Response on Review 1

First of all, we want to thank reviewer 1 for reviewing this work of me and my co-authors. Thank you for putting a great deal of effort and discovering many details. Your feedback has greatly helped improving this work and publication. In the following, I will go through and respond to your comments.

Specific comments: (reviewer comments in italic, responses in normal font, changes in blue)

1. A key finding in this paper is that activation of vegetation dynamics in Noah-MP and ECLand does not improve model performance across several variables and metrics. Such a result appears at odds with the necessary testing that accompanies model development – new features are rarely implemented if performance is decreased. As such, I would like to see further exploration of the discrepancies between the findings from the model development team and this paper. For instance, the development of ECLand (and CHTESSEL) has papers by Miguel Nogueira, in addition to those from Souhail Boussetta, that explore model performance. It is important to synthesise such publications in the manuscript and investigate reasons for any divergence of results. As an example, if model iterations are tested against variables such as land surface temperature over water and carbon fluxes during development then perhaps the authors' results indicate a need for a broader testing process.

Indeed, model development needs testing and novel model modules are only incorporated when there is an improvement or at least no deterioration in model performance in the variables that were chosen for evaluation. The choice of variables for model evaluation is really important. Since ECLand mostly is used in Climate Projections, the most important variable is land surface temperature which then is used for model evaluation. However, changes in vegetation representation barely affect energy balance calculations, especially not on a coarser temporal resolution that often is used for model evaluation. As a result, it is hardly surprising that model performance in our investigation diverges from published results during model testing. Thus, yes, we wanted to indicate that model testing needs a broader spectrum of target variables and different temporal resolutions. The work of Nogueira et al. (2020) is interesting but they focused more on updating land cover fractions and vegetation type clumping which had an important effect on land surface temperature. We extended the discussion on model performance and the importance of vegetation-related variables in ECLand.

2. The methodology uses LAI taken from MODIS as both an input and the observation against which the model performance is analysed. Such analysis is circular and, although it could prove useful in determining how model outputs change based on inputs, should be discussed further within the paper. It could also be avoided by using independent datasets – were efforts extended to identify alternative sources of LAI data, beyond the on-site data used for DE-HoH? There are other remote sensing products and certainly other flux sites with on-site LAI measurements.

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 evolution. 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. However, we could show that even with the same source of the data the dynamic simulations are not fitting the observations.
In the revision, we now provide more details on the MODIS LAI data and highlighted the differences between data used for input and for evaluation. (Lines 217-220)
Yes, the evaluation would really benefit from using on-site LAI data from more than one site. We were very thankful for having an additional LAI data source at all. I (first author) tried reaching out to the FLUXNET community via their contact form several times but never had any responses.

3. The authors investigate the behaviour of two models, ECLand and Noah-MP. However, the analysis is limited in places due to the static implementation of Noah-MP not producing output for carbon fluxes. With a large array of LSMs to choose from, further justification for selecting a model that can only partially contribute to the manuscript is required.

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. Although Noah-MP provides no GPP and NEE output for the static runs, it still is interesting to look at the LAI-GPP relationship within the model that we did for Figure 8. Nonetheless, we need to be more careful with absolute statements that we did. We adjusted the abstract and the discussion according to that.

4. The site selection for the study was performed such that sites with the same IGBP PFT class that fell within the same aridity bracket as already selected sites were dropped. This takes place to "avoid including more than one representative site for each combination of aridity and vegetation type". This data selection criteria, in addition to the sensible choice of excluding sites with less than 5 years of data, results in 22 sites being used out of a potential of 212 FLUXNET2015 sites available (of which 120 have 5 years or more of data). More information on this is required. What were the possible consequences of having two or more sites in the same PFT and aridity classes? One would assume that including more sites could provide additional insight into the reasons for model performance – namely helping provide strength to any statements made around the role of PFT and aridity interaction. In addition, the study is then such that 5 grassland sites are included but only one savannah site and one mixed forest site. How might this unbalanced dataset affect the interpretation of results?

Representative site selection is an important issue and we gave it detailed consideration when designing the analysis. When looking at the global distribution of FLUXNET sites, many of them are located in temperate climate on the Northern Hemisphere. Including all sites with more than 5 years would create an overrepresentation of regions with high density in sites, resulting in an imbalance of PFT-aridity combinations for model evaluation with especially (semi-)arid short vegetation being underrepresented (which is one of the limitations Martens et al. (2020) and Nogueira et al. (2021) faced in their study). Thus, we needed some sort of filter algorithm to avoid that overall model performance is either shifted towards better or worse performance due to this imbalance.

Savannah types are indeed separated within IGBP PFT, but this is not done in the models. Accordingly, I did not separate them either when selecting the sites, meaning that SAV and WSA belong to the same group within this selection process. I also merged PFT type MF with DBF since, after the selection via the aridity index, only two MF remained which is critically few (this is mentioned in the manuscript in line 91). Other possible sites had to be removed due to low-quality in soil moisture data (mentioned in line 89). Unfortunately, there are not enough sites available to create a second set of the same structure, as some aridity-PFT combinations are really rare. We are aware that such a second set would be helpful for strengthening and reproducing our findings. We now explained in more detail why and how site selection was done, and adapted Figure 1 in accordance with the model PFTs.

5. Unfortunately, the manuscript would benefit from an additional thorough proofread. I have included many small issues in the Technical Corrections below, but there are sure to be some I have missed. Some errors are particularly important to address for a scientific publication – for instance, there is a reference to a dataset having an author of "The PLOS ONE Staff" as the bibliography reference is to a correction of the original dataset article. Aside from this, the Results and Discussion section is hard to follow in places and frequently jumps from discussing one topic to another before returning to the original topic. It is not always clear which figure or model run is being discussed. Figure 8 is introduced multiple times, and some statements are duplicated. I would suggest the authors restructure the manuscript to have separate sections for the results and the discussion as this should provide additional structure and clarity. This should be combined with the addition of more quantitative results, providing numbers to support statements, and further explanation of how results are reached. For example, on line 223, "simulation results were unaffected by the type of LAI forcing with vegetation dynamics switched on" could be explained further by clarifying that this is seen from the symbols being in similar locations in Figures 2c and 2d. Supporting this statement with the mean difference in model performance across the sites would also help satisfy the need for more quantitative results.

We took care for mistakes in citations and linguistic deficits and separated Results and Discussion section. We thoroughly proof-read the manuscript.

Technical Corrections

Thank you for careful reading and writing down the propositions with this detail! I will only respond to those that exceed language.

