EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-641-RC2, 2022

Comment on egusphere-2022-641

Anonymous Referee #2

Referee comment on "Towards an ensemble-based evaluation of land surface models in light of uncertain forcings and observations" by Vivek K. Arora et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-641-RC2, 2022

General comments:

Arora et al. evaluate land model uncertainty using an ensemble of simulations with different model structure, forcings, and observations. This type of model study is useful to understanding quantities like the land carbon sink in the context of these uncertainties. The results show that biogeophysical variables like runoff and sensible heat flux are most impacted by meteorological forcing, while biogeochemical variables like vegetation biomass are most impacted by having an interactive nitrogen cycle. This is not necessarily surprising, but useful to have summarized here. The results on net atmosphere-land CO2 flux being independent of land carbon state are interesting and could be highlighted more visually. The benchmarking is also useful, and hopefully the AMBER tool can be shared more widely.

The premise that each of the 8 simulations is "equally probable" (abstract) or "equally likely" (introduction) needs more explanation. For example, is the model simulation without carbon-nitrogen coupling as likely as the simulation with this coupling? Are the different datasets equally plausible representations given the details discussed in sections 3.1-3.2? Some discussion on these points would be helpful. While the NBP results (Figure 9) show that the simulations are all within the historical uncertainty range, it is unclear how you would know that a priori to determine which structure, forcings, and observations to sample in an ensemble like this.

There are also a large number of figures and condensing or selecting the most salient results as main text figures would help with length and clarity. For example, do the full timeseries plots of all variables need to be included in the main text? Especially since there are clear groupings of variables that show more sensitivity to, for example, meteorological forcing vs. N cycle. I also felt that some of the ensemble mean figures (i.e., Figures A2-A16) were more interesting than the timeseries plots with all 8 simulations (i.e., Figures 3-9) because they nicely summarize the strongest effects for different variables. In general, the figure organization, number of figures, and placement

of figure in main vs. appendix could be improved to highlight the most interesting results.

In general, the presentation quality needs improvement to better communicate the results before I can recommend this paper for publication. I have highlighted some areas in the specific comments below.

Specific comments:

Line 33 and following: Some useful references to include here would be:

- Fisher and Koven 2020, https://doi.org/10.1029/2018MS00145
- Kyker-Snowman et al. 2022, https://doi.org/10.1111/gcb.15894
- Bonan and Doney 2018, https://doi.org/10.1126/science.aam8328

Lines 110-111: Is AMBER available to the community? Here it is listed as "open-source", but following the link in Seiler et al. (2021b) leads to a dead end: https://cran.r-project.org/web/packages/amber/index.html. Suggest adding an updated link for AMBER to the Code/data availability section in this manuscript.

Lines 177-180: There are other physical processes that could benefit from using a larger number of PFTs, for example, sensible and latent heat flux calculations.

Section 3.1: Here I started to get a little confused with specific land cover datasets. Line 219 states "two observation-based data sets are used", a remotely-sensed product (assuming that is ESA CCI) and the "LUH product as part of TRENDY". The next paragraph describes the process of generating land cover data with the "older" GLC 2000 product, with some information from LUH. Figure 1 compares these three datasets and a fourth one which is based on ESA CCI. Then line 337 (section 3.4) states that the two land cover reconstructions used in the model simulations are GLC 2000 and ESA CCI. Please clarify in the text which datasets are used for land cover in this study and how they are used.

Line 255: What land cover data is used for years prior to 1992 in the case of the simulations with ESA CCI?

Lines 349-351: Do the different end dates for the simulations with different meteorological forcings affect the analysis?

Lines 367-368: Some justification for the doubled weighting of S_rmse would be helpful here, even if a brief sentence/reference. One could argue the other scores also have "importance".

Lines 450-452: This sentence should be rephrased/expanded on since it doesn't add much on its own. Or it could be removed, as Table 3 summarizes the differences in cv values.

Line 529: Curious why the simulations show a land carbon source in the 1930s? Is that realistic?

Line 571-573: Is there a reason to include the SG250m dataset here since the model compares better with HWSD?

