1 General

Beyond the revisions made to the manuscript as response to the reviewers, copied again below, some further minor changes were made:

- A second affiliation was added for Chris D. Jones
- In preparation for final publications, all figures were converted into pdf format, which may result in some minor visual changes
- correcting minor grammar or spelling errors

2 Reply to comments from Referee #1 (Christopher Reyer)

We thank the referee for the constructive comments. We have revised the manuscript according to all review comments we have received. A point-wise reply is given below, with the original comments in italic, and our answers in bold text.

The paper is very well written, maybe a little dense with all the material and the "unconventional structure" but overall it works. See my comments in the attached file, I hope they are useful.

Thank you for your kind comment. The linguistic errors you spotted have been corrected, while further comments and questions to the contents are listed below with point-wise replies.

overall comment: what error metrics are you using? it seems first only visual inspection and then for the summary some error metrics? This needs clarification and some discussion I think.

This is correct, but we feel this is already clarified in the paper, the first part mainly considers means and standard deviations, while the specific metrics in the later section is also described in the text.

Abstract: start "in medias res", maybe one sentence setting the scene for the readers would be good.

An introductory sentence to the abstract has been added: "Simulation of the carbon cycle in climate models is important due to its impact on climate change, but many weaknesses in its reproduction were found in previous models."

End of abstract: role of extreme events? evaluate how models capture importance of fire, dieback of forests etc. on carbon cycling?

While this would certainly be interesting to analyse, this paper is focused to analyse the nitrogen cycle in the mean states, and thus looking at extreme events is beyond the scope of the paper. We added a mention for extreme events to be looked at in future studies, in line with the investigation at higher temporal resolution.

75ff: these two sentences ""Jones et al... weaknesses" should be included earlier in the text. seems odd to have these two sentences after the "objective" of the paper. I understsnd you want to justify why only performinig a global analysis but I think this point can still be made even if the order is changed.

As requested, we have moved these sentences before the introduction of the paper objective.

106ff: Can you provide a reference or an argumentation why the different realisations have similar carbon cycle performance?

We have previously seen this in Gier et al. 2020, and further argumentations about the use of only one realization is given e.g. in Anav et al. 2013. References have been added to this sentence in the paper.

ESMValTool: I think ilamb used tobe the tool to evaluate carbon and related variables fo global models. Why not using ilamb? Do you know it and did you check it?

While ILAMB is commonly used for carbon cycle evaluation in ESMs, ESMValTool has also been used for this (including several chapters of the IPCC AR6). Compared to the standard diagnostics in ILAMB, ESMValTool has the possibility of adding your own diagnostics for targeted analysis, which can be added to its public pool of diagnostics for easily reproducible analysis in the future. The diagnostics in ESMValTool already include the metrics of Anav et al. 2013, which this paper builds on, thus making use of ESMValTool a natural extension. The figures from the diagnostics of this paper will be included in ESMValTool. Furthermore, several of the authors are part of the development team for ESMValTool and thus have a deeper understanding of ESMValTool than ILAMB.

219ff: It can also be two-sided, you have to check that the datasets you are using are really only providing one-sided LAI

The cmor standard is to use the upper leaf surface area and thus a one-sided LAI. Output from CMIP models uploaded to the ESGF follows the cmor standard and thus provide one-sided LAI. The reference datasets also provide one-sided LAI.

252ff: I think such detailed descriptions are better placed in the figure caption where they help to understand the figure.

All detailed descriptions pertaining to figures have been moved to the corresponding captions.

395ff: this should come in the methods

We do not think this belongs to methods but would rather keep this as introductory sentence for the NBP part.

397ff: it is very unsatisfactory that there is no explanation why thee errors occur that lead to the removal of entire models from the analysis. is this rather a technical error or some interactions of model processes? - also what removval

Removal pertains to the removal of these datasets from the computation of the multi-model mean, while the individual results in the scatterplots and performance metrics plots are kept. Thanks to a community comment by David Wårlind, the behaviour of the EC-Earth3 models has been explained due to land-use transitions happening on the 31st of December in LPJ-GUESS, resulting in the strong land sources found in December. After consulting with developers for MIROC-ESM, the rapid changes in MIROC-ESM and MIROC-ESM-CHEM should be a combined effect of several model characteristics. Due to the stochastic nature of the dynamic vegetation model, the death of plants can have a drastic effect. Furthermore, the land use change emission module implemented can cause changes in NBP in the tropics, which is amplified due to the high climate sensitivity of this model. These explanations have been added to the text.