4. Line 8: "and use more … "Hohes Holz"." is overly detailed for an abstract.

True. Made it more general.

5. Line 9: "current implementation" – the current implementation of what exactly?

L9 "Current implementation" meaning the model source code as is it published currently.

Added "of dynamic vegetation" to be more clear.

11. Line 30 and throughout the manuscript: "like Best et al. (2015) or Krinner et al. (2018)." I would suggest that citations are included in the usual manner with the conjunctions *implied. Line 30 would hence become "Such works that introduce individual evaluation schemes are often accompanied by studies that perform comparisons between them (Best et al., 2015; Krinner et al., 2018)."*

done

12. Line 33: "than an ensemble of LSMs". With 'ensemble' onen having a specific meaning within the LSM community, I would suggest changing this line to read "than any single LSM", or similar, to more accurately reflect the findings of Best et al.

done

13. Line 34: "This does not allow to judge whether the investigated method achieved a (dis-)satisfactory performance". The authors of benchmarking studies would likely disagree with this statement, with one purpose of benchmarking being to assess whether performance is satisfactory against various a-priori expectations.

Here we are referring to benchmarking studies that use relative metrics to create a rank order of the models, like the ones of PLUMBER. Those do not provide information on whether the best model in this ranking really achieves good fit with observations since the absolute metrics are not shown. We have stated which types of studies we mean specifically in the introduction.

17. Line 47 and throughout the manuscript: FLUXNET Multi-Tree Ensembles are one type of product, considered "a precursor to FLUXCOM" as stated in the Jung et al. paper cited here. I would suggest referring simply to FLUXCOM, a well-known product in the community.

Done

20. Line 52: How do LSMs misrepresent water-sensitive regions? This statement should be expounded upon.

This paragraph was changed completely.

21. Line 57: "Currently, most LSMs are not able to represent a direct vegetation control on surface exchange". Do vegetation parameters not influence transpiration within LSMs and therefore exert a control on the land-atmosphere exchange of water?

We added more explanation.

2ti. Line 62: "… by LSMs helps to shed light on the known discrepancies". "helps" should probably be "would help". Which discrepancies are being referred to here?

We added more explanation.

28. Line 66i: "… that can only be executed for a limited set of models". This needs clarification – is this due to time constraints within the study, or is there some other characteristic of certain models that exclude them from such studies?

We added further explanation.

31. Line 71: What is meant by “different patterns” and “possible misrepresentations of the observations”? This could likely be stated more clearly.

We changed to “What are the mechanics behind modeled temporal patterns in vegetation dynamics and occurring misfits to the observations?”

32. Line 74: This statement is superfluous and could either be moved or removed.

Was removed

34. Line 80: The data from Trabucco and Zomer (2018) should be referred to as the CGIAR-CSI Global-Aridity and Global-PET Database.

done

35. Line 87: “… we assumed [the sites] to be neither very predictable nor very unpredictable in total …”. This statement is unclear in both its meaning and implications.

Was removed

36. Figure 1: Reference the IGBP classification scheme used, as well as where this data was obtained for each site (e.g., from the FLUXNET website, or within the site netCDFs). I believe some of the sites in this study have 19 years of data available – why does the scale have an upper limit of 18? Furthermore, a continuous color scale could likely be used here to allow differentiation between adjacent numbers (and clarify which color each label belongs to).

Fig. 1 reference was added. The longest time series within the selected sites was 199ti-2014 which is 18 years. We refrained from using a continuous color scale because the observational time can only be full years and, thus, the sites have distinct duration classes. We adapted the resolution of the scale.

38. Line 95: Should the soil water content have an abbreviation introduced here in the same vein as the fluxes?

We wanted to use as less abbreviations as possible to assure readability especially in Results and Discussions section. Additionally, the part where soil water content is referred to is limited which made it unnecessary to use an abbreviation.

40. Line 99: Rather than “We adopted the same procedure …”, simply say that you also excluded timesteps where L <= 0.

done

41. Line 101: How long did the gap-filled periods need to be to be excluded from the model performance analysis? How might this affect the results from the analysis?

We excluded gap-filled periods that were longer than one month from model evaluation.

42. Line 105: Why were only these four quality flags allowed? Other studies indicate any value less than 64 is usable (e.g., Fang et al., 2012; Ma and Liang, 2022).

More information on quality flags of MODIS was added. Working with MODIS is challenging and you would need to select quality flags for each site separately. After revisiting MODIS flags as part of the revision, I decided to exclude flag 65 but pro 73, 81 and 97 especially in order to keep some values during winter (see Fig. R1). Thus, we also redid the model runs, affected by those adaptations.

