some overlap between the "Results" and "Discussion and Conclusions" sections. But this is by purpose: The whole assessment is meant as a reference for all scientists interested in and/or working with this model. Therefore we designed the "Discussion and Conclusions" section as a quick reference to the main findings. Thanks to the reviewer's comment we now realize that this purpose is not clear from the section title so that we changed it to *"Summary and Conclusions"*.

As you can see from my comments below I am quite critical of the chosen variables to assess the models with, as well as the use of single reference (observation-based) datasets to compare the models with. I would usually request that more datasets get incorporated as I view the use of only single datasets, when others exist readily, as not useful to truly evaluate a model. However in this MS I see the model results are sufficiently poor as to not make much difference. I don't mean that in a disparaging sense, most coupled models perform relatively poorly against reference datasets. Yet, I suggest for future papers to not rely upon single reference datasets unless only one really exists.
Our paper is indeed not a full evaluation of model performance. Instead it is meant as a first assessment documenting the plausibility of simulation results. As a global Earth system model, ICON-ESM is not meant to reproduce observations in detail, but only general patterns. Obvious reasons for this are the usage of global parametrizations and missing representation of much process detail. But a particular challenge for globally coupled models is the mutual imprint of model bias between separately developed large model components (atmosphere, land, ocean, cryosphere) that makes it necessary to tune components against bias in other components. Note that ICON-Land is not meant as a separate Dynamic Global Vegetation Model (DGVM) but developed to be the land component of an Earth system model. Thereby in the interest of an overall acceptable performance individual components must generally perform poorly against observations. For such models it is

therefore in particular in early stages of model development (ICON is brand-new!) part of the modelling strategy to aim only at reproducing system behaviour in the large. Assuming that different observational data sets for the same variable all reproduce the large scale patterns, the inclusion of further data sets would only in exceptional cases reveal further relevant insight into system behaviour and would in addition further reduce the readability of our paper.

You needn't cite this or anything, but I would point out that land surface models typically do poorly for LAI, e.g. see Seiler et al. 2022, so the model poor LAI is by no means unique.
Seiler, C., Melton, J. R., Arora, V. K., Sitch, S., Friedlingstein, P., Anthoni, P., Goll, D., Jain, A. K., Joetzjer, E., Lienert, S., Lombardozzi, D., Luyssaert, S., Nabel, J. E. M. S., Tian, H., Vuichard, N., Walker, A. P., Yuan, W., and Zaehle, S.: Are terrestrial biosphere models fit for simulating the global land carbon sink?, J. Adv. Model. Earth Syst., 14, https://doi.org/10.1029/2021ms002946, 2022.
We agree and think it makes sense to mention this as it gives context to the assessment. We included the sentence
"This is not unique for our model (see Seiler et al., 2022)"
in line 629 and the corresponding reference in line 863.

line 20 - what does 'weaker' FAPAR mean? Lower?
We added *"lower"* in line 21.

Why compare against NPP rather than the more directly observable GPP? GPP has at least four independent global products that could be compared to. NPP 'observations' on a global scale are always going to be a fairly derived/modelled product. Similar question for WUE, why not consider the ET and GPP themselves rather than a variable derived from them?
From the perspective of global carbon cycle modelling (which is next step in the development of ICON-ESM) one must get NPP right, because its NPP not GPP that determines carbon storage. Indeed, a precondition to get NPP right is to obtain proper values for GPP and autotrophic respiration. And since GPP data products are from an observational perspective more reliable than NPP products we understand the reviewer's suggestion. Nevertheless, for the purpose of our paper to document the present model development stage also in view of further model development towards inclusion of the global carbon cycle we can at the moment easily do without GPP but not without NPP. And the paper is already quit long so that we decided to not consider GPP in addition. Concerning WUE, the situation is quite similar. From a modelling perspective it is for a first assessment more interesting to get insight into an integrated quantity like WUE than its individual components.

Lines 53 - 55 seems to imply that the land carbon cycle is switched off(?), if so, how do you do NPP? Even after fully reading the paper I still found this part puzzling. The paper seems to suggest no C cycle then it proceeds to compare LAI (so leaves made from C) and NPP (C again).

We understand that the reviewer is puzzled: The JSBACH phenology model is not carbon based but rather phenomenological describing the development of the LAI by a combination of phases of logistic growth towards a prescribed PFT-specific maximum LAI and exponential drop of leaves. And the onset and end dates of theses phases are triggered by the local environmental conditions (temperature, relative soil moisture). Because of this peculiarity of the JSBACH phenology model the growth and maintenance respiration cannot be determined from the amount of allocated carbon in living tissues. Instead the calculation of these respiration terms follows the implementation of the BETHY model [1], where they are calculated as a prescribed fraction of GPP, in exceptional cases modified by prescribed limits to allocation (Reick et al., 2021). This is the reason why NPP belongs as the other variables assessed in our paper to the 'fast' processes of JSBACH that can be assessed without considering carbon storage in tissue pools. To prevent this confusion, we replaced the imprecise wording "land carbon cycle" in line 55 so that the sentence now reads

"*These variables represent processes that are fast compared to the cycling of carbon between land storage pools or climate induced biogeographical changes in landcover.*".
Moreover, we added to the NPP section 2.3.6 of the manuscript the sentence:
"Note that autotrophic respiration is calculated as a prescribed fraction of GPP instead from carbon allocated in the different plant tissues. Accordingly, NPP belongs as the other variables considered here to the 'fast' variables of JSBACH that can be assessed without consideration of the cycling of carbon between different storage pools."
in line 295.

[1] W. Knorr, Annual and Internannual CO_2 Exchanges of the Terrestrial Biosphere: Process-Based Simulations and Uncertainties, Global Ecology and Biogeography 9 (2000), 225-252.

