Review

General comments

The manuscript 'Does dynamically modelled leaf area improve predictions of land surface water and carbon fluxes? - Insights into dynamic vegetation modules' by Westermann et al. aims to address the sensitivity of simulated carbon and water fluxes as well as leaf area index itself to prescribing or dynamically simulating the leaf area index in two terrestrial biosphere models as well as exploring the sensitivity to different prescribed LAI datasets. Simulations are conducted and compared against a suite of FLUXNET2015 sites. The authors find that LAI is sensitive to their simulations set-up (i.e. prescribed vs simulated LAI) but simulated carbon, water and energy fluxes are not strongly impacted. I think the research questions addressed are interesting and novel, and I appreciate that conducting the experiments and associated analysis were a lot of work. However, I have some major concerns before the manuscript can be considered for publication. I want to stress that I am aware a lengthy review can feel disheartening but my comments are meant to be helpful for revising the manuscript:) So please don't read it too negatively.

- My main concern is that the methods description is a bit confusing and I'm not sure I quite understand the experiment set-up.
    - Do you have two sets of experiments, where 1) you test the impact of LAI datasets on simulated LAI and carbon/water/energy fluxes and 2) where you 'simply' switch on/off the dynamic vegetation module? If so why is 1) not part of your research questions in the introduction? In parts of your manuscripts it reads like you prescribe LAI but it is dynamically simulated at the same time which I don't understand (see for e.g. caption Fig. 7)?
    - Where is LAI as an input driver coming from in the LUTs? How does it differ between the experiments (LUT vs your LAI? Is LUT also based on MODIS?) Not everyone is necessarily familiar with the look up tables of the specific models chosen for this study so it'd be good to clarify this.
    - The section where you compare LAI across your experiments almost seemed a bit circular to me, and I would suggest to reduce the emphasis on LAI and focus more on the simulated fluxes where you can avoid the interdependence of input and output LAI during evaluation (and this is also appropriate given the title of the manuscript). Alternative remotely sensed LAI datasets are available, although this comparison of course also would be a bit unfair.
    - Towards the end of your results/discussion section you describe what's happening in the model and how this explains some of your model results which is great! I think it could help your manuscript if in the methods the model descriptions had more detail too for the relevant processes
    - Throughout your manuscript it would help readability if you had specific experiment names that are consistently italic (or any other distinct formatting) like you attempted in L158.
    - Other small things I would like to suggest are:
        - Split the Results and Discussion section - the way it is written now, it is a bit of a back and forth and hard to follow.
        - Include a table with ALL experiments listed (expand Table 1).
        - Part of my confusion also stems from how the results are presented, and simply including headings could really help. Something like: 3

Results/ 3.1 LAI /3.1.1 Impact of different LAI initial files/ 3.1.2 Prescribed vs dynamic LAI (and so on for the remaining variables).

- You need to be more careful with the metrics you used for model evaluation, see below for specific comments
- I was also a bit surprised about your model selection? Why did you choose a model that couldn't provide all necessary outputs for all simulations you conducted?
- Why did you initialize your model simulations differently (ECLand vs Noah-MP)?
- You report that dynamically simulated vegetation leads to a lower model performance, at least in LAI. One thing I wondered is whether your model simulates the 'right' vegetation type for each site you considered (or do you define the vegetation type that is simulated)? You also point out multiple times how forests tend to show better model performance than shorter vegetation types, but you don't offer any explanations why that might be the case?

- The overall structure of the introduction makes sense to me but sometimes it is hard to tell what message you try to convey? I know this is quite wishy washy but I provided more specific comments below.