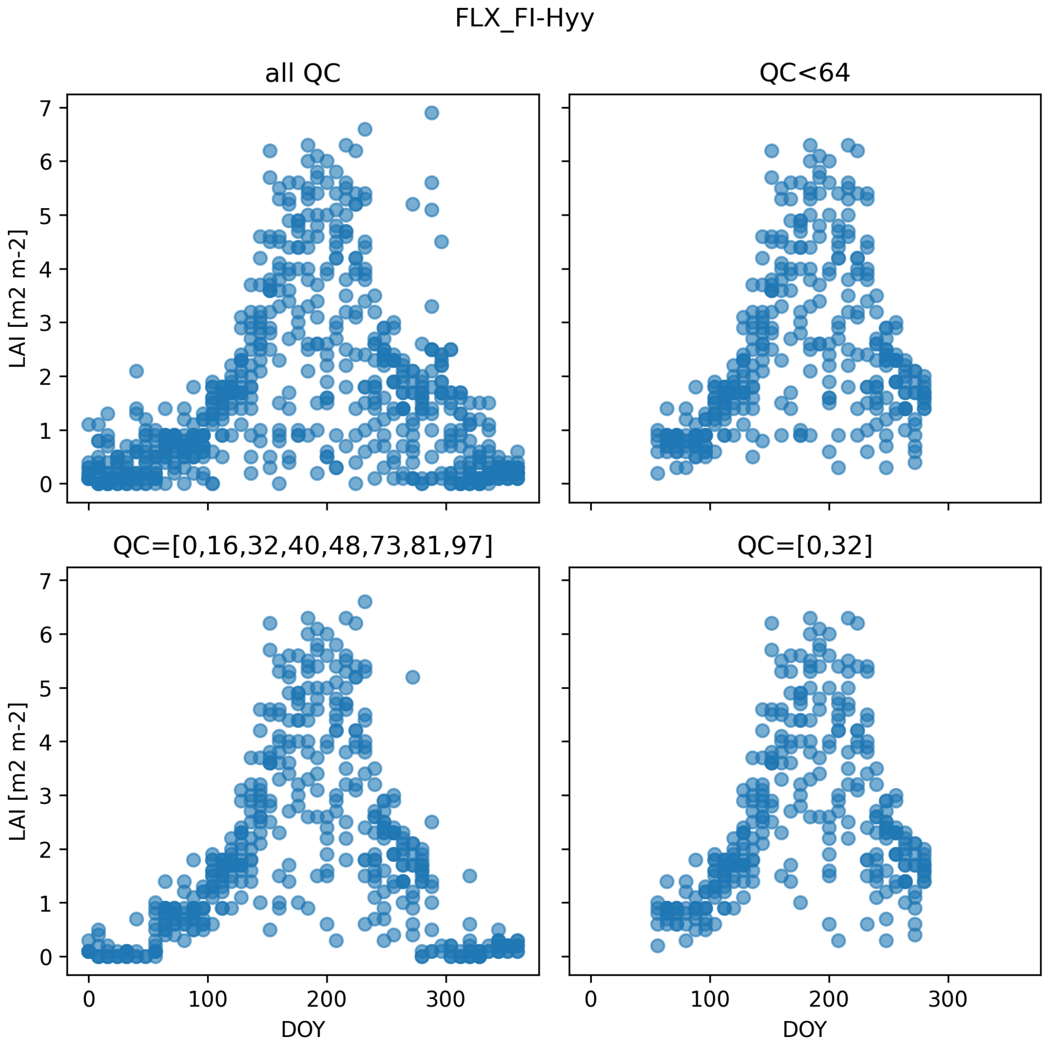

Figure R1: MODIS data points for FI-Hyy when using all quality flags (top left), quality flags less than ti4 as recommended by Fang et al. (2012) (top right), specific quality flags in our selection (bottom left) and only “high quality” flags (bottom right).

44. Line 107: Cite papers to provide assurance that using a Savgol filter is suitable for this purpose (e.g., Cao et al., 2018; Chen et al., 2004; Huang et al., 2021).

Done

45. Line 110: “Each following year … that specific year.” is ambiguous in terms of which year is being used as the forcing.

Removed

46. Line 112: “If LAI values for more than one month were not available” – did these months need to be consecutive?

Yes, this refers to consecutive months. Information added.

48. Table 1 and throughout the manuscript: Is there any potential for providing shorter labels for the “Terms” from this table? While they are descriptive which is useful, they can be unwieldy in length.

This is true. But descriptive simulation settings also prevent confusion, so we prefer to keep these.

49. Line 117: “Due to the use … from smoothing. Gaps were left as they were”. This needs more explanation. Why are the same data from earlier (with QC flags of 48 and 65) not used here?

Good point. I needed to keep data with other QC for creating the climatology to ensure that each month got a value for that site, which was a challenge especially for the single-year simulations. But since the trustability of these data points is low, they were left out for the temporal higher resolved evaluation.

53. Line 131: Explain what “IFS cycle “CY46R1” means.

Changed “cycle” into “version” to make clear that this is the version name/number.

54. Line 131: How were the IGBP PFT classes from FLUXNET2015 mapped onto the 19 vegetation types within ECLand? If the two classification schemes do not exactly match (or say, the ECLand types are taken from default data based on lat/lon of the site), were any tests performed to confirm that the classes aligned in a suitable manner?

We initialized the model with the closest possible fit to the on-site conditions without changing any parameters. For ECLand, we had a global setup that we used based on ERA5. We did not adapt the parameters in the global setup. Additional tests were not conducted. We incorporated the vegetation classes for ECLand into Table 2 to enhance clarity. However, it is true that tile fractioning in ECLand into high and low vegetation in the default setup might bias the evaluation with point measurements that belong to only one of these vegetation types. So, I checked and found that the vegetation type in the initial files from the global setup did not match the FLUXNET classification for some sites. Thus, we adapted that and repeated the experiment. Substantial changes of the resulting model performance occurred for some sites (e.g. AT-Neu, BE-Lon) but the general outcome of the study was not affected.

57. Line 135: Does “respective cover” refer to the fractional cover of each of the two vegetation heights?

Yes, “respective cover” means the fraction of each vegetation type on the grid cell. Wording was changed to “fractional cover”.

59. Line 149: See the comment for line 131 regarding the PFT classes for ECLand. The same holds here for Noah-MP.

Same as above in comment 54, here as well, we assured model setup to fit as closely as possible the on-site conditions.

62. Line 152: “Stomatal resistance is controlled by photosynthesis”. Is this statement true for Noah-MP? I would think it is more of a coupled relationship where stomatal resistance can also be controlled by e.g. vapour pressure deficit which in turn would decrease the level of photosynthesis by limiting the available intercellular CO2.

Yes, as it is explained in Niu et al. (2011) section 4.2. Changed to “Among others, stomatal resistance is predominantly controlled by photosynthesis (Niu et al., 2011)…”.

63. Line 160: While the height of flux tower would ideally be dependent on the vegetation height, this isn’t always true – towers can be situated within the canopy or many meters above it.

To be honest, the information that the tower ends in the vegetation canopy, is new to me especially since the aim of the network is to capture fluxes of the respective vegetation type. I checked the given measurement heights of the sites I chose, and two of them might look suspicious but I don’t know the vegetation on-site.

64. Line 161: How deep was the uppermost soil layer?

The uppermost soil layer for Noah-MP is 0.1 m and for ECLand 0.07m. We added this information.

65. Line 162: Was the ten-year spin up sufficient to reach a steady state in each model? What variables were used to check that such a steady state had been reached?

Steady state was not checked quantitatively but qualitatively.

66. Line 164: What “initial data” was taken from ERA-5? Why was the FLUXNET2015 data not suitable?

The initial files contain information on soil, tile fractioning, LAI climatology, state variables at the time of the start of the simulation. For the latter, I could have replaced them by measured values from Fluxnet2015 but the values adapt during the spin-up anyways.

67. Line 172: Why are the soil data averaged from neighbouring cells?

This was initially done to have better representation of the general conditions surrounding the tower. We checked and it would have been not necessary since soil type within the neighboring cells was the same as for the grid cell of interest, so we removed this averaging process in the revision and only work the grid cell where the tower is located.

69. Line 177: What aspect of the model meant that the temperate vegetation did not regrow? Is the requirement to have green vegetation fraction set to 1 a detriment to the results or their interpretability?