Line 574 and following: More discussion of Figure 11 is needed – there are a lot of model/data comparisons here that are summarized very briefly as "the model is overall able to capture the latitudinal distribution of most land surface quantities". For example, the aboveground biomass observations are very different from each other, and different from the model spread.

Line 599: The fact that the interactive N cycle degrades model performance for certain variables is an interesting result that merits some discussion. Some readers may be surprised that something that is essentially a model improvement for more realistic process representation doesn't necessarily improve performance.

Lines 601-602: Thanks for including the full AMBER results. This sentence and link should probably be moved (or repeated) in the Code/data availability section.

Conclusions section: The first two paragraphs of this section could benefit from linking back to specific results in this study with the results placed in the context of other studies (as is done in the third paragraph). Especially for the second paragraph, since model tuning was not covered in detail in the introduction.

Line 642 and following: Curious why the effect of the interactive N cycle is discussed here but the other factors are not?

Code/data availability: There are no references to the code/data used in this manuscript (e.g., the simulation output, how to access the observational datasets, or the code used to generate the analysis and figures.)

Table 2: Suggest also grouping by variables, so it is easy for the reader to see which variables have multiple globally gridded and/or in situ sources. This relates to the calculation of benchmark scores in Lines 383-385 where "at least two sets of observation-based data for a given quantity" are needed.

Table 3: Suggest adding the dominant source(s) of spread for each variable (e.g., met forcing, land cover data, N cycle) to summarize that information across variables. Figure 12 does some of this, but it could be improved for presentation quality as noted below. In addition, there are 14 variables listed in this table, while Figure 12 includes 16 variables and the text mentions 19 variables used for benchmarking and calculating scores. Is there a reason for these differences?

Figure 3 (and following analogous figures): Suggest specifying the exact years shown in these timeseries plots. I believe the end years are different for the different metrological forcings (e.g., 2016 vs. 2019) but it is difficult to see because the lines are very small. Also, please describe in the figure caption what the numbers are in the upper right part of the plot. What is the difference between the bold colored lines and the lighter/less bold lines?

Figures 3-5 (and A2-A5): The data in these figures for 1701-1900 is very repetitive, given the fact that these years use the meteorological forcing from 1901-1925 repeated. The timeseries plots could be shortened to show only 1900 onwards to focus on the most interesting parts of the historical timeseries.

Figure 9: Please add something about the TRENDY models / grey boxes in panel a) to the caption here.

Figure 10: Some additional explanation (in figure caption or text) on how the horizontal and vertical whiskers were calculated would be helpful.

Figure 11: The colors are confusing here. The caption says the model mean is in "dark purple" but it looks more like magenta/purple-red and the dark purple line looks like it is showing an observational dataset (e.g., GEOCARBON in panel a)). Line 577 lists additional colors in regard to this figure. Suggest adding dashes to the observational lines to better distinguish from the model and avoid relying on interpreting specific color choices. Please also explain the box plots in the figure caption.

Figure 12: This figure is a helpful summary of the results, but it needs improvement for presentation quality. For example, the "GLC2000-ESACCI", etc. does not need to be repeated down the column and may not even need to be included since the orange

shading denotes that section as "Effect of land cover". The average scores could be incorporated visually to "provide context" (which was somewhat unclear from the figure caption). Please explain the error bars in the figure caption.

Technical corrections:

Line 87: Community Land Model should be capitalized.

Line 93: Tian et al. (2004) used CLM2 coupled to CAM2, whereas Lawrence and Chase (2007) used CLM3 within CCSM3. Suggest rephrasing this to clarify.

Line 107: Should be "simulated" instead of "simulate".

Lines 353-354: Missing units for grid size (km?).

Line 407: Here the supplemental figures started being referred to as Figure SX instead of Figure AX – please adjust to match.

Lines 518-520: Move "or the net atmosphere-land CO2 flux" up to the first mention of NBP/Figure 9 since Figure 9 uses the latter term and not NBP.

Line 546: Should be "differences" instead of "difference".

Lines 622-624: Check figure references here, I believe both are incorrect.

Line 624: Should be "20 years" instead of "20-year".