# Response on Referee #1's second round review

First of all, we want to thank reviewer 1 for accomplishing a second round of reviewing this work of me and my co-authors. In the following, I will go through and respond to your comments. Please note that reviewer comments are in italic, our responses in normal font and an explanation of changes/adaptations made by us in blue font. In case, your suggested terms/phrases were incorporated without further adaptations, I refrained from copying them again to this document.

**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.*

   Thank you for providing further ideas for dealing with the limitation that Noah-MP is only partly suitable for answering our research questions. We now decided to keep the intention and the research questions of the manuscript as they are but, instead, changed the source code of Noah-MP a bit so that GPP and NEE are calculated also for the static simulations. As a result, GPP is calculated from the current conditions for photosynthesis and NEE results from that GPP and an estimate of respiration of the current available biomass. All other variables that usually would have been dynamically computed within the same function were reset to their values before calling the

function. We checked whether our adaptation had an impact on the predictions of latent heat flux and could not find any.
We explained that adaptation in the method section (L160-161), added some results for model performance of NEE and GPP (L295-308), included the LAI-NEE and LAI-GPP relationships for Noah-MP in Figure 7 showing the elasticity and changed the paragraph L434-448 in the discussion section a bit according to the new results.

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.*

   Yes, this issue is specifically addressed in Haughton et al. (2018a), a publication we are aware of and taken into consideration. Those authors found no relationship between the uniqueness of a site and its PFT or data record length and, thus, both characteristics do not legitimize site selection. However, at least for dryness (aridity) there is some evidence that it determines the uniqueness of a site. Furthermore, they also said that it is more important to communicate how and why a certain site selection was done, which we did. To address the comment, we now explicitly checked the predictability measures from Haughton et al. (2018a) for our selected sites. They had a RMSE uniqueness between 0.03 and 0.41 for NEE, 0.04 and 0.43 for sensible heat flux, and 0.04 and 0.30 for latent heat flux. Note that this is in a comparable range for all three variables, avoiding that one of our target variables could be more predictable on average. Second, this confirms that our dataset includes sites with medium to high predictability. Indeed, our selection ended up not including sites with low predictability which is against the suggestion by the authors to include the whole spectrum of predictability in a model evaluation study. At the same time, our selection strategy yielded sites that are less unique and, thus, more representative. In other words, site uniqueness did not affect our results. Indeed, we argue that less predictable sites would not allow enhanced insights: since dynamic vegetation modules of the two LSMs we tested here are already challenged to predict vegetation processes even for (very) predictable sites, they are unexpected to improve for sites with less predictability which yields no additional information. To check whether site predictability affects the sensitivity of the model performance to vegetation dynamics, we tested whether greater changes in model performance for e.g. NEE coincides with less predictability and whether more unique sites show greater sensitivity to activating vegetation dynamics, and found no evidence for both. We argue that our site selection captures "adequate diversity of site characteristics" for the purpose our investigation as proposed in Haughton et al., (2018a).

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.*

   Agreed, it was expected for the static runs with MODIS climatology to have good performance (since it's the same data but averaged and with coarser temporal resolution for input)
   However, we also had the expectation that dynamic runs should better capture single values from MODIS resulting in higher performance, even higher than with MODIS climatology which is not the case. In order to show that, we needed the assessment of the static runs with MODIS climatology as a basis for comparison. We cannot state that dynamic vegetation modules perform worse than a more site-specific climatology if we wouldn't have done this test. We have explicitly stated in the paper why we use those runs (L219-222). Thus, the message is: best performance can be achieved by using a more site-specific climatology, since dynamic vegetation modules still need improvements before they will do better than that.

**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.*

   Changed to "the land surface" (L1).

5. *Line 3: "some of these models". Some of which models? "Some land surface models" or similar would clarify this.*

   Changed to "Some land surface models" (L3).

6. *Line 7: add an "and" between "the FLUXNET2015 dataset" and "the MODIS leaf area".*

   Done (L8).

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.*

   Agreed, latent heat flux is largely determined by the vegetation but, in the end, it is a flux of water and, thus, part of the water cycle.
   Changed to "the performance regarding variables of the carbon and water cycle was unrelated for both models" to make it more clearly (L11).

8. *Line 93: "we assumed them to be neither very predictable nor very unpredictable in total" - I think this needs clarification.*

This relates to the predictability of FLUXNET sites found by Haughton et al. (2018a). Please see the answer to comment 2.
The sentence is now reformulated to “We were left with 24 sites, covering a wide range of site characteristics as recommended by Haughton et al. (2018a) including aridity, vegetation types and observation periods (Fig. 1)” (L92-94) and left out the reference to site predictability since this is not the focus of our site selection.

9. *Line 103: Include the citation to the FLUXNET website.*

   Citation added as “(fluxnet.org, 2020)” inclusively bibliography entry (L102).

10. *Line 108: spelling of "tends".*

    Done (L107).

11. *Line 110: Cite the Climate Data Store properly.*

    Citation added as “(Copernicus, 2018)” inclusively bibliography entry (L109).

12. *Line 180: It should be "Noah-MP" not "the Noah-MP" and "a global soil grid" not "the global soil grid".*

    Done (L181).

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.*

    Absolutely right.
    Added those values in brackets and explained in the table caption as “The values in brackets for Noah-MP initial LAI refer to sites on the Southern Hemisphere due to shifted seasons” (Tab. 1).

14. *Line 189: spelling of "therefore"*

    Part of the sentence was changed to “…with using matric potential limitation” (L190).

15. *Line 189: "Other options were used as their defaults" is not clear. I recommend "All other settings used default configurations" or similar.*

    All other options for the Noah-MP simulations are now explicitly listed in Table 2.

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.*

    Added “…since the symbols in Figure 2 c+d have the same positions” to that sentence (L258-259).

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.*

    Axis for correlation coefficient extended in the new Figure 2. The same appeared for Figure 4, Figure A1 and Figure A2.

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.*

    The "increased" here refers to the change in model performance by using MODIS climatology compared to using default climatology.
    Changed to "…using *MODIS climatology* instead of *default climatology* in static simulations…" (L268).

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.*

    In principle, it would be possible to quantify the arrow length since it is approximately the hypotenuse of the change in correlation coefficient and the change in normalized standard deviation. However, this quantity has no interpretable meaning.