This is about the minimum green vegetation fraction. The formulation was misleading in the manuscript. Setting the minimum green vegetation fraction to 1% assures that there is still a small amount of biomass after the winter, which is essential for the model to generate spring growth. Without any biomass (i.e. leaves) there would be no location for photosynthesis to take place (zero leaf area * high potential photosynthesis still is zero). Changed to “Minimum green vegetation fraction was set to 1 % to ensure that not the whole vegetation cover dies during winter which would hinder temperate short vegetation from growing in spring.”

71. Line 180: Is the implicit temperature time scheme for the surface temperature?

Yes, vegetation canopy surface temperature is meant and added.

72. Table 2: Why is the IGBP class OSH in this table when it does not feature in Figure 1? How was the initial LAI changed for sites in the Southern Hemisphere, namely the Australian sites? With DBF LAI set to 0.0, it seems clear that the Noah-MP initial LAI is based on a year startng on 1 January in the Northern Hemisphere, yet the Australian sites are potentially at the peak of their growing season in January.

72. Tab. 2 removed OSH from the table.

75. Equation 1: I do not think this is needed as Pearson’s correlation coefficient is widely used and available in many programming languages.

True, was removed

78. Equation 2: This is justified as avoiding division by 0 or values very close to zero. However, this doesn’t strictly follow from the formulation of the divisor. If the observations have very low variance or are very biased towards 0 values, then conceivably the mean minus the minimum could still be a very small number.

Agreed. But division by 0 is successfully avoided and in case of very low variance, the numerator is also small, which results in reasonable values of the relative bias and not like 3000%.

82*. Line 204: Was an abbreviation considered for the normalized standard deviation to improve the ease of referring to it, and bring it in-line with the other two metrics which are referenced with a single letter?*

Done

88. Line 220: “a bunch of” should be avoided – what was the actual number of sites?

This part was reformulated

89. Line 225: This is confusing wording as it is difficult to determine whether the authors are referring to all the sites, one specific site, or just some sites.

This part was reformulated

90. Line 228: “whether the predicted LAI fit better … was random”. Was the difference in performance random with respect to the sites’ classes or aridity? It might be better to say that there was no clear rela6onship between the difference in performance and the site characteris6cs explored.

Agreed, we changed “random” to “ambiguous”

91. Line 231 and throughout the manuscript: Define which classes are meant by “short or sparse vegetation types”.

Good point, thanks. We added in brackets which vegetation types we refer to.

93. Line 234: Comparing the static simulations across Figure 2 (and the other Taylor diagrams) is difficult as the end of the arrows are hard to locate, especially with respect to the site that the arrows represent when the arrows are clustered.

I changed the symbols to be a bit smaller so that they have less overlap, in hope that helps. But, we refrained from using thicker arrows because that could also be counter-productive by blocking the symbols.

94. Line 236: “With activated vegetation dynamics … in the Taylor diagram”. This statement implies that performance improves for all sites and all LAI forcings, yet

This statement refers to using MODIS climatology as LAI forcing. We added this only by referring to the respective figure parts because otherwise it would be too repetitive in this paragraph where we are already talking about MODIS climatology.

97. Line 242: Figures 3d-f are referenced but Figure 3 does not have sub-labels.

Sublabels added.

99. *Line 249: The total LAI is disaggregated into high and low, yet the model is run with either only high or low vegetation. How does this impact results, as one can imagine this results in lower LAI than truth.*

One grid cell in ECLand is split into high and low vegetation fraction with their LAI values. Meaning, one spot cannot have both vegetation types (there is no layering). The resulting LAI is then the weighted mean according the high and low vegetation fraction. Thus, if a grid cell in our setup is only a high vegetation type, resulting LAI is higher than for a grid cell that has also a low vegetation type fraction. This is the closest we can get to the footprint of flux tower observations.

101. Figure 2: Why does the arrow of US-Var extend outside of the plot domain in Figure 3c? How does US-GLE in Figure 3c have no change in either standard deviation or correlation yet an extremely large change in the relative bias? This would imply a simple shin in magnitude in the LAI output which would be striking if caused by the switch to dynamic vegetation.

Correlation coefficient for static Noah-MP LAI for US-Var was negative, so the arrow starts there. US-GLE has no change in standard deviation or correlation is because, as an evergreen forest, default LAI is constant throughout the year and, thus, correlation coefficient cannot be calculated.

102. Figure 2 and others: How were the aridity brackets defined for the color coding?

Fig. 2 Quantitative limits of the aridity classes based on Ashaolu & Ilorin (2018). Added that information and citation to the caption of Figure 2.

103. Figure 3: Since other figures are in color, I would suggest this figure also use color to differentiate between static and dynamic to help visually distinguish between the two.

done

104. Line 270 and throughout the manuscript: More consistency in the used definition of "model performance" would be good and can be aided by being more explicit about the metric currently being discussed.

The term model performance aims to include all the metrics that are discussed here. "Lower model performance" in general means that the majority of the metrics show deterioration. In other cases, the explicit metric is referred to.

105. Line 279: More explanation of how the opposing NEE biases indicate differences in respiration estimates is required.

We added more information about the model structure to the Methods section

107. Figure 4: Why is AU-Stp outside of the plot area for Figure 4a? The axes should be extended so that the site falls within the plot area.

done

110. Line 304: "Findings from this study … modelling carbon and energy fluxes". This is a strong statement about the impact of this work and requires more discussion to support *it. Which processes within ECLand has this study iden6fied as requiring further development? How has the study provided evidence for how these processes should be improved within the model?*

More insights and evidence for this are given in section 3.3. This statement is now in section "Implications".

112. Line 309: "Stevens … with static ECLand" is not needed as the precise results from these other studies are not critical to the discussion. Instead, these two papers could simply be cited to support the prior statement that the results are comparable to other studies. If the exact values from the prior studies are mentioned, then it would be good to also state the same metric values from this study explicitly.

Unfortunately, using the same metrics from other studies is not possible since they basically do not have them.

113. Line 312: Without being explicit about the methodology used for the literature review, it is also not necessary to state that no other studies were found. This is semi-implicit (if even required) in only having the two above cita6ons.

Ok, was deleted.

114. Line 318: "… points appeared to have the largest arrows". This statement could be supported quantitatively with a measure of length for the arrows, equivalent to the degree of performance difference between the two model runs.

Results and Discussions are now separated in the revision (as stated before), the "new" results part includes more quantities.

115. Line 322: "… no trend regarding vegetation type or site aridity can be seen …". Were any statistical tests to check for a trend performed here? If not, then changing "trend" to "rela6onship" might be preferable.