404ff: Again can this not go into the figure caption? and the text focusses on the results shown on the plot and not its description?

Moved.

458: but more recent papers by nabuurs et al. NCC (for example) also start shwoing a declininig sink, possibly turninig into source. but mostly after 2015 I think.

We are aware of this, but as our analysis ends in 2005, this is beyond the timeframe analysed in the paper and thus not mentioned.

574: What other improvement were made in the models that could have an effect? and what about those models who still do not have a n-cycle but still improved? did you find any of those?

Many improvements have been made, see the appendix on land model descriptions, as well as the mention in the abstract. Models also improved without the inclusion of a nitrogen cycle, as can be seen when comparing the non-Nitrogen cycle multi-model means for CMIP5 and CMIP6, and looking at individual models in figures 16 and 17. This paper only focused on the impacts of the nitrogen cycle, without claiming it to be the only reason for model improvement. In the overall CMIP performance section we show a general improvement in models participating in CMIP6 over the predecessors in CMIP5 irrespective of their inclusion of a nitrogen cycle.

705ff: this is very short and contains little actal information compared ot the other description. can the descriptions be harmonized to describe for every model the same processes,set-up, submodels etc are described? and also make a clear statement if certain processes are not included?

The amount of information available for the different land models differs, but we focused on reporting the same quantities, with additional information on added processes and differences between model generations. For more information, we refer to the model papers given in Table 1 and the corresponding model description sections.

Figure 1: How can the MMM lines be outside of the st of the MMM? or am Imisinterpreting the plot?

The multi-model means for nitrogen and non-nitrogen cycle models share the same color, but have different line styles. Their standard deviations are shown with a different hatching style and for CMIP5 only two models contribute to the nitrogen cycle multi-model mean, so its standard deviation is very small, leading to your misconception.

3 Reply to comments from Anonymous Referee #2

We thank the referee for the constructive comments. We have revised the manuscript according to all review comments we have received. A point-wise reply is given below, with the original comments in italic, and our answers in bold text.

The paper evaluates whether CMIP6 models are better in reproducing several selected observations as compared to CMIP5. The authors use the ESMValTool to judge model quality, and focus on LAI, GPP and NBP, and vegetation and soil carbon stocks. The period covered is 1986 to 2005, the spatial resolution is rather course (2 deg x 2 deg) by necessity. It is a pity that they exclude the years 2006-2014 for the CMIP6 models since their ability to reproduce more recent observations is also a relevant topic.

The paper is very careful in drawing clear conclusions, and the authors demonstrate that they are aware of the limitations of their data and the whole study. E.g., calling products like MTE or GLASS "observations" is problematic - for their construction, a lot of modelling is involved, and the uncertainty introduced prior to ESM runs is not easy to control. A drastic example is *for soil and vegetation carbon, where in Figure 15 the observations appear as a single small star; in fact, could you quantify the uncertainty reliably, that star would probably cover a good fraction of the plane shown in Fig. 15. The models are also not constraining the values very much - with vegetation carbon differing by a factor of 3, and soil carbon by a factor of 6 between them.*

In general, when comparing GPP and NBP means and their trends (Figs. 6 and 12), the ESMs are simply not doing well with a large spread and gross deviations from the observations even if the latter are quite constrained. Miraculously, the MMMs are often closer to the observations, but not always (the Christiansen 2018 study is not claiming that the MMMs are closer, only that the ensemble mean error is smaller than the errors of the individual models). However, in some cases, the authors achieve this desired property only by excluding models considered as outliers, which is merely a matter of taste, given the huge discrepancies existing anyhow. The ESMs have a very long way to go before one could consider them as "good" in the ordinary sense of modelling.