20. *Line 294: Similar to comment 19, can the "scattered more closely" be quantified?*

    Quantified as the average deviance from nstd=1. Added "…Noah-MP seemed to capture NEE representations better as the mean deviance from a normalized standard deviation of 1 was 0.33 (ECLand: 0.39) …" (L309-310).

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?*

    The "best" performing sites here referred to their position in the Taylor diagram and, thus, to the combination of correlation coefficient and normalized standard deviation. Including the relative bias into that ranking would give a different picture.
    Adapted the sentence to "…best sites regarding NEE correlation and variance were forests…" (L311).

22. *Figure 3: Caption uses "die" rather than "the".*

    Done (Fig. 3).

23. *Line 314: Why does CH-Oe2 exhibit such improved performance compared to all other sites? Does this have any lessons for model development?*

The site CH-Oe2 shows improved performance not for all variables but consistently large changes in model performance. After checking the initial file for the ECLand setup, the reason can be found in low LAI values of 0.16 throughout the year for the default climatology. The grid cell with the flux tower in the global setup, where the initial file was taken from, contained no short vegetation. However, during the setup of the default climatology simulations, the vegetation on that grid cell was assigned to be low crops (coverage of the low vegetation = 1) to fit the sites descriptions. This means that the low LAI values were used as climatology since they remained unchanged (which is one of the conditions of the default setup). Consequently, switching on vegetation dynamics or replacing the LAI climatology had a large impact on modelled fluxes and their performance. There is not really a lesson to learn for model development rather for preparing land use datasets for model setup since they might miss some patches of different vegetation types.
Added the sentence "The big exception appeared for CH-Oe2 which was caused by its default LAI climatology that did not fit the vegetation type" (L328-329).

24. *Line 330: "Despite being low" - to what is this referring?*

To the improvement in model performance for soil moisture when activating vegetation dynamics.
Changed to "Some sites showed improvement of soil moisture prediction by activating vegetation dynamics for both models although the improvement was very weak" (L345-346).

25. *Line 351: "less uncertainty" is not the terminology to be used here. Maybe "weaker"?*

What is meant here is the variability of the values and, thus, the scattering around the regression line.
Changed to "…considerable less variability compared to the observations" (L383).

26. *Line 355: This reads as though it is introducing Figure 8 but Figure 8 has already been discussed in the previous paragraph.*

The statements of the first paragraph in section 3.3 were distributed in different locations now. The third sentence (former Line 344-346) became the last sentence of that paragraph (Line 365-367). Instead, explanation of Figure 8 (former Line 351-353) was placed after introducing the Figure (Line 362-365). The sentences "In general, ECLand shows a linear relationship with considerable less uncertainty compared to the observations" and "The slope and intercept of the linear regression is dependent on the choice of static or dynamic vegetation" were moved to Line 382-383. The sentence "In contrast, Noah-MP shows a non-linear relationship with a pronounced hysteresis" (former Line 358) was merged with the sentence "Noah-MP shows a marked hysteresis effect at all sites except the tropical one (Fig. 8 e-h)" from former Line 364 to "Noah-MP shows a non-linear relationship with a pronounced hysteresis effect at all sites except the tropical one (Fig. 8 e-h)" in Line 377. The sentences "This hysteresis is related to the partitioning of GPP to the carbon pools in the plants. Noah-MP uses a non-linear function for allocation of GPP to the leaves that limits the maximum LAI the model can

grow" from former Line 349-350 were moved to Discussion section 4.3 and merged with the sentences in former Line 516-519 (now Line 537-540).

27. *Line 366: I would check the literature for examples of MODIS LAI being inaccurate for tropical sites.*

    Now included in Discussion section 4.4 where limitation of MODIS data is discussed by "Noisy and uncertain LAI data from MODIS for tropical forests was already reported among the literature (Weiss et al., 2007; Garrigues et al., 2008; Xiao et al., 2016; Zhang et al., 2024)" (L595-596).

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.*

    Done (Fig. 8).

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.*

    Done (Fig. 8).

30. *Line 380: I would argue that comparison of modelled and observed fluxes on a daily basis is performed more frequently than "rarely".*

    Agreed regarding heat fluxes. However, for specifically LAI, NEE and GPP, it really was challenging to find comparable model evaluation studies to discuss with, especially for ECLand. Thus, in my opinion, "rarely" is the right phrase for that.

31. *Line 410: What is the Noah-MP Crop module?*

    Noah-MP-Crop is an extension of Noah-MP developed by Liu et al. (2016) that explicitly considers field management practices for certain crops. The reference is given in the text (L493).

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?*

    We were trying to check our results for plausibility by including studies in our discussion that did data assimilation of LAI within their model runs. In our study, we have not done data assimilation per se, but rather updated the LAI input on an annual basis in the MODIS single-year setup where we could not find any effect of this additional adaptation. I guess the formulation of the statement was misleading here.
    Changed to "…but did not have an effect for the annual resolution applied here" (L432).

33. *Line 425: "low predictive efficiencies" is unusual terminology. Maybe "low predictability" or "low predictive power"?*

    Changed to “low predictive power” (L443).

34. *Line 441: "inclusively LAI" should be "inclusive of LAI".*

    Done (L457).

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.*

    True.
    Changed to “Overall, we recommend using MODIS climatology forcing for static simulations which yielded the best model performances for carbon and water fluxes. This might be valid for other remote sensing LAI products as well but would need to be tested beforehand” (L582-584).

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.*

    The methodology of measurements and quality control is standardized within FLUXNET. Thus, uncertainty from measurements should be comparable/transferable from the sites Haughton et al. (2016) considered to others. Additionally, by citing their work we did not want to claim that our sites have the same level of uncertainty but rather giving an evidence that it was already shown that poor model performance likely has other sources than the uncertainty of measured data.
    Changed to “…but Haughton et al. (2016) demonstrated that observational errors, in general, are unlikely to cause poor model performance” (L585-586).

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?*

    Surely, this was one of the aims in this study but, of course, our investigation cannot absolutely answer this question, since we tested only one remote sensing product and also had on-site LAI measurements from only one site.
    Added “…since we were limited in data sources” (L611). For suggesting future directions, we changed the last statement to “Additionally, they might be a good starting point for a similar intensive investigation with other land surface models or other alternative LAI climatology” (L625-626).

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.*

Done (L627-633).

Response on Referee #2's review

First of all, we thank Reviewer 2 for giving detailed feedback to our work. The critical remarks have helped us to improve this publication. We are sorry that we omitted to submit a complete list of the response and we have now added this below.