Was changed to "relationship"

116. Line 329: It would be good to explore the low EF / high NEE performance in forests in more detail. What processes are likely to be responsible for this mismatch in model performance? It is findings such as these that, with further discussion, would support the statement from my comment 115.

We considered this and decided against intensifying the discussion on NEE-EF relationship. Clearly, more consideration of the processes would be good. However, we already have an in-depth discussion about how LAI and turbulent fluxes are related in the models, and we believe that many of those points already touch on this relation.

117. Line 335: I would include the soil moisture plots in the appendix.

done

118. Line 340: Slightly more explanation for how the underes6ma6on of GPP/LAI could cause the poor EF performance is needed. A few words on the linking mechanisms would be sufficient.

In the EF calculation, LE is in the numerator. Thus, lowering LE reduces EF. When LAI is modelled to be small, the transpiration can only be low (water balance) or, equally, LE is smaller (energy balance) because less energy is used to transpire water. With an underestimation of LAI also the EF representation deteriorates. We added "because the energy fraction that is used for transpiration is underestimated"

122. Line 344: Activating vegetation dynamics in Noah-MP arguably had more than "a small impact" on LE and EF for certain sites. AU-DaS noticeably has significant displacement in position between static and dynamic runs in Figures 5d and 6d. Similarly, comparing the position between static runs for AU-DaS with default and MODIS LAI, there is clearly a large difference in model performance.

Yes, true, there were some exceptions. We mentioned some.

123. Line 346: I would suggest more information on the possible causes of disagreement between Ma et al. and this study. Why might different results have been reached? I would also replace "already concluded" with "found", otherwise it reads as if the authors are dismissing their own results!

Disagreement between statements from Ma et al. (2017) and our study is low. Only the bias values vary. One possible reason might be the differing timescale for the evaluation (daily vs. monthly/annual). Added "…which could be due to the differing timescales for model evaluation".

125. Line 357: To what measurements does "optimal values" refer?

"Optimal values" refers to the values for soil characteristics in look-up tables. Rephrased to "…optimal values for soil parameters are still uncertain".

126. Line 361: This statement is not clear.

Was rephrased

128. Line 363: "Surprisingly, the model quality of those actually closely related variables was independent". This sentence needs work. What does model quality mean? Which variables are considered closely related, and why? How does this affect the confidence in the results?

Agreed, the phrasing was ambiguous and needed more explanation, which was now added. But the finding, that model performance in LAI and in LE seems to be independent of each other although LE values depend on LAI values, does not affect the confidence in the results since it **is** one of the results.

132. Figure 7: Keep the x axes constant across the nine plots. This ensures that comparison between the plots is easy and does not mask the differences in performance. This is also the case for the other figures – where the point of subfigures is to allow comparison between them, ensure that all scales are consistent as this provides ease of comparison. It is also necessary to describe what each element of the boxplots represents.

Thank you! Indeed in this Figure the x axes were not consistent, and this was changed now. For other figures, I could not relate that criticism. The Figure caption is extended.

133. Figure 8: Which LAI is used for the models in this figure? I would suggest less transparency for the MAM and SON points, or just use different colors. Moving the range indicators outside of the ploting area would ensure they do not cover points on the plots.

LAI in Figure 8e-p is the model output. Added to the caption.

135. Line 407: I would suggest changing the units that GPP is reported in such that the values do not need to be reported at so many decimal places.

Good suggestion, done.

136. Line 408: Are the MODIS values of LAI varying between 1 and 7 realistic? It should be clear whether the authors believe the LAI or GPP is the most likely reason for the two variables to not align.

This comment likely refers to the tropical site (GF-Guy) selected for the Figure 8. We were also concerned about this large range of LAI values. However, we handled data quality as careful as possible and used only days with high standard quality flags. Some information on that can be found in the new “Limitations” part.

139. Line 419: This sentence makes it unclear which sites were being discussed previously – the start of the paragraph indicates that all of the sites are being discussed but then here it is stated that similar behaviour is seen at a specific site.

Was rephrased.

145. Line 444: Is the 11% in the model? If so, how does this compare to observations?

Yes, the 11% of assimilation in the model goes into dark respiration. Checking that ratio for the observations is tricky, and we think beyond the already very detailed analysis presented here.

148. Line 458: “However, an evaluation of the representativeness of key variables like lead area index or net ecosystem exchange is rarely done”. I would agree this is frequently a part of model evaluation, and therefore needs to be more specifically worded to accurately infer what the authors are saying.

We added “…on high temporal resolution” to be more specific.

157. Tables A1 – A6: What are the column headings? How do they relate to the different model runs?

The headings were edited.

158. Table A6: This appears to disagree with the statement made at line 334 that model performance for soil moisture is insensitive to LAI forcing or vegetation dynamics. Assuming that each column in Table A6 is one of the different model runs, then sites such as US-SRM (relative bias of ECLand varies from 314% to 552%) appear to have quite varying performance, even if it is consistently poor.

There are some exceptions but for the majority of the sites, soil moisture did not respond to changes in LAI. We just mentioned the majority since the overall manuscript is long and provides much information anyways and we tried not to overload the Results section with small details.

References

Ashaolu, Eniola & Iroye, Kayode. (2018). Rainfall and potential evapotranspiration patterns and their effects on climatic water balance in the Western Lithoral Hydrological Zone of Nigeria. Ruhuna Journal of Science. 9. 92-11ti. 10.4038/rjs.v9i2.45.

Souhail Boussetta , Gianpaolo Balsamo , Anton Beljaars , Tomas Kral & Lionel Jarlan (2013) 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: 10.1080/014311ti1.2012.71ti543

Fang, H., Wei, S., and Liang, S.: Validation of MODIS and CYCLOPES LAI products using global field measurement data, Remote Sensing
of Environment, 119, 43–54, https://doi.org/10.101ti/j.rse.2011.12.00ti, 2012.

Ma, N., Niu, G.-Y., Xia, Y., Cai, X., Zhang, Y., Ma, Y., & Fang, Y. (2017). A systematic evaluation of Noah-MP in simulating land-atmosphere energy, water, and carbon exchanges over the continental United States. Journal of Geophysical Research: Atmospheres, 122, 12,245–12,2ti8. https://doi.org/10.1002/ 2017JD027597

Niu, G.-Y., et al. (2011), The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res., 11ti, D12109, doi:10.1029/2010JD015139.