This is all not the fault of the authors who are simply reporting the current status of the climate models. The distinction between models with and without a nitrogen cycle, with or without dynamic vegetation, and either with prescribed concentrations or emission-driven makes a lot of sense. It is surprising that the effect of more "process realism" (models incorporating the N and potentially also the P cycle, have dynamic vegetation, and are emission-driven) is by and large absent, or at least do not lead to the clear conclusion to be a must-have. Yes, the GPP is reduced overall when N is a limiting nutrient as expected, but the model quality is not necessarily improved. There are still some striking differences, beginning with the notorious overestimation of LAI and than of GPP as a consequence, and so on.

The paper is rather thoroughly written, has a decent and balanced representation of the state of affairs of CMIP6 and CMIP5, and avoids bold claims about the improvements made from version 5 to 6. It can be easily accepted for publication, apart from some minor points addressed below and in particular in the attached pdf which contains 32 comments and corrections, please consider them all. They still amount only to something between "technical corrections" and "minor revisions".

Thank you for this detailed summary of the paper. The specific comments and corrections in the attached pdf were very helpful in revising the paper.

Specific comments:

l. 35f: how could you know that the models show "improved climate projections" when their present-day performance is worse? Wouldn't that require to gaze into the crystal ball?

We have changed the keyword from improved to self-consistent to more accurately reflect the meaning we wanted to portray.

l. 359: "The models are clustered around the GLASS trend" (ref. to Fig. 6) - of course they are, but with differences of up to 300% ! In short, the models are unable to reproduce the observed GPP trend!

We have reworded and removed clustered from the sentence, instead saying that the model range is centered on the glass trend but models show a large range.

l. 397-402: any ideas why the EC-Earth models have such a "strange December"? Or the "popping up in random months" for MIROC? This almost sounds like bugs in the code. Could you comment on that in the paper? What do the model authors say?

We have added explanations for the model behaviours to the text, and reworded the issues in a more scientific way. The EC-Earth models land source in December is due to the annual update of land use transitions, as explained in a community comment by David Wårlind. After consulting with developers for MIROC-ESM,

the rapid flux changes in MIROC-ESM and MIROC-ESM-CHEM arise due to the stochastic nature of the dynamic vegetation model, which can have drastic effects upon plant death, combined with the land use change emission module and the high climate sensitivity.

Figure 10: this is really tough for the eye with so many lines on top of each other. Choose another way of presenting, or delete the Figure.

We have removed the figure and changed the corresponding text accordingly.

The paper is quite lengthy. A suggestion to shorten it a bit: in the main text, there are detailed legend descriptions for some Figures which are either repeated in the proper figure legend, or in one case (Fig. 18) absent from the figure legend. Delete from the main text and keep it only in the Figure legends.

We have removed the lengthy descriptions from the text and added the missing parts to the corresponding figure captions. Many thanks for spotting the absence of the description in the caption on Figure 18.

l. 591f: a classical modelling dilemma is reported here: more process realism (like dynamic vegetation cover) leads to worse performance. If the main purpose is to reproduce the observations better, shouldn't one follow a parsimonious path and not include these additional processes? This is a rather sharp version of Occam's razor; but still, you are recommending to include them all ("only these models...can account for future changes"), why? This needs a proper justification.

The parts we are recommending to include in all future models are the inclusion of the nitrogen cycle and to use emission driven instead of concentration driven simulations, as written in the conclusions. These do not show worse performance for the present day performance.

However, a very common aim for climate models is a better representation of the future rather than merely reproducing the observable present state. If models can only reproduce the present due to including observationally derived data, their future projections will also rely on estimations obtained from other models and not be fully self-consistent. It is known from paleo observations that large changes in climate, such as during the Paleocene-Eocene Thermal Maximum, have a significant impact on vegetation and its global distribution, as well as the global carbon cycle (e.g. McInerney and Wing 2011). To be able to simulate these changes due to climate change in the future, the implementation of dynamic vegetation alongside similar processes becomes a basic requirement. If added process realism helps models perform future projections while already not showing adverse effects in present-day comparisons, it should experience widespread implementation. Those processes like dynamic vegetation, which currently lead to a much worse performance in present-day comparisons require more iterations before such general recommendations can be made, but will ultimately be required to produce meaningful future projections for investigating changes in the terrestrial carbon cycle. Similarly, the first simulations with an interactive nitrogen cycle showed a bad performance, as seen in the CMIP5 models with nitrogen cycle in this paper, but its implementation has matured in CMIP6 models to the point of earning our recommendation for inclusion in all models.