*Thanks a lot for the effort in revising the manuscript. Theoretically of course I'm happy to review the revised manuscript, but unfortunately the response document is set up in a way that makes it very hard to track down what changes have been made. The requested format of the author's response is online where it says*
*'The author's response in case of "minor" or "major" revisions must be submitted as one separate *.pdf file (indicating page and line numbers) structured in a clear and easy-to-follow sequence: (1) comments from referees/public, (2) author's response, and (3) author's changes in manuscript. [...]'*

*I would greatly appreciate it if this is how the author's response was structured too (see also here https://www.biogeosciences.net/for_authors/).*

*At least for my comments (referee 2) the specific comments I made were left out and 'only' the authors' responses were included so it's not immediately obvious which comments the authors are referring to, and doing this to shorten the overall document is not an argument to make a response letter incomprehensible. I would prefer having a long document and being able to follow the responses and how they are addressed in the revised manuscript.*

*I further noticed that the actual changes in the revised manuscript were not included in the response letter for any of the three reviews. I don't think every new word added needs to be defended, but at the same time responses like 'information added' or similar are not sufficient to indicate which changes were made specifically in the manuscript. Lastly, I'm not clear what the color coding in the response letter means? I assume this is to indicate what changes have been made, but if this could be made more clear that would be very helpful.*

*All reviewers are volunteering their time to review this manuscript and having to go back and forth between three documents (the original review, your responses, and the tracked changes manuscript) to get a sense of how concerns are addressed and which changes were made is quite time consuming and unnecessary given this information could just be combined into one document.*

We acknowledge and understand this comment. We apologize for omitting to send a proper response together with the last revision. We now attached below the point-to-point response to the first review of reviewer 2 below. We also give a short explanation of the formatting.

*I did briefly go through the responses and wanted to point out that I am still not satisfied by the justification for the model set up. This needs to be phrased more carefully. Including a model because 'it is still interesting to look at' isn't exactly a powerful argument - but maybe this is elaborated on in the revised manuscript which I have not read. Another thing I would like to point out is that there seems to be some contradiction: In the introduction the authors pose three research questions that read like they are aiming to generalize their results*

*derived based on a single model (given Noah-MP couldn't be run with static vegetation), but then they argue in the response to reviews that there is 'no chance in directly transferring results and conclusions' from their model to others models. This makes me wonder what the point of the study is if the study set up and results are so model specific that it is not possible to derive any conclusions for land surface models in general?*

To address this comment, we have now included additional runs with Noah-MP that mimic the static vegetation and yield output on carbon fluxes. This was done by using the dynamic version (which produces output for the carbon fluxes), but resetting the LAI every year to the original. We double checked, whether the output corresponds to the original static runs based on the water fluxes, which are output in either version. In this way we are able to show comparable results for both models, even if static Noah-MP is specifically not made to look at carbon fluxes. With regard to model selection: We really made an effort to look into the code in those two widely used models in order to understand whether and how the surprising results came to place. We agree that it would be great to do the same for many more models. But at the same time, this almost forensic investigation cost a great deal of time as not all of the reasons are obvious from the model description. We hope that our insight from those two exemplary models inspires more research in this direction. Notwithstanding that, we have rephrased the research question to specifically address ECLand and Noah-MP in the revised manuscript.

## Response on Referee #2's 1st review

Thank you for giving this detailed feedback. We think addressing them has improved the manuscript in both revisions. Below we give the missing point-to-point response of the 1st review by reviewer 2. They also include changes made during the second most recent revision.

The reviewer comments are in italic, our responses in normal font and an explanation of changes/adaptations made by us in blue font.

**General comments**

- *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)?*

  Our main focus was testing whether switching on dynamic vegetation in the models enhance their performance regarding the target variables. We changed the LAI source in order to find out whether this more site-related information as initial input "helps" the model in their prediction of LAI and NEE. However, we did not aim for doing data assimilation since there are many investigations published on that. Information on LAI is always required for initializing the models independently of whether the runs are with dynamic or static vegetation.

We handled the terminology and the descriptions throughout the manuscript more carefully.

- *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 default climatology in the initial file (what I refer as LUT LAI) of ECLand is already based on MODIS values (as mentioned in the manuscript L195-196). A time span from 2000 to 2008 and disaggregation of the gridded values for LAI was used to create that climatology (Boussetta et al., 2013). LAI values in the look-up tables of Noah-MP are defined for the plant functional types (PFTs). I could not find any information from where these values were generated from or which time span these values were taken from or how individual LAI climatology within one PFT was merged. In the default setup, this LUT LAI was used (default climatology). For the other setups, those values in the LUT were replaced by "our" LAI values from MODIS (L197).

- *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.*

  The LAI from MODIS used for model input and model evaluation is not identical. Model input is a LAI climatology on monthly basis resulting from multi-year average MODIS values. Model evaluation is done with the daily MODIS values which are 8-day means. For the static runs, this comparison provides the information whether an incorporation of more site-specific climatology results in higher representativeness of local LAI development. For the dynamic simulations, comparing modeled LAI with daily MODIS values is used to examine whether the models are able to capture inter- and intra-annual LAI dynamics. Surprisingly, we found that even with the same source of the data the dynamic simulations are not fitting the observations.
  We provided more details on the MODIS LAI data and highlighted the differences between data used for input and for evaluation (L198-222).

- *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.*

  Addressing this comment, we extended explanation of model processes concerning dynamic vegetation and added important equations to the appendix for the last revision.

- *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.*

  Thank you for the advice.
  Done.

- *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.*

  Addressing this comment, splitting Results and Discussion section was done already in the last revision.

- *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?*

  We chose ECLand and Noah-MP because both models can be and are widely used for coupling them as LSMs with established climate projection models. The fact that Noah-MP is only partly suitable to answer the research questions of this study was brought up more often by now. Thus, we decided to adapt the source code of Noah-MP a bit in a way that NEE and GPP are calculated also for the simulations with static vegetation. This is done by calling the function that usually processes the carbon dynamics but resetting all dynamically calculated variables afterwards to their values before calling that function. This gives estimated for GPP and NEE for the current atmospheric (photosynthetic) conditions of the current available biomass, without changing those. We checked whether our adaptation had an impact on the predictions of latent heat flux and could not find any.
  We explained that adaptation in the method section (L160-161), added some results for model performance of NEE and GPP (L295-308), included the LAI-NEE and LAI-GPP relationships for Noah-MP in Figure 7 showing the elasticity and changed the paragraph L434-448 in the discussion section a bit according to the new results.