Nogueira, M., Albergel, C., Boussetta, S., Johannsen, F., Trigo, I. F., Ermida, S. L., Martins, J. P. A., and Dutra, E.: Role of vegetation in representing land surface temperature in the CHTESSEL (CY45R1) and SURFEX-ISBA (v8.1) land surface models: a case study over Iberia, Geosci. Model Dev., 13, 3975–3993, https://doi.org/10.5194/gmd-13-3975-2020, 2020.

Response on Review 2

First of all, I want to thank you for reviewing the work of me and my co-authors. You have spent a lot of effort and emphasized many details. Your criticism is important for improving this work and publication and your hints are useful for reaching this goal. In the following, I will go through and respond to your comments.

General comments (reviewer comment in italic, response in plain text, adaptations in blue):

- *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. Prescribing the LAI is always for initializing the models independent of dynamic vegetation. In the model simulations themselves, this prescribed LAI is only used in the model runs with static vegetation but is in any case part of the initial input of the models independently whether it is used or not. We handled the terminology and the descriptions within 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. 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). These values are also based on MODIS observations which were disaggregated to the different PFTs on each observational grid cell (Oleson et al., 2010). I could not find any information from which time span these values were taken or how individual LAI climatology within one PFT were merged. In the default setup, this LUT LAI was used. For the other setups, those values in the LUT were replaced by "our" LAI values from MODIS.

- *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 evolution. 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. However, we could show 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 (L190-220).

- *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.*

  We extended explanation of model processes concerning dynamic vegetation and added important equations to the appendix.

- *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.*

  Splitting Results and Discussion section is done.

- *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. Although Noah-MP provides no GPP and NEE output for the static runs, it still is interesting to look at the LAI-GPP relationship within the model that we did for Figure 8. Nonetheless, we tried to be more careful with absolute statements and adjusted the abstract and the discussion according to that.

- *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 that information (L162-163).

- *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 sure for Noah-MP, since there is only one vegetation type on the grid cell. For ECLand we would have needed to adapt vegetation to be either high or low vegetation in the initial file. We did this now but it didn't change much. Regarding the model performance of short vegetation types, we could 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, in general, 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.

Specific comments (for brevity here we only give the responses to the specific comments. The line numbers refer to those in the original submission):

- L8: "More detailed information" refers to the on-site LAI. Changed to "…regarding leaf area…"
- L13-14: We didn't 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.
- L21: done
- L24-25: reformulated
- L26-36: 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.
- L29-31: changed "them" to "models" to make it more clearly
- L34-35: added "…since all methods could have a poor individual model performance but there will still be one that performs best, resulting in the highest rank" to explain the disadvantage of only presenting normalized metrics.
- L40: 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. Slightly reformulated.

- L41: 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. Reformulated.
- L45: This is a topic sentence and several works are cited in the following sentences.
- L46ff: 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. However, we have shortened this paragraph a bit.
- L66-67: it's especially interesting because both models are still under development especially with respect to freshly introduced modules like that for vegetation dynamics.
- L80: 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, referring to the aridity index that was created by Budyko & Miller.
- L83: 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. Added an info on logarithmic scale.
- L87: 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.
- L97: 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: 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.
- L97-98: The ERA5 product I 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. This information is included now.
- L100-101: 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. Information added.

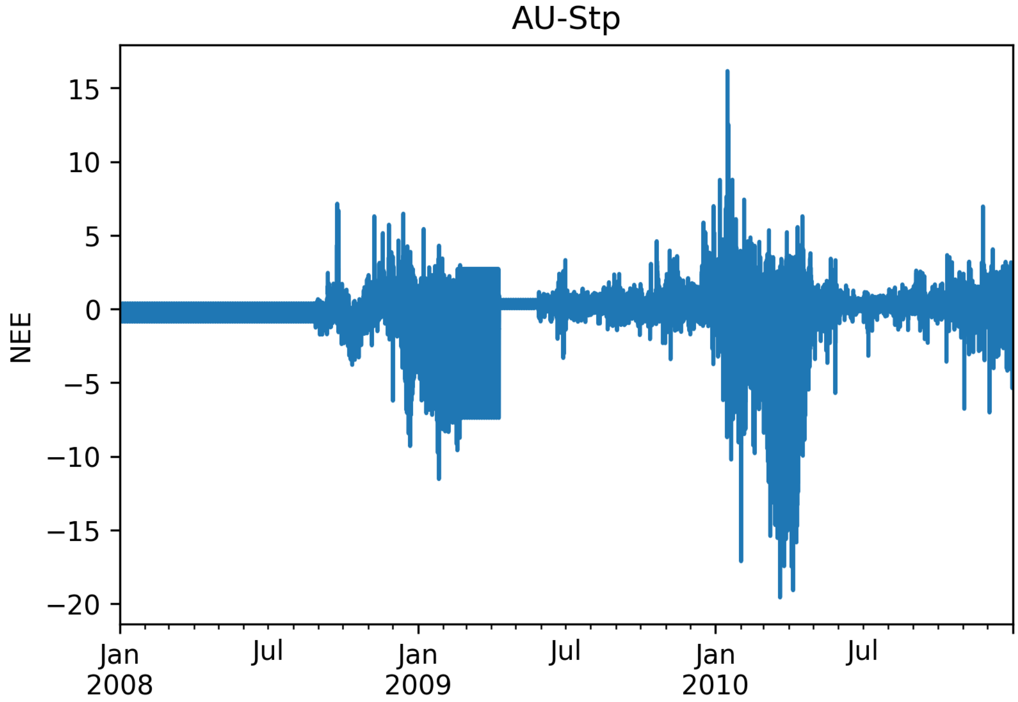

Figure R1: NEE time series for FLUXNET site AU-Stp, exemplarily. Gap-filling from MDS can be identified visually, in this case from January to August 2008 and from March to May 2009. These intervals were left out for model evaluation.

- L104: Temporal resolution is 8 days. There are different MODIS datasets available. The one we used, MOD15A2H, has a spatial resolution of 500 m. Information added.
- L105: Done.
- L106-107: 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.