Specific comments

L7 Maybe change to 'We compare model results with **observed** fluxes from the FLUXNET [...] or similar
L8 MODIS leaf area **index** ?
L8 More detailed information? What does this mean? If the only additional output is LAI, you might as well explicitly state this here but in general I think this is weirdly specific for an abstract the way it is written now
L13-14 This is not really a reason that explains weak model performance, but just another way to phrase poor model performance! I think also the abstract is not a place for speculation but you should clearly state what the drivers for poor model performance are based on your study

L21 You could already state in the first paragraph where LSMs are used (you do give the CMIP example in L26 but LSMs are also used in meteorology models, reanalysis [...]) before diving into the more specific applications to motivate your study, and from there go to your model validation topic
L26-L36 In general this paragraph discusses schemes to evaluate model performance, which I think is a useful topic for your introduction! But I'm not sure what the key point is you're trying to make here. Are you trying to build up to presenting a new evaluation scheme in your paper?
L24-25 This doesn't offer a lot of information. For example, you could mention why more features are added to LSMs to give this sentence more value
L28 I suggest 'global, regional, and site scale'
L29-31 I'm not sure I understand this. Do you mean that evaluation schemes are compared against each other? Or model - obs comparisons?

L34-35 This is also unclear to me. Did they evaluate some LSM ensemble average against statistical methods? I also don't understand how not reporting individual model performance is linked to only having normalized metrics, and also why normalized metrics are not useful. Could you explain this a bit more?

L37 Replace 'had a closer look' with explored, investigated or something similar?
L40 This is unclear to me. Do you mean that they didn't find an error in the observations they compared the model simulations to?
L41 I get your point but to me this almost reads like model-observation comparisons can't help identify areas of uncertainty at all which is not true (see e.g. Whitley et al., 2016 for one of the many studies that identify reasons for inter-model differences and mismatches with obs)

L45 Could you name a few references to support your statement here?
L46-47 This seems a bit out of place and unnecessary to focus on a single benchmarking dataset that as far as I can tell you're not even using in your study? I think the first two sentences are a bit dis-connected from the rest of the paragraph anyway where now all of a sudden you focus on drought and water-limitation (why?)

L66-67 Why did you choose ECLand and Noah-MP? Other than them being used by others
L74-75 I think you can drop this:)
L80 Why did you invert the Aridity Index? I know you took it from a dataset but it'd be good to know how it is calculated
L83 Why +- 0.1 log AI? Is this a common threshold?
L87 I'm not sure the predictability of your sites necessarily follows from your selection procedure. But maybe I misunderstand - can you explain this a bit more?
Fig.1 Can you also explain the colorbar in the Figure caption?
L97 Why was precipitation set to zero in the gapfilling?
L97 Is using a Kalman filter a common procedure for gapfilling? How is gapfilling achieved in FLUXNET?
L97-98 ERA5 is relatively coarse if I'm not mistaken, and it'd be worthwhile mentioning the shortcomings and the reason for choosing this reanalysis? I assume you chose it because of the high (3h?) temporal resolution which would also explain the 3h cut-off between Kalman Filter and using ERA5
L100-101 What's the cut-off for 'longer gap filled' periods?
L103 You already defined LAI in L59
L104 What is the temporal and spatial resolution of MODIS LAI?
L105 I know you point to the documentation but could you please explain the choice of quality flags (and what they mean) in the manuscript?
L106-107 I'm not sure I'm following. How does selecting certain quality flags help having the 'same amount of data in a month'?
L112 You quite happily use LUT LAI as if this was a commonly known dataset but personally I have no idea where the data in your look up tables are derived from. Is this look up table derived from MODIS, which years were used for it [...] - you don't have to state every little detail here but perhaps there is a reference with a description of how this look up table was derived?
Table 1 The way you described *MODIS climatological LAI* and *MODIS single-year LAI* in the table is not super clear. The first one is just a monthly climatology (? based on which years),

the second is the first year of the simulation period on monthly timesteps? Could you rephrase this?
L127 ‘under-development’ does this mean the model isn’t ‘ready’ yet?
L129 Is the daily resolution really an appropriate reason to choose those quality flags? Why did you avoid smoothing your data for LAI?
L133-134 So only two vegetation types per grid cell are simulated?
L134-142 It’s great you explain a bit of the model, but which photosynthesis model is used? It would also be good to know how LAI interacts with GPP/NEE, and also with heat fluxes because these relationships are key to your analysis - but of course I appreciate you can’t explain every line of the model code here
L149 ‘taken from LUT’ I would replace this with ‘were prescribed’ or similar
L152 ‘Thereby’ I’m not sure this follows from the previous sentence
L150-156 Again, I think it would be interesting to know which photosynthesis scheme Noah-MP employs, and to get a rough idea how LAI interacts with GPP/NEE and heat fluxes etc in the model

