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

General Comments

In this manuscript, Westermann et al. test a range of implementations of vegetation dynamics, namely dynamic vs static LAI from different data sources, within two land surface models, ECLand and Noah-MP. By contrasting the model performance across three different metrics for these different model setups, the authors identify not only which implementations produce better predictions of various fluxes but also potential reasons for these differences in model performance. In particular, the authors find that implementing dynamic vegetation in the two models actually decreased model performance with respect to the analysed fluxes. This is an interesting finding that is of importance to those both performing modelling studies and utilising the outputs of these models in model intercomparison studies.

Unfortunately, due to the number of comments I have, I must recommend major revisions to this manuscript before it is suitable for publication in Biogeosciences. In general, additional work is required to ensure the manuscript is well structured with clear and concise results and discussion. I am sympathetic to the fact that the length of this review and the number of technical corrections may be disheartening, but the majority of these are no more than the result of a further proofread and should be simple to address for the main author. I hope that these comments will provide the necessary guidance for bringing this submission up to the standards required by Biogeosciences.

Specific Comments

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.

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 still 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.

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.

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?

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.

Technical Corrections

1. Line 2: Superfluous "and".

2. Line 5: "improves" should be "improve".

3. Line 7: "range in" should be "range of".

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

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

6. Line 9 and elsewhere: the use of “e.g.” is not ideal as it is not clear what other processes or variables are to be inferred. I would amend this to detail the results more explicitly.

7. Line 10: “while Noah-MP improved it only for some sites” should be more along the lines of “while performance improved in Noah-MP only for some sites”.

8. Line 13: “One reason, we showed here, might …” would read better as “We show that one potential reason for this might …”.

9. Line 17: “For both, water and carbon fluxes” should be “For both water and carbon fluxes”.

10. Line 28: What does “a.o.” mean? I do not believe this to be a standard English abbreviation.

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).”

12. Line 33: “than an ensemble of LSMs”. With ‘ensemble’ often 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.

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.

14. Line 37: “… a closer look on the cause …” should be “… a closer look at the cause … “.

15. Line 44-45: “Nonetheless, ... to be further explored” is a difficult sentence to parse. I recommend rewording this.

16. Line 46: “thereby” is not needed.

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.

18. Line 48: The sentence containing “… is crucial to make use of LSMs …” is unusually worded and therefore difficult to parse.

19. Line 52: “… that LSMs do misrepresent …” should be “… that LSMs misrepresent …”.

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

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?

22. Line 58: “… amongst others because …” should be “amongst other reasons …”.

23. Line 60: Missing comma after “files”.

24. Line 61: Missing “a” between “as” and “prognostic”.

25. Line 62: Missing “of” between “understanding” and “how”.

26. 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?

27. Line 64: “Here, we investigate model performances for water and carbon fluxes especially with focus on vegetation processes.” This sentence could be “Here, we investigate model performance for water and carbon fluxes with a focus on vegetation processes.”

28. Line 66: “… 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?

29. Line 69: “those LSMs” should be “the LSMs”.

30. Line 70: “Does improving one variable, compromise performance in the other or improves it along with it?” could be written more clearly and without the unnecessary comma. For example, “Do improvements in model performance for one variable compromise performance for other variables?”

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

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

33. Line 77 and throught manuscript: There is no space between FLUXNET and 2015.

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

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.

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).

37. Line 94: “e.g.” should be “i.e.”

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

39. Line 98: The Climate Data Store citation is an unusual format – stick to the normal style and create a bibliography item for this dataset.

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

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?

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).

43. Line 105: "as trade-off" should be "as a trade-off".

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).

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

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

47. Line 113: Move the "also" from before "on-site" to before "available".

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.

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?

50. Line 125: Typo of "represents".

51. Line 127: I would change "an under-development vegetation dynamic module" to "a vegetation dynamics module currently under development".

52. Line 130: Delete "in".

53. Line 131: Explain what "IFS cycle "CY46R1" means.

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?

55. Line 131: “parameter” should be “parameters” or “parameter values” (or similar).

56. Line 131: “stomata resistance to water and carbon flux” should be just “stomatal resistance”.

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

58. Line 135 and throughout the manuscript: “… to be used for the vertical exchange with the atmosphere” is superfluous text. There are instances throughout the manuscript where slight revisions of text could improve clarity and conciseness.

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

60. Line 151: Missing “the” between “between” and “canopy”.

61. Line 152: Unnecessary “thereby”.

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.

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.

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

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?

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

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

68. Line 175 and footnotes: Do not use footnotes. Biogeosciences journal guidelines specifically say to avoid them. Just cite this dataset as usual with a bibliography entry.

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?

70. Line 180: Delete “therefor”.

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

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 starting on 1 January in the Northern Hemisphere, yet the Australian sites are potentially at the peak of their growing season in January.

73. Line 183: “transferred” would be better as “aggregated”.

74. Line 184: Unnecessary comma after “output”.

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

76. Line 191: Is “co-domain” the correct term? It usually refers to the domain of a dependent variable. “Domain” is likely the proper word here.

77. Line 192: Typo of “therefore”.

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.

79. Line 199: “To account for” should likely be “to analyse” or similar.

80. Line 200: “variable to that” should be “variable on that”.

81. Line 204: “e.g.” should be “i.e.” as the authors list every metric.

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?