Line 60 - this point about making sure the implementation is free of defects could actually go in the abstract. Otherwise the last line of the abstract sounds a bit optimistic since the rest of the abstract seems to suggest that v3 and v4 share many of the same biases, which makes a reader wonder why there is so much optimism in the abstract's last sentence. To be more plain - reading the abstract without knowing that v3 and v4 are almost identical except for the framework implemented in makes one think the problems in JSBACH must be really stubborn to not improve much between versions! This point seems to have confused Ref #2 as their first comment implies they missed that one of the main motivations for this paper is to ensure the implementation was correct.
True. As suggested we include this point in the abstract in line 23 by replacing its last sentences by the new formulation:
"Overall, the biases found in the different assessment variables are either already known from the previous implementation in MPI-ESM1.2, or have changed because of the coupling with the new atmospheric component ICON-A. As discussed, there is a good perspective to mitigate these biases by an improved processes representation. Accordingly, this study demonstrates the technically successful completion of the re-implementation of JSBACH into ICON-ESM-V1."

Table 1 should list number of snow layers so the older versions two layers can be listed alongside the newer versions five.
True. We listed the number of snow layers now in table 1.

line 142 - how can wood turnover not be implemented but you can calculate NPP? Doesn't the NPP contribute to plant tissues which would need to turnover to present a continual buildup. I still don't get this part.
Please see our answer to your comment on lines 53-55 of our manuscript.

line 198 - 'skin reservoir on the surface' = ponded water?
"Skin reservoir" is the hydrological terminus technicus for the storage pool of water from precipitation and dew on the canopy and the ground (see e.g. [2]). However, to avoid confusion about this term we changed the sentence to
"In JSBACH, TWS is the sum of water stored as snow and water on the surface and the canopy, as soil water and soil ice, and as runoff."
in line 200.

[2] P. Viterbo, L. Illari (1994). The impact of changes in the runoff formulation of a general circulation model on surface and near-surface parameters, Journal of hydrology, *155*(3-4), 325-336.

line 199 - 'as or in' - fix, confusing as is.
We deleted "or in".

line 244 - 'a high LAI typically reduces albedo', really? Grass with an LAI of 2 will still be lighter (higher albdeo) than needle leaf evergreen forest with LAI of 4 so I am not sure what is meant there.
Sorry, but as far as we see your example confirms what we are saying: a high LAI (needle leaf evergreen forest with 4) has a lower albedo than a low LAI (grass with 2). Nevertheless, we agree that this statement is too bold: how albedo is affected depends not on the LAI alone but also on the albedo of the underlying ground and the presence of snow ('snow masking'). Therefore increasing LAI may, depending on the situation, raise or lower the albedo. We thus drop the reference to albedo in the respective sentence and keep only the remarks on transpiration and productivity. The modified sentence thus reads:
"A high LAI typically enhances transpiration of water and in general also primary productivity."
in line 247.

Why compare against only one global LAI product? Is there reason to suppose that product is without biases? Similar question about the other variables. It is not sufficient to compare against only one observation-based dataset if other reliable ones are available. To do otherwise assumes that all 'observations' are without bias - a naive assumption at best.
We agree that it would indeed be naive to assume that "all 'observations' are without bias". But as already mentioned above, our modelling strategy is not to get simulation results correct in detail, but only in the large. Our assumption is that for this purpose any generally accepted observational product will do.

Why move from two snow layers to five? I realize the full logic might be outside the scope of this paper but it puzzles me as to what would be gained. Is it just to have finer discretization?
Yes, we intended a better representation of the heat fluxes into the ground. Anyhow, the term two layer for snow on/in the soil and on the canopy is misleading. Actually there is only one snow layer on the soil and the snow on the canopy is addressed separately in our paper. Therefore we changed "two" to *"one"* in line 132 and in table 1.

Figure 1 is a nice inclusion
Thanks.

Fig 6 - does 'year' in the top two plots stand for annual mean? Perhaps make that more clear. Same in fig 7, table 4, and perhaps others.
We replaced "year" in all tables and figures with *"annual"*.

line 409 - RSM is not defined prior to use.
We realized that the "relative" it not necessary in the term precipitation-RSM feedback. Therefore we replaced RSM in line 418 with *"soil moisture (SM)"* and "precipitation-RSM feedback" with *"precipitation-SM feedback"* throughout the paper. We included *"Relative Soil Moisture"* at its first appearance in line 469.

line 414 - it is easy to miss that the 'general rule' comment is for January, I would suggest rewording to make clear it is not intended to be a year round phenomenon.
We changed this sentence to
"However, for both models a higher TCC has a warming effect in January – as expected from the general rule that a cloud cover tends to warm the surface in winter (Chen et al., 2000)."
in line 422.

Fig 11 - max amount of water possible above wilting point? Is this field capacity or ? It is better to use more standard terms as the max amount of water possible above wilting could include inches of water ponded on the surface - the definition given is not excluding that.

Yes, 'maximum amount of soil water' is field capacity. We thus reformulated the respective sentence in the caption of Fig. 11 into
"RSM is calculated as the ratio between the amount of soil water above wilting point and the amount of water between field capacity and wilting point. RSM thus characterizes the soil conditions concerning water stress of plants (the plant usable field capacity)."

line 513 - 'The above named opposing northern mid latitudes NPP biases'. Consider revising for clarity, since it is new para it is hard to know what the above named is.

We reformulated the respective sentence into:
"The shift in NPP bias from JSBACHv3 to JSBACHv4 seen in northern mid latitudes in July has several reasons."
now in line 521.

L 562 - fix ref to appendix
Done.