Second round review of "Does dynamically modeled leaf area improve predictions of land surface water and carbon fluxes? – Insights into dynamic vegetation modules" by Westermann et al.

In their major revisions, Westermann et al. have substantially improved this manuscript which investigates the model performance of two land surface models (LSMs) when utilising static vs dynamic vegetation representations. The initial round of reviews resulted in many reviewer comments to be addressed and, in general, these have been resolved. The authors must be commended for their effort in the responses. The manuscript is cleaner and the message clearer.

Despite the improvements realised in the first round of author revisions, I still have some reservations regarding the implementation of this study. These are detailed further below along with some additional minor technical comments. As such, I regretfully recommend another round of revision before the manuscript can be published.

General Comments

1. I still find the choice of Noah-MP in the study perplexing. One of the main thrusts of the manuscript, as evidenced by the title, is gaining insights into the carbon fluxes of eddy-covariance sites when using LSMs with different vegetation initial conditions and representations. However, since Noah-MP does not produce GPP/NEE output when run statically, a sizeable amount of potential data/analysis is lacking. For example, Figures 3, 4, 7, 8, and A1 and Tables A2 and A3 highlight this conspicuous unavailability of information, where effectively only a single LSM is being used to assess the static/dynamic vegetation influence on carbon fluxes.
   In their response to this issue in the first round of reviews, the authors justify their choice of Noah-MP and ECLand as "both models can be and are widely used for coupling them as LSMs with established climate projection models." This is true for many LSMs and as such is not particularly convincing. However, I understand that access to resources and expertise for particular models is a challenge and the authors likely used the two models they could reliably run. In light of this, I believe there are two potential avenues for addressing my concerns, and I would be very interested if either is acceptable to the authors. Firstly, the manuscript could be amended to focus more on the water fluxes since all model runs output the relevant data for these. The carbon fluxes would then be additional/supplementary information and the missing outputs would not be as detrimental to the message. This option would result in a change of title

and reordering of the manuscript to address latent heat, evaporative fraction and soil moisture first.

Alternatively, the manuscript could instead focus on specifically assessing ECLand, and use Noah-MP as a benchmark. Again, this would reduce the impact of the missing carbon fluxes from the static Noah-MP. As well as changes to the text to focus more on ECLand, this would require the figures to be rearranged to emphasise the ECLand results.

Neither of these options would necessitate additional model runs, and I believe would minimise the amount of work required from the authors to mitigate the apparent issues of missing NEE/GPP data. Of course, it is possible that this aspect of the manuscript has been considered appropriately addressed in the author response by the other two reviewers, in which case I am happy to defer to the majority.

2. In a similar vein to the above, I am also unconvinced by the authors' response to the comments on their site selection criteria. I agree with the authors that the FLUXNET dataset is biased both geographically and relative to vegetation types. However, I would argue that individual sites can exhibit unique behaviour and may not be representative of their "aridity - PFT" class, and that this is far more likely to influence results negatively than by (further) introducing the well-known and understood biases from the heavily skewed location of FLUXNET sites. There are ways to reduce the impact of the PFT-aridity class imbalance that would exist if more sites were selected. For instance, contributions to the aggregate model performance metrics could be weighted by PFT-aridity class size. In fact, most figures and results in the paper are discussed on a site-by-site basis (e.g., the Taylor diagrams) and so the class imbalance would not present issues here.

3. The comparison of LAI output from the static model runs to MODIS LAI is another aspect that continues to trouble me regarding the applicability of the results from this study. For the dynamic runs, it is understandable as the vegetation evolves away from the initial inputs and so the use of MODIS data as an initial condition avoids any circular comparisons. However, under static runs, it would appear to me that the study is simply comparing the same MODIS data at different levels of time aggregation with zero influence from the models. Hence I struggle to derive any messages from this analysis for future model development.