83. Line 205 and throughout the manuscript: An abbreviation has been introduced for latent heat, so it could be used here. This is frequently the case throughout the manuscript.

84. Line 206: Delete “as” before “independent”.

85. Line 209: Again, cite the code as is standard with a bibliography entry rather than a footnote.

86. Line 212: “LAI model” should be “model LAI”.

87. Line 212: Consider changing “The point of optimal model performance is indicated with a star” to “The location an optimal model would occupy is indicated …”.

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

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.

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 relationship between the difference in performance and the site characteristics explored.

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

92. Line 231: “Especially short … performance for LAI” is a difficult sentence to parse.

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.

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 clearly for the default LAI, model performance decreases for AU-Stp and US-Ton.

95. Line 238: “did not contribute to improve LAI” would be better as “did not result in improved LAI”.

96. Line 240: This relates back to the restructuring of the manuscript but starting a new paragraph before “Figure 3” would improve clarity.

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

98. Line 249: Delete “the” from before “disaggregating”.

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.

100. Line 255: “Updating the LAI forcing …” is a sentence that appears misplaced.

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 shift in magnitude in the LAI output which would be striking if caused by the switch to dynamic vegetation.

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

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.

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.

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

106. Line 297: Delete “thereby” after “types”.

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.

108. Line 303: Delete “fluxes” before “predictive”.

109. Line 304: The “is” after ECLand should be “are”.

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 identified as requiring further development? How has the study provided evidence for how these processes should be improved within the model?

111. Line 309: “Statistical measures” is a broad term. I would recommend replacing with the specific metrics that were calculated and explored in this study.

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.

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 citations.

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.

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 “relationship” might be preferable.

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.

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

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

119. Line 341: Delete “and” from before “might also be the reason …”.

120. Line 342: Add “and” before “sensible”.

121. Line 343: Change “has the potential in improving” to “has potential for improving”.

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.

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!

124. Line 350: Delete “more” from before “sufficient”.

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

126. Line 361: This statement is not clear.

127. Line 363: Add “a” before “metric” and replace “the bar plots of Fig.” with “Figure”.

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?

129. Line 372: Typo of “or” as “of”.

130. Line 372: Move “do” from after “LAI” to after “sites”.

131. Line 377: Capitalise L in “ECland”.

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.

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 plotting area would ensure they do not cover points on the plots.

134. Line 398: It is “an evergreen”, not “a evergreen”.

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.

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.

137. Line 412: Replace “depends next to LAI also” with “also depends”.

138. Line 418: Add “of” between “values about”.

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.

140. Line 433: Add “a” between “shows” and “similar”.

141. Line 435: Replace “govern this daily” with “govern these daily”.

142. Line 436: Replace “GPP relates linear to LAI” with “GPP is linearly related to LAI”.

143. Line 439: Replace “phase” with “phases” and add “the” between “biomass from” and “previous time”.

144. Line 441: Add “the” in two places – between “part of” and “senescent biomass” and also between “reduced in” and “case of”.

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

146. Line 446: Replace “minimize net primary production or even produce negative values” with “reduce net primary production, even producing negative values”.

147. Line 454: Delete “However, ”.

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.

149. Line 467: Replace “… higher variability in the ecosystem exchange especially of short or sparse vegetation but this was predominantly …” with “… higher variability in ecosystem exchange, especially that of short or sparse vegetation, but this was predominantly …“.

150. Line 468: It is “a negligible” not “an negligible”.

151. Line 473: It should be “observations”.

152. Line 474: Replace “relation” with “relationship”.

153. Line 475: Replace “linear” with “linearly”.

154. Line 477: Replace “… pinpoints to the reasons of model behavior … ” with “… pinpoints the reasons for model behavior …”.

155. Line 477: In general, the conclusion is very long. I would recommend synthesising the study impacts in more detail in a Discussion section and keeping the conclusion a short summary of this.

156. Figure A1: Even though the sub-panel d would be identical to sub-panel c as stated in the caption, I would still include it. The space is free anyway so there is no cost to this, but it will emphasise the similarity of the two plots, especially if the caption still mentions that they are identical.

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

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.

Bibliography

Cao, R., Chen, Y., Shen, M., Chen, J., Zhou, J., Wang, C., & Yang, W. (2018). A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter. *Remote Sensing of Environment*, *217*, 244–257. https://doi.org/10.1016/j.rse.2018.08.022

Chen, J., Jönsson, Per., Tamura, M., Gu, Z., Matsushita, B., & Eklundh, L. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment*, *91*(3), 332–344. https://doi.org/10.1016/j.rse.2004.03.014

Fang, H., Wei, S., & Liang, S. (2012). Validation of MODIS and CYCLOPES LAI products using global field measurement data. *Remote Sensing of Environment*, *119*, 43–54. https://doi.org/10.1016/j.rse.2011.12.006

Huang, A., Shen, R., Di, W., & Han, H. (2021). A methodology to reconstruct LAI time series data based on generative adversarial network and improved Savitzky-Golay filter. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102633. https://doi.org/10.1016/j.jag.2021.102633

Ma, H., & Liang, S. (2022). Development of the GLASS 250-m leaf area index product (version 6) from MODIS data using the bidirectional LSTM deep learning model. *Remote Sensing of Environment*, *273*, 112985. https://doi.org/10.1016/j.rse.2022.112985