L164 What does global initial data mean here? Did you conduct the model spin up using ERA5? Why didn’t you choose the meteorology that comes with the flux sites (like you did with Noah-MP)?
L166 Which site information - do they all report vegetation height and you used a certain threshold?
L169 Is there a reference for the van Genuchten soil hydrologic params?
L171 Could you please update this reference - PLOS ONE is a journal and not the dataset name, and the author of the dataset surely is not ‘The PLOS One Staff’!
L172 So the initial conditions for the two model runs are not identical? Why not?
L173 I would have thought each gridcell has 8 neighboring gridcells?
Table 2 Could you include the full name of the vegetation types and also the actual name of the Noah-MP vegetation classes? I’m not sure I mentioned it elsewhere but a similar table for ECLand would be great too
L183 What does ‘transferred’ mean here? Did you calculate daily averages/sums?
L188 You are also using the Pearson Corr. to assess non-linear relationships but this is not a suitable metric (Pearson Corr. only describes linear relationships; see further below)
L189-190 Which symbol represents the normalized standard deviation metric?
L191-193 I’m not sure I understand why you subtract the minimum from $\bar{x}$? I also wonder how you interpret the relative bias when x_{min} differs strongly from $\bar{x}$?
L199-200 So you try to understand the relationship between output variables within the same model simulation?
L199-207 I don’t really understand how you calculate the elasticity. What is the ‘slope of their correlation’? Is there a reference that uses the same definition for elasticity you could include? Is the -0.1 - 0.1 threshold common (reference)?

L213 ‘The model performance of the dynamic run is shown with the symbol’ - which symbol ?
L217-218 Can you name some references to support ‘in line with results in the available literature’? It would also be useful to see some metric -> same for the next sentence about Noah-MP. How much more biased??

L222 Here e.g. you compare the LUT LAI runs with your generated LAI forcing but in order to fully understand the performance difference it would be useful to know where this mysterious LUT LAI comes from??
L224 ‘simulation results were unaffected by the type of LAI forcing with vegetation dynamics switched on’ I must have completely misunderstood your experiment set-up, but I thought when vegetation dynamics are switched on, LAI is simulated (not prescribed?)
L225 ‘but not necessarily for all sites’ can you quantify this more?
L226 Here for example (and across the entire section) it would be helpful to stick to the nomenclature for your experiment you defined in the table earlier, and highlight it with a different font type
L226 ‘increased variance’ compared to what?
L227 ‘was random’ - what does this mean?
L231 What are sparse vegetation types (i.e. did you define it anywhere)?
L243-244 Can you also give a value range for ECLand - it’d help the comparison between the models
L246-252 Here you kind of say a bit about where original LUT is coming from, and this belongs in the methods section already
L255 Why are you not showing the single-year LAI results - could just be in the appendix?
Fig 2 How did you set threshold to separate the AI into different aridity classes? This belongs in the methods
Fig 3. I think you could use more distinct colors here, at least on my copy it’s quite hard to see the difference in the shade of gray. Also why are there only single symbols for NEE and GPP (for dynamic simulations I think?) for Noah-MP? I also don’t understand why you have different values for different prescribed LAI data for the dynamic simulations - but this is because I didn’t quite understand the experiment set-up, as I said above. Does LAI not emerge from the spin-up (and therefore doesn’t need to be prescribed)? The values do look very similar across all experiments. One more small thing, but the red dotted line is hard to see for the Pearson correlation and Normalized STD
L265 Here you dive into the impact of LAI-set-up on ecosystem exchange variables and it would be useful to see in the methods (at least to some degree) how LAI is actually linked to those variables?
L281-282 Then why did you use Noah-MP in this study? This kind of defeats the purpose of your study, at least according to the title
L284 I’m not sure this is the right conclusion - did you also find an overestimation of GPP? If NEE is right for the wrong reasons for the historical period that doesn’t mean you can have a lot of confidence in simulated NEE in future scenarios
L289 ‘being independent of the prescribed LAI forcing’ - so this is just the dynamic simulation that does not have any prescribed LAI?
L289-290 Not sure consistent is the right word here, those are two v different experiments
L291 How does the onsite LAI actually compare with the MODIS LAI? MODIS LAI can be biased for some flux sites (so a reason for the lower performance could just be that the model was tuned to match MODIS, and MODIS and on-site LAI are actually wildly different)
L306-307 Where is this recommendation coming from? Did you not say earlier there is no performance difference depending on the LAI forcing used?
L309-316 A lot of these studies look at different temporal resolutions, and I actually wondered too how your model simulations would do if you looked at climatologies, or annually aggregated values - this could help you identify whether the models get the broad

