EGUsphere, author comment AC2
https://doi.org/10.5194/egusphere-2022-641-AC2, 2022

Reply to Referee #2

Vivek K. Arora et al.

Author 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-AC2, 2022

Reviewer # 2

We thank Referee #2 for their helpful comments. Our replies to their comments are shown in bold below.

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.

Thank you for your overall positive feedback. That the net atmosphere-land CO2 flux is largely independent of the land carbon state is downplayed by referee #1 and they suggested that this conclusion is overstated. While our framework studies land cover uncertainty, it does not take into account land use change (LUC) uncertainty (as noted by referee #1) and therefore we would like to both clarify this caveat to our conclusions and err on the side of caution and not highlight this conclusion more visually.

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.

If given the opportunity to revise our manuscript, we will expand on this. This is indeed the conundrum, it is not known a priori which model structure, forcing, and observation to sample. It is difficult to conclude which meteorological

driving data are more reliable, which land cover is more realistic (as the large spread in the vegetated, tree, and grass area from TRENDY models illustrates), and which model version is indeed better. Hence, the conclusion is that an ensemble-based approach allows a more robust evaluation of a model.

In our case, some general conclusions can be made. For example, we have more confidence in the ESA-CCI based land cover than in the GLC2000 based land cover because the reclassification of 37 ESA-CCI land cover classes to CLASSIC's nine plant functional types has been more thoroughly vetted against high-resolution land cover data compared to the GLC2000. We will, however, remove the phrase "equally probable" since it is difficult to defend and include discussion around this when revising our manuscript.

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 time-series 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 time-series 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.

Swapping figures in the appendix with those in the main text is also suggested by referee #1. We will swap these figures, condense the total number of figures by combining zonal and time series plots for a given variable, and drop some less interesting variables.

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

Thank you for pointing out these additional references which we will include.

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.

We have now replaced the CRAN link with the following link using Zenodo and will include this when revising our manuscript.
https://doi.org/10.5281/zenodo.5670387
The link provides the source code as well as the scripts required for reproducing the computational environment, which takes care of all dependencies with other R-packages.

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.

Yes, in theory, a larger number of PFTs should allow the modelling of PFT-dependent processes more realistically. However, for CLASSIC (and for other LSMs too) latent heat flux is primarily a function of available energy and precipitation. In CLASSIC, large changes in leaf area index (LAI) do not change total latent heat flux considerably since the partitioning of evapotranspiration into its sub-components (transpiration, soil evaporation, and evaporation/sublimation of intercepted rain/snow) changes. For example, a decrease in transpiration and evaporation of intercepted precipitation, due to a decrease in LAI, is compensated by an increase in soil evaporation. As such then biogeochemical processes benefit more in terms of realism than physical processes when the number of PFTs is increased. We will add this clarification around these sentences when revising our manuscript.

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.

The methodology for generating land cover from 1700 to the present day requires a remotely sensed snapshot of present-day land cover (e.g. GLC 2000 (which represents the year 2000) and ESA CCI (1992-2018) land cover products with their 20-40 land cover classes that need to be remapped/reclassified to model's nine PFTs) and an estimate of the change in crop area over the historical period (1700-2018) (LUH). These are the three data sets used in our study. Figure 1 in the manuscript shows the model's total vegetated, tree, and grass area when using the GLC 2000 (blue line) and ESA CCI (red line) land cover products compared against Li et al. (2018) (dotted black line) and other TRENDY models (grey lines). We can see how this can be confusing and will clarify this when revising our manuscript.

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

As mentioned above the 1992-2018 ESA CCI land cover provides a snapshot of the present-day land cover. Although there is some interannual variability between these years overall the total vegetated area doesn't change much. We chose the year 1992 from these data, reclassified its 37 land cover classes to CLASSIC's nine PFTs, replaced the crop PFTs for the present day with those from the LUH product, adjusted the natural PFTs accordingly, and then went back in time to 1700 using crop area from the LUH data. This yields a reconstruction of the historical land cover with CLASSIC's nine PFTs based on the ESA CCI land cover and crop area changing over the historical period from the LUH product.

We will clarify this when revising our manuscript.

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

No, they do not. However, we will make the time period the same for a consistent comparison.

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".

We agree that the decision to give twice as much weight to S_rmse is somewhat subjective. This follows from Collier et al. (2018) but we will make a note of subjectiveness in this decision.

Collier, N. et al. 2018. "The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation." *Journal of Advances in Modeling Earth Systems* 10 (11): 2731–54.

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.

We will remove this sentence when revising our manuscript.

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

In Figure 9a, before about 1970 the model simulates both a land carbon sink and source in response to interannual variability in meteorological data.