Technical Comments

4. Line 1: “the surface” is not clear. It would be better to use “the Earth’s surface” or “the land surface”, for example.
5. Line 3: “some of these models”. Some of which models? “Some land surface models” or similar would clarify this.
6. Line 7: add an “and” between “the FLUXNET2015 dataset” and “the MODIS leaf area”.
7. Line 11: I would argue that latent heat flux is both a vegetation- and hydrology- related variable and therefore this sentence is not quite correct.
8. Line 93: “we assumed them to be neither very predictable nor very unpredictable in total” - I think this needs clarification.
9. Line 103: Include the citation to the FLUXNET website.
10. Line 108: spelling of “tends”.
11. Line 110: Cite the Climate Data Store properly.
12. Line 180: It should be “Noah-MP” not “the Noah-MP" and “a global soil grid” not “the global soil grid”.
13. Line 182: Initialising all LAI values based on Table 1 for model runs starting on January 1st would misrepresent the four Australian sites and may cause model performance issues.
14. Line 189: spelling of “therefore”
15. Line 189: “Other options were used as their defaults” is not clear. I recommend “All other settings used default configurations” or similar.
16. Line 257 - 259: I would suggest explicitly explaining how this follows from the figures e.g., that the symbols are in the same location / there are no arrows.
17. Figure 2: This was raised in the first-round reviews, but the arrows should not extend beyond the plot area. I understand that this is because the normalised standard deviation of the static run falls outside the plot limits, but this is not acceptable. The axes must be extended such that the arrows are fully located within the plot area.
18. Line 269: It is unconventional to refer to the performance in the static runs as having “increased” when these runs are the ‘baseline’, and this data is plotted as the beginning of arrows.
19. Line 285: Is it possible to quantify the increase in arrow length in Figure 4? This would be preferable to the qualitative use of “longer arrows” in this instance.
20. Line 294: Similar to comment 19, can the “scattered more closely” be quantified?
21. Line 296: Is it not the case that the sites with the “best” performance depends on how one prioritises the metrics, or are the 12 sites mentioned the best performing across all three metrics used?
22. Figure 3: Caption uses “die” rather than “the”.
23. Line 314: Why does CH-Oe2 exhibit such improved performance compared to all other sites? Does this have any lessons for model development?

24. Line 330: “Despite being low” - to what is this referring?
25. Line 351: “less uncertainty” is not the terminology to be used here. Maybe “weaker”?
26. Line 355: This reads as though it is introducing Figure 8 but Figure 8 has already been discussed in the previous paragraph.
27. Line 366: I would check the literature for examples of MODIS LAI being inaccurate for tropical sites.
28. Figure 8: I suggest rearranging the panels so that the facets are, in descending order, “Observation”, “Static ECLand”, “Dynamic ECLand”, “Dynamic Noah-MP". This keeps the ECLand runs next to each other, but also places the dynamic runs adjacent to each other as well.
29. Figure 8: The caption refers to the fitted linear regression models as “applied as additional information” which does not read correctly. I would delete “as additional information” in this instance.
30. Line 380: I would argue that comparison of modelled and observed fluxes on a daily basis is performed more frequently than “rarely”.
31. Line 410: What is the Noah-MP Crop module?
32. Line 420: In which scenario was a frequent reset of LAI applied to ECLand as compared to the other studies? I do not follow where this was applied and had no effect?
33. Line 425: “low predictive efficiencies” is unusual terminology. Maybe “low predictability” or “low predictive power”?
34. Line 441: “inclusively LAI” should be “inclusive of LAI”.
35. Line 565: Why do the authors suggest “alternative remote sensing LAI products”? No other products were tested in this study and such products may not perform well.
36. Line 568: Haughton et al. (2016) explicitly checked the sites used in PLUMBER for observational errors. This study shares only three sites with the PLUMBER study and therefore this citation likely shouldn’t be used in support here.
37. Line 590: “Using alternative input ... but this needs to be evaluated in more detail”. Is this not what was investigated in this manuscript? What additional detail should be checked in any future studies? What are the authors’ suggestions to model developers?
38. Code and Data Availability: I suggest including the datasets in the bibliography and citing them here properly rather than the current use of weblinks. Proper citations would ensure reproducibility by containing additional information such as dataset versions, date accessed, etc.