patterns right. But having said that, you already cover a lot of ground with your study so this is more of a curiosity rather than an actual suggestion for the revision
L329-330 Interesting indeed - do you have any idea why this is happening?
L335 Again - why not include a figure in the appendix to show these results?
L361-362 Here it is so much clearer what your elasticity metric is meant to do! In general, I really like this part of the analysis
L351 How are vegetation and soil moisture state variables coupled?
L354-359 I think this is all valuable and true, and explains the general soil moisture bias well. But it doesn't explain your actual results, i.e. not seeing an impact of static vs dynamic vegetation on soil moisture
Fig 7 Where is the footnote for elasticity?
L389-396 Why didn't you show LAI-GPP elasticity in Fig 7 when this is such a strong focus in this paragraph? Do you have an idea whether it is more realistic to have a linear or non-linear LAI-GPP relationship (i.e. what do studies say that look at observed LAI-GPP relationships)?
Fig 8 What are the arrows in the figure? Some of the relationships in this figure (as you point out earlier) are clearly non-linear, and therefore using the Pearson correlation coefficient is not suitable. I would also suggest to change the GPP units to avoid having so many decimals - in this figure and also in the text
L403 You need to define σ_r
L409-410 You already say in the methods that you are masking MODIS using quality flags (and as I said there, please be more precise about what the flags mean)
L421-431 I like that you are explaining the model relationships here and link them with the results
L434 typo 'dyanmic'
L454-455 Can you give a reference for this ('real LAI'), also maybe replace real with 'observed' or similar
L461-462 I think this is a bit harsh and not true; depending on the focus (i.e. if it's purely benchmarking) of course this can happen, but there are also many papers that actually explain reasons for mismatches between obs and model simulations, and also differences across members of the LSM ensembles (see e.g. Whitley et al. 2016)
L467 Why did you expect that dynamic vegetation would improve ecosystem exchange especially for short vegetation?
L480-481 If you don't know how representative your models are compared to other LSMs then why did you choose them? I see both are part of the PLUMBER2 experiment - how do they compare to other LSMs there (see Abramowitz et al., 2024)?

TableA1-A5 are not referenced in the manuscript

You did not include the data and code availability statement!!

References
Abramowitz, G., Ukkola, A., Hobeichi, S., Cranko Page, J., Lipson, M., De Kauwe, M., Green, S., Brenner, C., Frame, J., Nearing, G., Clark, M., Best, M., Anthoni, P., Arduini, G., Boussetta, S., Caldararu, S., Cho, K., Cuntz, M., Fairbairn, D., Ferguson, C., Kim, H., Kim, Y., Knauer, J., Lawrence, D., Luo, X., Malyshev, S., Nitta, T., Ogee, J., Oleson, K., Ottlé, C., Peylin, P., de Rosnay, P., Rumbold, H., Su, B., Vuichard, N., Walker, A., Wang-Faivre, X., Wang, Y., and Zeng, Y.: On the predictability of turbulent fluxes from land: PLUMBER2 MIP

experimental description and preliminary results, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-3084, 2024.

Whitley, R., Beringer, J., Hutley, L. B., Abramowitz, G., De Kauwe, M. G., Duursma, R., Evans, B., Haverd, V., Li, L., Ryu, Y., Smith, B., Wang, Y.-P., Williams, M., and Yu, Q.: A model inter-comparison study to examine limiting factors in modelling Australian tropical savannas, Biogeosciences, 13, 3245–3265, https://doi.org/10.5194/bg-13-3245-2016, 2016.