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

There are multiple observations available for certain variables. There are times when it is obvious which observation-based data set is better or more appropriate but there are times when it is not. In the case of soil carbon, since CLASSIC and most land models do not include processes to represent peatland and permafrost carbon at high latitudes, it is clear that the HWSD data set is more appropriate for comparison with the model output. The idea behind using both data sets is to illustrate this concept and we will clarify this.

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.

Thank you for pointing this out. We will discuss all panels of Figure 11 in the revised manuscript. In the context of aboveground biomass, the GEOCARBON data set uses two products, one for the extratropics and the other for the tropics to create a global aboveground biomass product. The Zhang product is based on

10 biomass maps. Both products are described in detail in section 2.3.3. of the following paper. We will include this information when revising our manuscript.

Seiler, C., et al. (2022) Are terrestrial biosphere models fit for simulating the global land carbon sink? Journal of Advances in Modeling Earth Systems, p.e2021MS002946. https://doi.org/10.1029/2021MS002946.

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.

Yes, we will include a discussion about why the inclusion of the N cycle degrades the model performance for some variables. The inclusion of the N cycle changes the maximum photosynthetic rate (Vcmax) to a prognostic variable for each PFT as opposed to being specified based on observations. This is analogous to running an atmospheric model with specified sea surface temperatures (SST) and sea ice concentrations (SIC) as opposed to using a full 3D ocean. Using a dynamic ocean allows future projections (since future SSTs and SICs are not known) but invariably degrades a model's performance for the present day since simulated SSTs and SICs will have their biases. Similarly, using an interactive N cycle allows to project future changes in Vcmax (based on changes in N availability) but also degrades CLASSIC's performance for the present day since simulated Vcmax has its own biases.

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.

We will move the AMBER link to 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.

Thank you for this suggestion. We will cover model tuning early on in the manuscript and link it to the second paragraph in the conclusions.

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

The effect of the interactive N cycle is discussed explicitly to outline model limitations and room for improvement when the interactive N cycle is switched on. We will also include a discussion of meteorological forcings and land cover when revising our manuscript.

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.)

The model code and documentation are available at **https://cccma.gitlab.io/classic_pages/ and this is mentioned in the manuscript. Observation-based data sets are available from their respective sources. If providing model output and the scripts used for analysis is a requirement for Biogeosciences we will upload these to Zenodo and provide a DOI.**

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.

Thanks for your suggestion. We will group by variable and reorganize Table 2 according to whether the data are globally gridded and/or in situ, or make note of this in an additional column.

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?

Indicating the dominant source of spread for each variable is a good suggestion. The reason for the different numbers of variables in Table 2 and Figure 12 is that the variables are bundled slightly differently. In Table 2 heterotrophic and autotrophic respiration are evaluated separately but in Figure 12 they are bundled in ecosystem respiration for comparison with observations. In addition, some variables are repeated in Figure 12 (e.g. ecosystem respiration and net ecosystem exchange) because their scores are statistically different when evaluating the effect of land cover, meteorological forcing, and an interactive N cycle. We will make note of this when revising our manuscript. The full list of 19 variables also includes the albedo and leaf area index. We did not include albedo since it shows very little variability across the eight simulations (cv=0.007) and the leaf area index is somewhat similar to vegetation biomass. We will add these two variables to Table 2 for clarity and completeness. In the current manuscript Figures A18 and 11 do compare zonally-averaged values of albedo and leaf area index, respectively, with observations. We will clarify this when revising our manuscript and when discussing Figure 11 in more detail.

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?

We will make the time periods consistent for comparison for GSWP3 and CRU-

JRA driven runs, and make it explicitly clear that the GSWP3 and CRU-JRA end in different years in the figure caption. We will also mention in the figure caption that the thin lines are individual years and the thick line is their 10-year running mean. We will also clarify the numbers in the upper right part of the plot (which are the mean from 1700-1720, the mean over the last 20 years of the historical period, and the difference between these values).

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.

Yes, we agree to make this change.

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

We will clarify in the figure caption that the grey boxes are the estimates based on the Global Carbon Project.

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

The vertical whiskers show the range of eight model scores when a given variable from all eight model simulations is compared to an observation-based data set. The horizontal whiskers show the range when three or more observation-based are compared to each other. When only two observation-based data sets are compared to each other there is only one benchmark score, and therefore there is no range. We will clarify this when revising our manuscript.

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.

We will modify these figures to make the colours more obvious and/or use lines with different patterns.

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.

We will modify Figure 12 to remove GLC2000-ESACCI and other similar

wordings. We will also think about how best to incorporate the average scores visually in the figure. The error bars denote the 95% confidence interval. We will add this info to 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".

Thank you for noting these minor corrections. We will incorporate these when revising our manuscript.