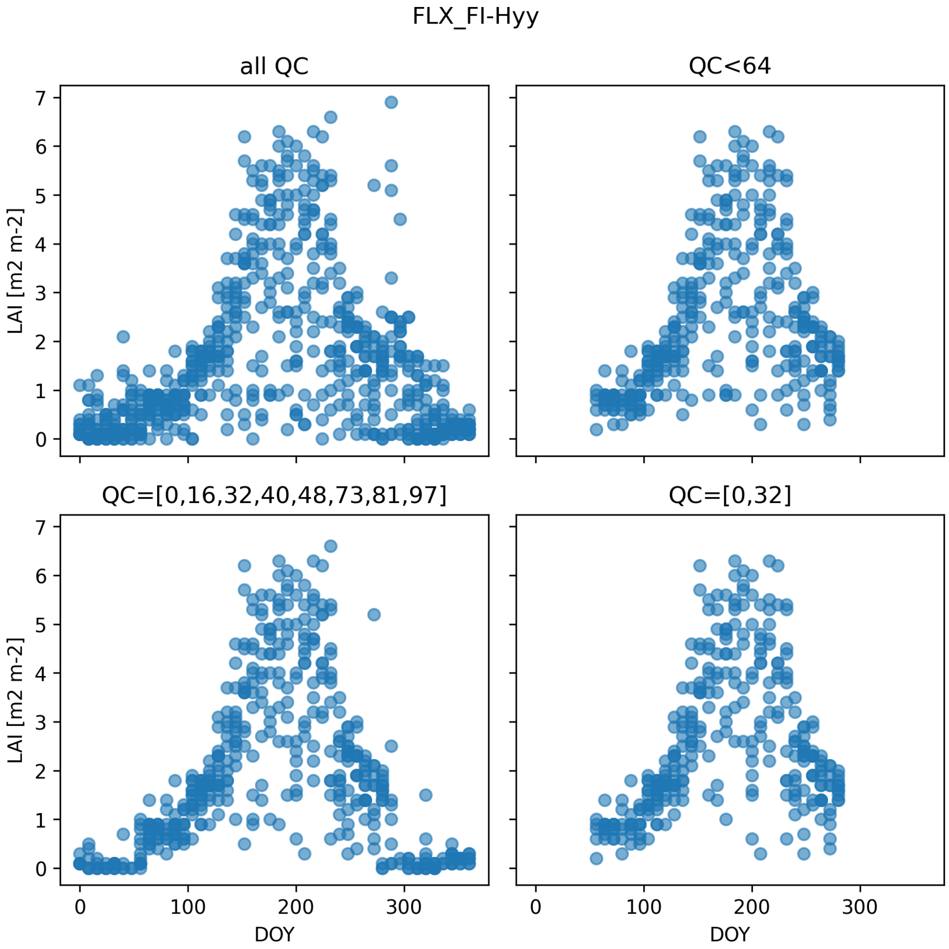

Figure R2: MODIS data points for FI-Hyy when using all quality flags (top left), quality flags less than 64 as recommended by Fang et al. (2012) (top right), specific quality flags in our selection (bottom left) and only "high quality" flags (bottom right). Using only the last category of flagged data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

- L112: See my explanation in point 2 of the general comments.
- Table 1: Rephrased it and added information on the timespan.
- L127: "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 119): From MODIS documentation, every value flagged higher than 0 has some uncertainty or limitation. QC=32 might have the least uncertainty after that. So, I 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.
- L133-134: Yes, at maximum. It could be even one or none.
- L134-142 + L150-156: We tried to leave model description as short as possible. However, more details on LAI-related processes might help and we included them. Extended process explanation and added important equations to the appendix.
- L164: 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.

- L166: 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.
- L169: added.
- L172: 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.
- L173: 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. So, we now stated and took the soil type of the grid cell of the tower location.
- Table 2: done
- L183: Yes, daily averages or sums (depend on variable). added
- L188: In principle, the relationship between observed and modelled values of a target variable is expected to be linear.
- L189-190: done
- L191-193: 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}.
- L199-200: Yes, exactly.
- L199-207: I tried to explain the elasticity in more detail. Unfortunately, I found no publication from environmental sciences that use the same metric, only from economics. Tried different explanation now.
- L213: All symbols that are in the Taylor plots. Changed to “symbols”.
- L217-218: shifted to and extended in discussion section
- L222: Here, we refer only to literature because Stevens et al. (2020) also replaced LUT LAI by MODIS LAI and compared model results.
- L224: 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: 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.
- L226: done
- L226: Increased variance in comparison with static ECLand simulations. Added “…compared to static simulations…”.
- L227: We could not find any tendencies regarding aridity or vegetation type to have positive or negative shift in relative bias. Replaced by “was ambiguous”
- L231: No, I 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…”
- L243-244: done
- L255: Model performance metrics for the MODIS single-year simulations are in the appendix tables. But they are not part of the Taylor plots.
- Fig 2: added.
- Fig 3: 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.
- L281-282: Although Noah-MP provides no GPP and NEE output for the static runs, it still is interesting to look at the LAI-GPP relationship within the model that we did for Figure 8. Apart from that, we still can look at LAI, latent heat flux and soil moisture.
- L284: 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.
- L289: yes
- L291: 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: 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. However, I just recognized by reading your comment that we can only recommend this for ECLand since for static Noah-MP we don’t know the actual performance regarding NEE and GPP. Shifted to “Implications”.
- L329-330: 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: done
- L351: 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.
- L354-359: 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.
- Fig. 7: Sorry that footnote was there by accident.
- L389-396: I added LAI-GPP and LAI-NEE elasticity for ECLand in Figure 7 but excluded NEE-LE and NEE-SM instead. 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.
- Fig. 8: We added description of the arrows to the figure caption. 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.
- L403: done
- L409-410: this is now part of the section “limitations”.
- L454-455: 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”.
- L461-462: deleted.
- L467: 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: Other models have processes implemented differently. So, there is no chance in directly transferring results and conclusions from these two models to others.

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

Souhail Boussetta , Gianpaolo Balsamo , Anton Beljaars , Tomas Kral & Lionel Jarlan (2013) 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

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.

Oleson, K. W., Lawrence, D. M., Bonan, G. B., Flanner, M. G., Kluzek, E., Lawrence, P. J., … Zeng, X. (2010). Technical Description of version 4.0 of the Community Land Model (CLM) (No. NCAR/TN-478+STR). University Corporation for Atmospheric Research. doi:10.5065/D6FB50WZ

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.

Response on Review 3

First of all, I want to thank you for reviewing our work. We tried our best to implement your suggestions and we think that the manuscript has improved substantially.

In the following, I will go through and respond to your comments. (reviewers comments in italic, answers in normal font, changes in blue)

I understand that the authors use some existing Plant Functional Type in the respective models and do not tune PFT parameters to fit the dominant or average plant traits of each fluxnet site. However, the authors still perform comparisons and evaluate the models against observational point-level ecosystem-specific data. Is this the case? I find this slightly inconsistent, since plant functional types conceptually are mostly meant to be used in larger-scale simulations representing the average characteristics of a vegetation category, rather than being compared with site-level data.