- *Why did you initialize your model simulations differently (ECLand vs Noah-MP)?*

  In principle, both models are initialized with the same values, fitting as close as possible to the on-site conditions. However, there are some technical differences in the model initialization which we described.
  We added "The models were set up as closely as possible to the available site information but there are some technical differences in the structure of the model input, i.e. in the initial files" (L165-166).

- *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?*

For Noah-MP, I agree with you since there is only one vegetation type on the grid cell. For ECLand this requires adapting vegetation to be either high or low vegetation in the initial file. We did not do it originally, but changed it now. It did not affect the results much. Regarding the model performance of short vegetation types, we can interpret a bit more. One possible reason could be that forests have less dynamics in their productivity compared to crops, grasslands or shrubs. Surely, trees have dynamics in their leaf mass and photosynthesis rate dependent on environmental impacts but, commonly, have access to deeper water resources and intrinsic carbon storages to at least partly overcome water scarcity. Shorter vegetation types cannot cope for limitations in this way, resulting in higher relative temporal variations.
This explanation can be found in the discussion section (L443-448).

**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*

  "More detailed information" refers to the on-site LAI.
  The sentence changed to "We compared model results with observations across a range of climate and vegetation types from the FLUXNET2015 dataset and the MODIS leaf area product, and used on-site measured leaf area from an additional site" (L7-8).

- *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*

  We did not aim to pinpoint poor model performance of the models themselves for single or all selected sites. The question of this investigation was whether model performance can be improved by dynamic vegetation. Since this is not the case, we provide possible explanations and misrepresentation of the relationship between LAI and GPP is the major one we figured here.
  Reformulated into "We show that one potential reason for this could be that the implemented ecosystem processes diverge from the observations in their seasonal patterns and variability" (L13-14).

- *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*

  Merged into "Traditionally, their main purpose has been to provide a surface component in coupled atmosphere-land models. LSMs are applied in meteorological models, reanalysis products or in the Coupled Model Intercomparison Project (CMIP)" (L21-23).

- *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*

  Changed to "There is active development within the land surface modeling community, with more and more features being added to existing models to make them more realistic (Blyth et al., 2021)" (L25-26).

- *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?*

  No, we don't want to come up with new evaluation schemes. Rather, we want to motivate why we did an analysis with only a few models and presenting absolute performance metrics, which seems like "a step back" in comparison with multi-model evaluations.

- *L28 I suggest 'global, regional, and site scale'*

  Done (L28).

- *L29-31 I'm not sure I understand this. Do you mean that evaluation schemes are compared against each other? Or model - obs comparisons?*

  This was referring to inter-model comparisons.
  Changed "them" to "models" (L30).

- *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?*

  Best et al. (2015) compared the model performance of several LSMs and simple statistical models and ranked them based on normalized (relative) statistical metrics. The disadvantage of only presenting normalized metrics is that in any case there will be one model with the highest rank although it could be that this model misrepresents the target variable but the others are doing even worse.
  Changed to "Using this method, Best et al. (2015) reported that simple statistical methods achieve a higher performance in energy partitioning at eddy-covariance sites than any single LSM tested. One limitation of that study is that they did not report metrics of individual model performance, but only normalized ones. This procedure does not allow to judge whether the investigated methods have achieved a (dis-)satisfactory performance, since all methods might have a poor individual model performance" (L32-36).

- *L37 Replace 'had a closer look' with explored, investigated or something similar?*

Replaced by "…more closely explored…" (L38).

- *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?*

  Of course, there will be always uncertainty in measured data but I am sure that they accounted for that. Haughton et al. (2016) were investigating reasons for the outcomes of the PLUMBER study that simple empirical models outperformed most LSMs. They excluded systematic bias of flux tower data, time scaling effects and lack of energy conservation in the data as potential causes and stated that processes within or parameterization of the LSMs themselves need to cause poor performance.
  Reformulated to "…and not related to errors in the observations" (L41).

- *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)*

  What we were trying to say with that sentence was that benchmarking or ranking models alone is no suitable tool to identify specific causes for a mismatch between model predictions and observations. Achieving this, needs a deeper look into single models and their individual performance.
  Changed to "Yet, specific reasons for this mismatch, for example over-parameterization, missing processes, calibration issues etc., cannot be identified by benchmarking studies or model rankings alone, but requires further investigation of individual model performance" (L41-43).

- *L45 Could you name a few references to support your statement here?*

  This is an introducing topic sentence and several works are cited in the following sentences (L47-48).

- *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?)*

  Since one of the motivations to have dynamic vegetation in LSMs is to better predict impacts of water scarcity and drought events on the vegetation, we found it would be valid to argue that current implemented and used LSMs struggle in making prediction that fit observations in these conditions.
  We have shortened this paragraph a bit (L47-61).

- *L66-67 Why did you choose ECLand and Noah-MP? Other than them being used by others*

  Both models are still under development especially with respect to freshly introduced modules like that for vegetation dynamics (L65).

- *L74-75 I think you can drop this:)*

  Deleted.

- *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*

  Aridity describes water deficit in long-term climate conditions. Following this, it is the ratio of annual potential evapotranspiration to annual precipitation, leading to larger values of this ratio meaning larger aridity of the site. However, the ratio in this dataset was calculated the other way around which is less intuitive. Also, since we planned to filter the sites on a logarithmic scale, inverting delivered the opportunity to include more semi-arid and arid sites which differ much between each other with respect to seasonality and vegetation dynamics while humid sites are more even.
  We explained a bit more by adding "…and inverted afterwards, bringing it back to the initial definition as the ratio of the long-term mean annual potential evapotranspiration to the long-term mean annual precipitation by Budyko (1974)" (L78-80).

- *L83 Why +- 0.1 log AI? Is this a common threshold?*

  It is not a common threshold but we needed to come up with one within our filter algorithm. The aridity indices of wetter sites are closer to each other than for drier sites. In order to not overrepresent dry sites within selection by using a threshold in absolute values of the aridity index, we transformed the aridity index to a logarithmic scale, creating almost linearity of the aridity index scale.
  We explained a bit more: "Next, other sites with similar aridity (±0.1 logarithmic aridity index) were dropped to avoid an overrepresentation of some vegetation type-aridity combinations due to heterogeneous site distribution within FLUXNET. We used logarithmic values to create a linear scale of the aridity index, avoiding an overrepresentation of drier sites within the selection process" (L84-87).