The sentence "However, paleo observations show that large changes in climate, such as during the Paleocene-Eocene Thermal Maximum, have a significant impact on vegetation and its global distribution (e.g. McInerney and Wing 2011)." has been inserted before l. 591 to justify the need for dynamic vegetation.

My congratulations to a really good paper.

Thank you very much.

Further comments in the pdf we'd like to address:

ESGF 88: it would be an advantage to see a link here, potentially as a footnote, or a reference to "Code and data availability" already here

We have added a reference in parenthesis ("see Code and data availability").

108: up to here, "project " and "experiment" were not mentioned, so it is unclear for the reader what you are referring to. This changes only later in the paper. Please define what you mean by "project" (different from an ESM) and "experiment" (e.g. including the N cycle is hardly an "experiment")

Projects have been changed to phases in the text, the projects were referring to CMIP5 and CMIP6. Experiments are a CMIP terminology referring to different simulations with preset forcings, in this case the experiments used are the concentration and emission driven simulations. We have changed the introduction in line 90 to initially refer to these as experiments instead of simulations.

254: just for clarity - this is the mean of LAI3g, LAI4g and GLASS pooled together?

That is correct.

Figure 2 (scatterplot): 1. What does the [1] in all the axes legends indicate? The unit of dimensionless quantities is usually denoted as [-] 2, What is the difference between "esmHistorical" and "historical"? 3. There is no greater than (>) in the legend, but in the Figure and in the caption

1: In the cmor standard, the dimensionless unit for LAI is given as 1, so to be consistent with the cmor standard we remain with [1] as unit for LAI. 2: We have added these experiment names in the data section, but also changed the legend to consistently use the "c" and "e" notation to refer to concentration and emission driven simulations respectively. 3: We have extended the legend to include both the filled and unfilled symbols corresponding to nitrogen and non-nitrogen cycle models, previously the notion of the filled symbols nitrogen cycle models was also mentioned in the caption.

Figure 7: The numbers at the colorbars below the figure fit well with the unit given in parenthesis [kg C m-2 yr-1]. You should report the overall mean also in the same unit, not as a total sum. Thus, the "e+14" makes little sense, or is it simply an error? Please correct.

The plots reported the global total flux, i.e. the sum, to be compared to the tables, but we have changed the numbers to report the global mean flux instead to avoid confusion.

4 Reply to Community Comment by David Wårlind

Why EC-Earth showed a very strong land source in December is due to that all land-use transitions is happening on the 31st of December in LPJ-GUESS. To not get a peak in CO2 concentrations on the 1st of January every year, the land-use fluxes are released to the atmosphere evenly on every day the following year in EC-Earth3-CC. But the LPJ-GUESS outputs are still following when it actually happens in the LPJ-GUESS.

Thank you very much for this clarification, we will add a sentence in the revised paper to mention this.

References

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R., & Zhu, Z. (2013). Evaluating the land and ocean components of the global carbon cycle in the cmip5 earth system models. *Journal of Climate*, *26*(18), 6801–6843. https://doi.org/10.1175/Jcli-D-12-00417.1

Gier, B. K., Buchwitz, M., Reuter, M., Cox, P. M., Friedlingstein, P., & Eyring, V. (2020). Spatially resolved evaluation of earth system models with satellite column-averaged CO_2. *Biogeosciences*, *17*(23), 6115–6144. https://doi.org/10.5194/bg-17-6115-2020

McInerney, F. A., & Wing, S. L. (2011). The paleocene-eocene thermal maximum: A perturbation of carbon cycle, climate, and biosphere with implications for the future. *Annual Review of Earth and Planetary Sciences*, *39*(Volume 39, 2011), 489–516. https://doi.org/https://doi.org/10.1146/annurev-earth-040610-133431