Indeed, the footprint of a flux tower observation has a smaller area than the grid cell of LSMs. Nonetheless, comparison of model output against point-level observations such as those from FLUXNET is a common way to perform model evaluation, especially, since most LSMs are able to be used on a wide range of spatial scales. FLUXNET delivers the basis for such a model evaluation on smaller scales.
PFTs are a concept to simplify the parameterization of vegetation that is expected to respond in a similar way to its environment. As a result, they should be transferable and representative for all subtypes of vegetation that are merged into one PFT. If not, they would have been separate groups. We set the vegetation of the considered grid cell within the model to the PFT that fit closest to the on-site conditions to minimize potential mismatches in parameterization.

If we have a model whose parameters are not set to fit the observations (see above), then why would we necessarily expect that switching on a dynamic vegetation module (which is also unparameterized) should increase model performance? One could argue that solely switching the environmental dependency of LAI on should justify this expectation, but isn't the environmental dependency of LAI on average also embedded in a prescribed climatology by definition?

Just to clarify, the dynamic vegetation modules are not unparameterized, only not additionally calibrated for that specific site. We agree that switching on dynamic vegetation introduces environmental dependency of LAI. If the model is allowed to adapt the vegetation (and its productivity and LAI) to environmental conditions, it can be expected that model predictions are closer to the observations compared to simulations with static vegetation. The climatology contains long-term seasonality of LAI. It represents the average temporal pattern of LAI that is adapted to the long-term mean environmental conditions. Intra- and interannual variability as a result of environmental conditions cannot be included into the LAI climatology. To cope for this, dynamic vegetation modules were implemented.

I deeply appreciate that the authors are extensively discussing past research, their interpretations, and their results in full detail, which increases transparency, but the manuscript is overall difficult to read. It would help a lot if the authors split the results from *the discussion points. The manuscript would also need further proof-reading, since one can easily still find mistakes scattered across the text.*

Thank you for this positive feedback. In the revision Results and Discussion have been separated. We also carefully proofread the manuscript.

The results suggest that model performance "regarding latent heat flux or soil moisture is independent of how LAI is represented" (Line 380). This is very counter-intuitive, and one would wonder whether this is the case because LAI is equally badly represented in all cases.

True, we were also surprised by this result. The answer is within section 3.3. This independency does not mean that the predicted values of latent heat flux do not change when LAI is changing. Rather, the model performance in latent heat flux does not change. Together, this means that the mismatch between modeled and observed latent heat flux might be small or large (depending on the site) but is in almost the same extent small or large with a different LAI representation, resulting in the same model performance. We tried to phrase this even more explicitly in the revision.

Given that LAI dynamics are the main focus of this study, it is important that the authors describe in more detail the LAI modules of the two models. They start doing so in Line 420 and discuss allocation, senescence etc., but they need to do this in a comprehensive manner in the method section, and not scattered across the text.

The Model description was extended and now explains all processes involved in modeling vegetation dynamically and important equations can be found in the appendix now.

I think it is really important to show in the appendix panels with mean annual LAI, GPP, and NEE climatologies for every site, showing prescribed and model-predicted LAI. This would help a lot the interested reader understand the dynamics at play.

Good point. It would be definitely interesting to show some time series plots. However, this is not possible for all sites and all model simulations without stretching the appendix a lot or risking that figures are too small to detect graphs. Also, selecting only single sites would add no additional information compared to the graphs of Fig. 8. We therefore decided not to follow this suggestion.

The selction criteria for the fluxnet sites used seem arbitrary. Why do the authors drop sites of roughly similar aridity index? If anything, including more sites would increase the robustness of their results.

Representative site selection is an important issue and we gave it detailed consideration when designing the analysis. When looking at the global distribution of FLUXNET sites, many of them are located in temperate climate on the Northern Hemisphere. Including all sites with more than 5 years would create an overrepresentation of regions with high density in sites, resulting in an imbalance of PFT-aridity combinations for model evaluation with especially (semi-)arid short vegetation being underrepresented. Thus, we needed some sort of filter algorithm to avoid that overall model performance is either shifted towards better or worse performance due to this imbalance. Also, some sites had to be removed due to low-quality in soil moisture data. Unfortunately, there are not enough sites available to create a second set of the same structure (e.g. a representative coverage of climate zones and PFTs), as some aridity-PFT combinations are really rare. We are aware that such a second set would be helpful for strengthening and reproducing our findings. We now explained in more detail how the systematic site selection was done, and adapted Figure 1 in accordance with the model PFTs.

In my understanding, in the switched-on dynamic vegetation runs LAI is freely estimated by the model and not constrained by some prescribed climatology. Is this indeed the case? From the results (e.g., Fig. 2) it feels as if there are different types of dynamic runs – one for each of the different possible LAI forcings. How is this the case? In my understanding, somehow these prescribed climatologies are still used to "initiate" the dynamic runs. What does that exactly mean? If this is indeed the case, still LAI is free to evolve, so why do we end up with different model performance in every run, just because initial LAI conditions have been different? If this is not the case and LAI climatologies are somehow fed also into the dynamic runs, then how does it make sense to validate these results against the forcing LAI climatology itself? The authors need to try and clarify their setup more.

Even when the models run with vegetation dynamics, the LAI climatology is still part of the initial files, independently whether it is used or not. For Noah-MP, these climatological values are not used for the dynamic setup. This is why we end up with the same model performance for all dynamic Noah-MP runs. In ECLand, the vegetation is not totally dynamic. For instance, with a fraction of 5% the prescribed LAI still merges into the LAI estimate for that simulation day (defined by Souhail Boussetta, used as a dynamic ECLand setup). Thus, also model performance of ECLand differs a bit depending on LAI climatology source. The two columns in the Figure should show in which direction and how much model performance shifts for the dynamic simulations compared to only relying on the prescribed LAI climatology of the simulations with static vegetation. We have now added more information on how LAI was used to the Methods section (Section 2.3)

Line 155-156: Was it not technically possible to maintain GPP and NEE estimation in the static Noah-MP runs, despite prescribing LAI? Or is there some other reason?

Static Noah-MP does not calculate GPP and NEE (missing values in the output file). This relates to model structure. LAI for the next time step is already known, so there is no need to estimate assimilation by photosynthesis or allocation to plant tissues. So yes, technically, this is not a model output.

Line 165: It is slightly unclear in which occasion ERA5 data are used

The default initial files were based on ERA5 dataset. We added more specific information.

Line 255: I would suggest that the reviewers show these results in supplementary material to ensure transparency (or don't mention at all)

Ok. The Taylor plot for soil moisture now is in the appendix.