- *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?*

  Haughton et al. (2018a) found out that, within the FLUXNET sites, drier sites (higher aridity index) and wetter sites with low temperature span tend to have higher predictability, meaning that it is easier to achieve good model performance. With our selection by aridity, we assured that we do not only include sites with high or low predictability.
  Since we don't want to focus too much on the predictability of the sites, but keeping it in mind, we changed the sentence to "We were left with 24 sites, covering a wide range of site characteristics as recommended by Haughton et al. (2018a) including aridity, vegetation types and observation periods (Fig. 1)" (L92-94).

- *Fig.1 Can you also explain the colorbar in the Figure caption?*

Added “The color scale represents the duration of the available time series in years” to the figure caption (Fig. 1).

- *L97 Why was precipitation set to zero in the gapfilling?*

  Filling missing precipitation data with zeros is the only option that is possible. We don’t know whether it rained that hour or day. However, the model input cannot handle missing values.

- *L97 Is using a Kalman filter a common procedure for gapfilling? How is gapfilling achieved in FLUXNET?*

  I do not know how common the Kalman filter is. Gapfilling for the TERENO site “Hohes Holz” was done with it. FLUXNET usually uses Marginal Distribution Sampling which is a really complicated algorithm to implement and to run. Additionally, it cannot fill large gaps as well, which can be seen in time series data from some of the FLUXNET sites (e.g. Fig. R1).

- *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*

  The ERA5 product we retrieved had 0.1° spatial and 1h temporal resolution and, thus, really helped with filling the gaps. The limit of 3h in using the Kalman filter evolved from the observation that the filter tends to overestimate the values when gaps are longer.
  Changed to “For longer gaps, the Kalman procedure tends to overestimate the observations which resulted in offsets at the end of the filling periods. Thus, filling data for these gaps was retrieved from the ERA5 (Hersbach et al., 2020) data product (Copernicus, 2018) with 0.1° spatial and 1 h temporal resolution” (L107-109).

- *L100-101 What’s the cut-off for ‘longer gap filled’ periods?*

  Longer periods where data is filled with Marginal Distribution Sampling (MDS) within the FLUXNET dataset can be seen visually because variability is unnaturally low (see Fig. R1). “Longer” in this respect means at least a month.
  Added “…we excluded gap filled periods that were longer than one month” (L112).

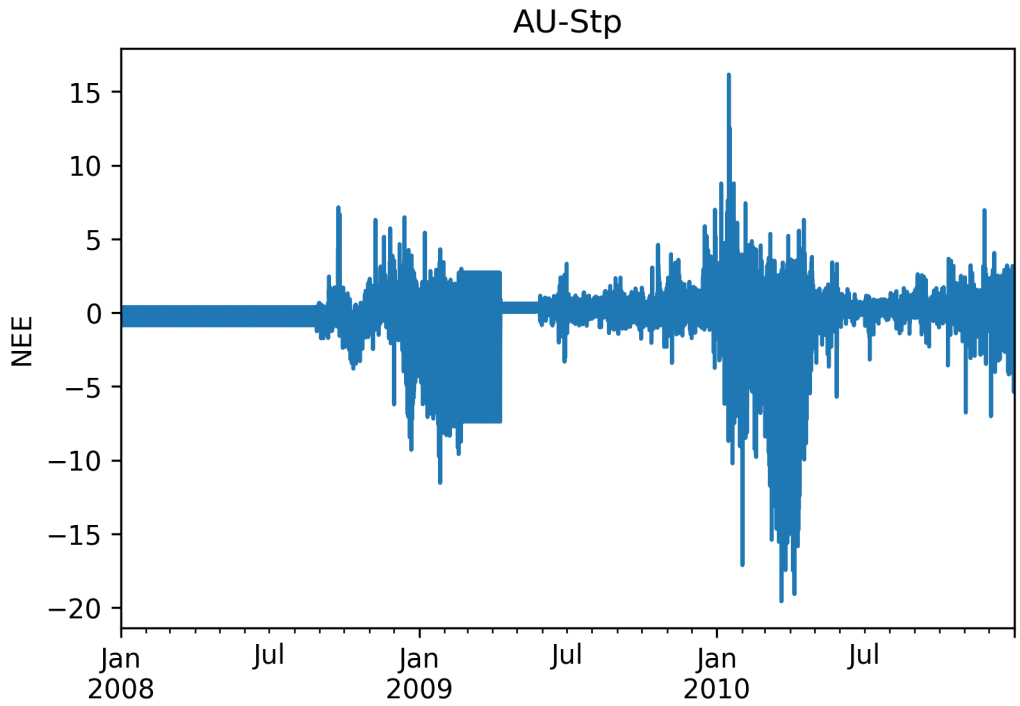


**Figure R1**: NEE time series for FLUXNET site AU-Stp, exemplarily. Gap-filling from MDS can be identified visually, in this case from January to August 2008 and from March to May 2009. These intervals were left out for model evaluation.

- *L103 You already defined LAI in L59*

  Deleted.

- *L104 What is the temporal and spatial resolution of MODIS LAI?*

  Temporal resolution is 8 days. There are different MODIS datasets available. The one we used, MOD15A2H, has a spatial resolution of 500 m.
  Added "One grid cell of 500 m x 500 m was selected per eddy covariance tower according to the site coordinates and LAI values with temporal resolution of eight days were extracted for the years 2000 to 2014" (L197-199).

- *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?*

  We explained usage and quality control of the MODIS LAI in more details now in L197-205.

- *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'?*

  Creating the LAI climatology means to calculate the average annual LAI cycle. For a 10-year time series of MODIS LAI, it might happen that some months have 30 values while other months have only 3 by selecting the same quality flags (i.e. 0 and 32) (see Fig. R2 for the site FI-Hyy). For example, a tropical site is covered by ITC cloudiness nearly at the same time of each year. Thus, all the values during that time have a lower quality flag

and would be excluded. It happened that we were left with some months without any LAI information, so we included a larger set of flagged data points for the climatology. As before, more explanation now can be found in L197-205.

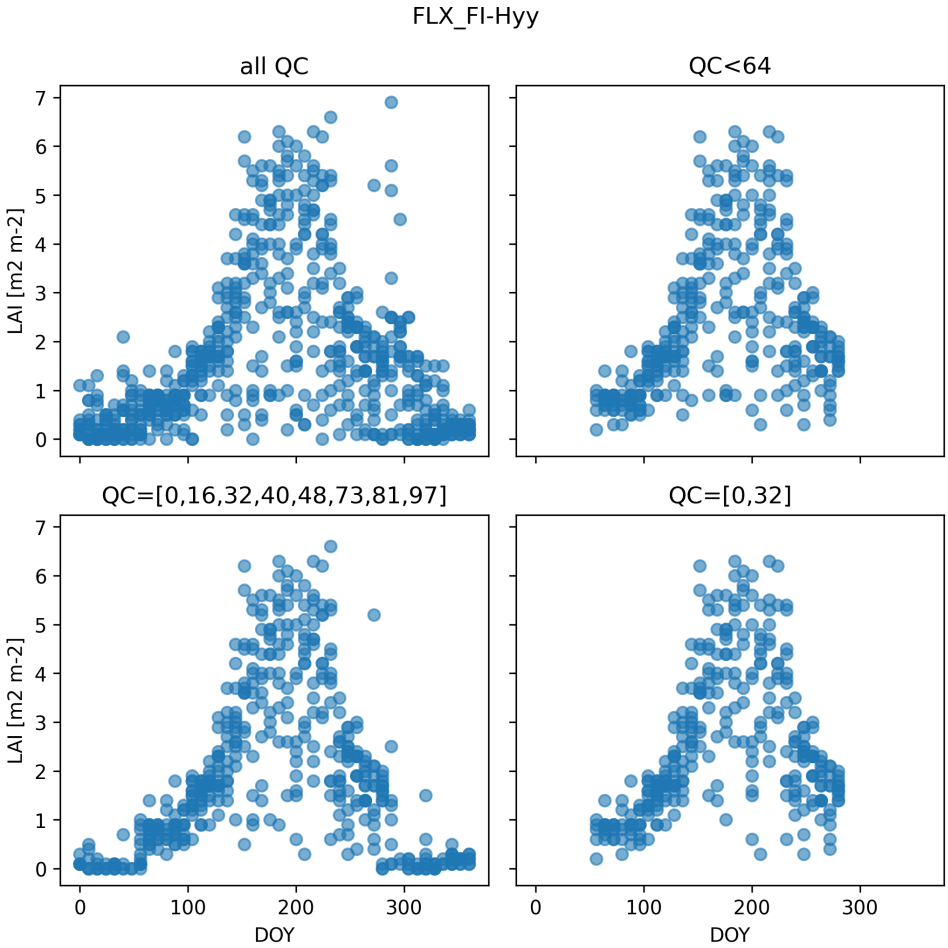


**Figure R2**: MODIS data points for FI-Hyy when using all quality flags (top left), quality flags less than 64 as recommended by Fang et al. (2012) (top right), specific quality flags in our selection (bottom left) and only "high quality" flags (bottom right). Using only the last category of flagged data (or even in this case all data points with QC<64) would have left us with no information on LAI during winter.

- *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?*

  See my explanation in point 2 of the general comments.

- *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?*

  Explanation extended and information on timespan added (now Table 3).

- *L127 'under-development' does this mean the model isn't 'ready' yet?*

  "Under-development" means that these models (and here especially the modules that incorporate dynamic vegetation to the models) are constantly extended and improved. We removed that word.

- *L129* (refers to L119) *Is the daily resolution really an appropriate reason to choose those quality flags? Why did you avoid smoothing your data for LAI?*

  From MODIS documentation, every value flagged higher than 0 has some uncertainty or limitation. QC=32 might have the least uncertainty after that. So, we limited the data used to these two flags for the single-day comparisons, to lower uncertainty in the data. Smoothing was not applied to capture also potential low or high peaks in the LAI data. Additionally, due to unequal gaps within the LAI time series of QC=0 and 32, the smoothing could distort the LAI values.
  Added in "Due to the usage of single day values, we solely used data with quality flags 0 (no issues) and 32 (saturated) to lower the uncertainty. Additionally, we refrained from smoothing to avoid an offset of the LAI values and left gaps as they were" (L216-218).

- *L133-134 So only two vegetation types per grid cell are simulated?*

  Yes, at maximum. It could be even one or none.

- *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*
  *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*

  We tried to leave model description as short as possible. However, more details on LAI-related processes might help and we included them.
  We extended the process explanations in L125-141 for ECLand and in L147-160 for Noah-MP and added important equations to the appendix.

- *L149 'taken from LUT' I would replace this with 'were prescribed' or similar*

  Changed to "…parameter values … of the 27 vegetation types are taken from look-up tables" (L147-148).

- *L152 'Thereby' I'm not sure this follows from the previous sentence*

  Changed to "Among others, …" (L151).

- *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)?*

Both models have two types of input: Initial files (with initial values for some variables to start with) for model setup and time series files with meteorological data for model runs. The initial files contain variables like vegetation type, deep soil temperature, soil layering, soil type, initial soil moisture, vegetation cover fraction and initial LAI value or LAI climatology which are not all present in the FLUXNET data. But the variables included in the initial files differ for both models that is the reason why it sounds like different setups but they are not. For ECLand, these initial data files were prepared for a global setup already and we could make use of that. For Noah-MP, no such setup existed and we created the initial files by ourselves by using the information we had. After model initialization followed the spin-up phase so that these initial values were not used any longer and became overwritten by actually modelled values.
Added the sentence "These initial files contain information on albedo, orography, soil type, surface roughness and monthly LAI which is not available in the FLUXNET metadata" to give more background on that (L171-172).

- *L166 Which site information - do they all report vegetation height and you used a certain threshold?*

  Clustering the vegetation into high or low vegetation type does not depend on vegetation height but on the vegetation type on-site. Forests in any case are high vegetation no matter how big the trees actually are.
  Added "Forests and savannas were treated as high vegetation types while grasslands and croplands were allocated to low vegetation types. The vegetation type that fits most to the FLUXNET characterization was selected (see Tab. 1)" (L174-176).

- *L169 Is there a reference for the van Genuchten soil hydrologic params?*

  Done (L178).

- *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'!*

  Done (L181).

- *L172 So the initial conditions for the two model runs are not identical? Why not?*

  The reason for the initial conditions of the two models being different is only because these initial files look different for both models and require slightly different set of variables. Apart from that, we kept initial conditions as close to each other and as close to on-site conditions as possible.
  Added "The models were set up as closely as possible to the available site information but there are some technical differences in the structure of the model input, i.e. in the initial files" (L164-165).

- *L173 I would have thought each gridcell has 8 neighboring gridcells?*

True. we checked whether soil type would change when including 8 neighboring cells compared to just 4 or even only the grid cell with the tower on it, but this was not the case.
Now, we took the soil type of the grid cell of the tower location and stated as follows: "… by selecting the grid cell including the flux tower location" (L181-182).

- *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*

  Done (now Table 1).

- *L183 What does 'transferred' mean here? Did you calculate daily averages/sums?*

  Yes, daily averages or sums (depend on variable).
  Changed to "…were averaged/summed to daily values for direct comparison" (L223).

- *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)*

  In principle, the relationship between observed and modelled values of a target variable is expected to be linear.

- *L189-190 Which symbol represents the normalized standard deviation metric?*

  Added $s_n$ (L229).

- *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 xmin differs strongly from $\bar{x}$ ?*

  A "normal" relative bias was not applicable since our target variables (i.e. LE, H, NEE and GPP) have values that vary around zero. This results in relative biases that are not only really large partly but also difficult to interpret (e.g. reaching 3000% of relative bias but not because the model estimate is far away from observation but rather because the mean value is close to zero). By subtracting the minimum, the distribution is shifted to positive values only, with the minimum value being zero. As a result, the relative bias really contains an information on how much the estimates deviate from the mean since the reference system is the codomain of the variable. This works independently of the distance between $x_{min}$ and $x_{mean}$.
  Changed to "For this purpose, the distribution of the observed values was shifted by their minimum, resulting in only positive values with a minimum of zero" (L232-233).

- *L199-200 So you try to understand the relationship between output variables within the same model simulation?*

  Yes, exactly.

Changed to “To investigate the sensitivity of dynamically modelled vegetation on the model performance, we checked how strongly the quality of the model simulation of one target variable (e.g. LE) depends on the model quality of another one (e.g LAI). For this, we used the elasticity as a metric” (L241-243).

- *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)?*

  Unfortunately, I found no publication from environmental sciences that use the same metric, only from economics.
  Changed to “Elasticity is calculated as ratio of the change in one statistical measure (analogous to equation 2) for two different target variables” (L243-244).

- *L213 ‘The model performance of the dynamic run is shown with the symbol’ - which symbol?*

  All symbols that are in the Taylor plots.
  Changed to “symbols” (L253).

- *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??*

  Shifted to discussion section 4.1 and extended to “For Noah-MP, model quality metrics were in the range of other studies (Brunsell et al., 2020; Li et al., 2022; Xu et al., 2021; Liang et al., 2020). However, dynamic LAI modelled by Noah-MP in our assessment with a mean of +70 % was more biased compared to a mean of +20 % for annual LAI values reported by Li et al. (2022)” (L392-395).

- *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??*

  Here, we refer only to literature because Stevens et al. (2020) also replaced LUT LAI by MODIS LAI and compared model results. However, as stated in L194-196 ECLand LUT LAI is already MODIS-based.

- *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?)*

  For the dynamic simulations, LAI is not prescribed but still part of the initial files. It is expectable that LAI predictions for the dynamic simulations are independent of the initial input. However, dynamic ECLand still incorporates prescribed LAI to 5% (RLAIINT=0.95 was defined by the developers’ team to be fully dynamic).

- *L225 'but not necessarily for all sites' can you quantify this more?*

  Changed the sentence into "For ECLand, this was also the case for many sites but not necessarily for all, e.g. AT-Neu and AU-How" by adding examples (L259-260).

- *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*

  Done.

- *L226 'increased variance' compared to what?*

  Increased variance in comparison with static ECLand simulations.
  Added "…compared to static simulations…" (L260-261).

- *L227 'was random' - what does this mean?*

  We could not find any tendencies regarding aridity or vegetation type to have positive or negative shift in relative bias.
  Replaced by "was ambiguous" (L263).

- *L231 What are sparse vegetation types (i.e. did you define it anywhere)?*

  No, we did not. Sparse vegetation are savannas and shrublands because they have no closed canopy surface.
  Added "Especially short (GRA+CRO) or sparse (SAV+WSA) vegetation types…" (L265).

- *L243-244 Can you also give a value range for ECLand - it'd help the comparison between the models*

  Added "…−30% on average…" (L280-281).

- *L246-252 Here you kind of say a bit about where original LUT is coming from, and this belongs in the methods section already*

  This is now discussed in L407-412.

- *L255 Why are you not showing the single-year LAI results - could just be in the appendix?*

  Relative bias for the MODIS single-year simulations are in the appendix tables. But they are not part of the Taylor plots.

- *Fig 2 How did you set threshold to separate the AI into different aridity classes? This belongs in the methods*

  We included the thresholds in the figure caption now as "The symbol colors indicate the site aridity (top right legend) as following: very humid - aridity index (AI) < 0.6, humid -

AI < 1.25, sub-humid - AI < 1.54, dry sub-humid - AI < 2, semi-arid - AI < 5, arid - AI ≥ 5 (Ashaolu and Iroye, 2018)" (Fig. 2).

- *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*

  Static Noah-MP produces no output for NEE and GPP which is according to model structure. Thus, only the values for the dynamic runs can be presented here.
  Colors changed, axes extended (Fig. 3).

- *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?*

  This should work now since model description was extended, also by including process equations.

- *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*

  Please see my answer to the seventh point of the general comments.

- *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*

  For most of the sites, GPP was overestimated with dynamic Noah-MP, but relative bias was predominantly small for forests (Tab. A3).
  We reformulated that conclusion and tried to be more precise: "Thus, although some previous studies found substantial uncertainties in modeled GPP for different vegetation types (Ma et al., 2017; Liang et al., 2020; Li et al., 2022), predicting ecosystem variables using dynamic Noah-MP could be useful at least for forests in studies when LAI climatology cannot be used such as climate change impact studies" (L449-452)

- *L289 'being independent of the prescribed LAI forcing' - so this is just the dynamic simulation that does not have any prescribed LAI?*

  In principle, the dynamic simulation has the prescribed LAI (since it is part of the setup file) but it is not using it and, thus, runs fully dynamically.

- *L289-290 Not sure consistent is the right word here, those are two v different experiments*

Changed to “In line with the finding that model performances of dynamic Noah-MP were independent…” (L315).

- *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)*

  On-site LAI and MODIS LAI were linearly correlated. MODIS LAI might be biased for some sites, but so might be on-site measured LAI due to technical limitations (scatter correction, saturation effect…). During development of the dynamic vegetation modules, a tuning of the parameter sets was done but not to MODIS LAI as target variable. However, mismatch between MODIS and on-site LAI is reflected in lower performance of NEE and GPP of the static ECLand simulations. The reason is unclear: It could be that on-site LAI does not reflect actual LAI but it could also be that calculations of GPP in relation to LAI do not match reality (similar to what we have shown in Fig. 8). For the dynamic ECLand runs, differences between MODIS and on-site LAI play only a minimal role since 95% of the LAI calculations come from dynamically predicted LAI and NEE and GPP predictions are even fully dynamically predicted.

- *L306-307 Where is this recommendation coming from? Did you not say earlier there is no performance difference depending on the LAI forcing used?*

  The performance is not different for the dynamic simulations. But for the static runs, it is. Thus, we recommended here to use static simulations with MODIS climatology forcing.
  Shifted to section “4.4 Implications” (L580-581).

- *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*

  Yes, I agree, definitively would be interesting to look at but, as you said, was beyond the scope of this investigation.

- *L329-330 Interesting indeed - do you have any idea why this is happening?*

  It might be that carbon and water transport processes are coupled not tightly enough. With NEE estimates fitting well, the photosynthetic activity also is good captured by the model. The demand of water by the photosynthesis might be underestimated by the model and, leading to less transpiration and, thus, also to a lower fraction of energy that is used for latent heat transport. Additionally, downward CO2 transport and upward water transport through turbulent fluxes occurs in the same eddies which is not captured by the model. These are just some ideas on that so far. We refrained from intensifying the discussion on that.

- *L335 Again - why not include a figure in the appendix to show these results?*

  Done (Fig. A2).

- *L351 How are vegetation and soil moisture state variables coupled?*

  Vegetation needs water for photosynthesis which stems from the soil. Thus, more photosynthetically active biomass extracts more water from the soil and, otherwise, less soil water restricts maximum plant productivity and biomass build-up.
  This is further discussed in L495-503.

- *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*

  Yes, you are right. The reason for unaffected soil moisture to vegetation dynamics still remains unclear. Referring to the point before, it could be due to the implemented interaction of carbon and water processes. First, the potential photosynthetic activity in dependence of leaf area and radiative conditions is calculated. Then, the limitation factor of extractable water is estimated according to available soil water and roots. Lastly, the photosynthetic activity is adapted to that restriction and transpiration rate adapted to conductivity and atmospheric conditions. As a result, the only included path is that soil moisture impacts photosynthetic activity and biomass build-up. But there is no feedback that more biomass needs/loses more water that will be taken from the soil because photosynthetic activity relates only to the carbon fluxes but not to the water fluxes.
  We added this explanation to the text (L495-503).

- *Fig 7 Where is the footnote for elasticity?*

  Sorry, that footnote was there by accident.
  Removed.

- *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)?*

  Other studies also found a linear relationship between LAI and GPP but with large variability. Some sites might be exceptions from the linearity (IT-Ren) where LAI-GPP relationship appears to be a non-linear saturation function.
  We now added LAI-GPP and LAI-NEE elasticity for ECLand in Figure 7 but excluded NEE-LE and NEE-SM instead (Fig. 7).

- *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*

Since the most probable in the observations LAI-GPP relation is a linear one, Pearson correlation coefficient is the statistical basis of this linear regression and also the measure for the relationships from the model output. We cannot compare different kinds of correlation coefficients.
We added description of the arrows to the figure caption with "The arrows represent the range of GPP and LAI values for the individual seasons" (Fig. 8).

- *L403 You need to define σr*

  Done (L371).

- 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)

  This is now part of the section 4.4 "Limitations" (L592-597).

- *L454-455 Can you give a reference for this ('real LAI'), also maybe replace real with 'observed' or similar*

  We cannot replace "real" by "observed" because we are not referring to any measured values here. The reality this sentence is referring to is the fact that trees do not immediately lose their leaves when they are faced to a few days of suboptimal conditions for photosynthesis.
  Replaced by "realistic" (L571).

- *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)*

  Not in the manuscript anymore.

- *L467 Why did you expect that dynamic vegetation would improve ecosystem exchange especially for short vegetation?*

  Compared to forests that are more resistant and resilient for e.g. water scarcity, short vegetation more dynamically and more instantly responds to environmental limitations for its growth. Thus, firstly, assuming the same LAI cycle for each year and, secondly, assuming a constant LAI values over a whole month as in the static model simulations, do not represent reality. Our expectation was that modelling vegetation dynamically would cope for that variability and, as a result, yield in better performance of observed ecosystem fluxes.

- *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)?*

  Other models have processes implemented differently. So, there is no chance in directly transferring results and conclusions from these two models to others.
  Added "…since they have processes implemented differently" (L624).

Literature


Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, Journal of Hydrometeorology, 16, 1425–1442, https://doi.org/10.1175/jhm-d-14-0158.1, 2015

Boussetta, S., Balsamo, G., Beljaars, A., Kral, T. & Jarlan, L.: Impact of a satellite-derived leaf area index monthly climatology in a global numerical weather prediction model, International Journal of Remote Sensing, 34:9-10, 3520-3542, DOI: 10.1080/01431161.2012.716543, 2013.

Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Santanello, J. A., J., Stevens, L. E., and Vuichard, N.: The plumbing of land surface models: is poor performance a result of methodology or data quality?, J Hydrometeorol, 17, 1705–1723, https://doi.org/10.1175/JHM-D-15-0171.1, 2016.

Haughton, N., Abramowitz, G., De Kauwe, M. G., and Pitman, A. J.: Does predictability of fluxes vary between FLUXNET sites?, Biogeosciences, 15, 4495–4513, https://doi.org/10.5194/bg-15-4495-2018, 2018a.

Stevens, D., Miranda, P. M. A., Orth, R., Boussetta, S., Balsamo, G., and Dutra, E.: Sensitivity of Surface Fluxes in the ECMWF Land Surface Model to the Remotely Sensed Leaf Area Index and Root Distribution: Evaluation with Tower Flux Data, Atmosphere, 11, https://doi.org/10.3390/atmos11121362, 2020.