The comments of the reviewer are in *italics*, and author responses in blue plain type.

Authors compare the behaviour of coupled terrestrial N and C cycles in five models that are contributing results to sixth phase of CMIP (CMIP6). The subject of the manuscript is of clear and significant interest to the Earth system modelling community as more and more land components in ESMs explicitly represent terrestrial N cycle and given the large spread among land C cycle models. However, in its current state the manuscript appears to be written hastily with several points unclear, statements that are weakly supported, some incorrect statements, and at places the analysis of results is as simple as which model produces high values of a given quantity and which low.

This paper was produced under a number of constraints that were sub-optimal, but of course the reviewer should judge the paper on its own merits. We assure the reviewer that this paper is not intended to be disrespectful of reviewer time or the community at large. We aimed to provide a clear appraisal of the models and their performance, without pretentions. These are new models and their collective and comparative performance is not commonly known, and the mechanisms behind the differences are still under investigation. We apologise that the reviewer found some parts difficult to understand and have used the comments to improve the paper's clarity.

I have three major comments.

First, nitrogen used efficiency (NUE) as introduced in equation (1) is simply C:N ratio. In the current literature NUE is typically defined as an efficiency indicator for the utilization of nitrogen in agriculture and food systems (Fageria and Baligar, 2005). That is, higher the NUE the lower amount of applied N enters the environment. I suggest, to avoid confusion with existing definition of NUE authors simply use C:N ratio in equation (1).

While it is true that in agronomic research NUE is defined as the efficiency of nitrogen recovery per unit added fertiliser, in ecosystem research it is, as in our paper, defined as the growing season integrated nitrogen requirement for growth, or in other words the net primary production per unit nitrogen uptake. While the unit is that of a simply C:N ratio, we note that there is a difference between the NUE and the C:N ratio of vegetation given the different turnover times of the various tissue types, nutrient retranslocation upon senescence as well as plant N inputs due to N fixation. As such, it is a relevant measure of performance of N-enabled terrestrial biosphere models, which has been used by the community. We appreciate this might cause some confusion for non-nitrogen modelling specialists, it seems sensible to follow the established method set for the terrestrial biosphere modelling community rather than switching to a definition used by agronomists.

Second, the authors have compared the results of two experiments, +CO2 and +N, from models with observation-based estimates. I feel that, the observation-based estimates and the experiments they were based on have not been properly introduced. Nor do the authors discuss limitations of these real world experiments whose results are used to evaluate models. For example, the results from +CO2 experiment used to evaluate models are based on the Baig et al. study which is a meta-analysis but a reader is never told about this. How many studies does this meta-analysis summarizes results from? Similarly, for the +N experiment, the LeBauer and Treseder (2008) study is also a meta-analysis. Both these meta-analyses, results from which are used to evaluate models, should be properly introduced and their limitations discussed. For example, the +CO2 type experiments done are based on instantaneous doubling of CO2 while in the real world CO2 is increasing gradually. Similarly, in +N experiments additional N application rates, I think, are increased instantaneously while in the real world N deposition rates have increased gradually. In addition, can the average *results from meta-analysis be used to evaluate the globally-averaged response. The photosynthesis theory says that the CO2 fertilization effect must be strongest in the tropics. How does one account for this? Were the studies used in meta-analysis uniformly distributed geographically speaking? As a modeller myself, I realize, the business of evaluating models is difficult but as long as limitations of observation-based estimates are mentioned, it allows readers (and authors too) to make a rationale and informed expectation of the extent to which observations and models should compare well with each other.*

We understand this concern and have made several changes to ameliorate it. The visual comparisons in (now) figures 2 and 3 have the potential to be misleading, so we have removed those and added a new table (table 3), which contains the same information about the observations but has more room for nuance. We have added a new sub-section to the methods with more details of the observational data used for comparison. We have extended the existing part of the Discussion, which discusses the limitations of the observations, with the reviewer's suggested topics above.

My third comment is that as a reader, after reading this manuscript, I am not sure if I know anything more about N cycling in models than I did before.

Since CMIP6 experiments have been so long in being published some well-informed readers, such as the reviewer, will have seen many of the results in conference presentations etc., therefore the results may feel less 'fresh' to them. However, that does not detract from the fact that this information has not been published (in a journal article) before.

The +N and +CO2 experiments are, to our knowledge, the first published results of this kind for a range of LSMs with nitrogen cycles which are used in CMIP6. There is a value in presenting what the state-of-the-art models are doing and comparing them to each other and available metrics, even if we cannot completely explain why the models differ. In the revised version we highlight the implications of fundamental model assumptions regarding spin-up and biological N fixation, which should be informative for future use of the models.

Further, many researchers on the fringes of either CMIP6 or nitrogen in LSMs will find this paper a useful summary of the key results and model features. While senior researchers such as the reviewer may find marginal benefit, there are many more junior researchers who will find it very useful.

I feel, the results from these models need to be analyzed and reported in a much more clever way to provide overarching conclusions.

As a modeller, the reviewer will be aware that model comparisons do not always provide the neat generalisations one might hope for. Unlike carbon, nitrogen model structures are very heterogeneous, and their effects are confounded by co-occuring differences in the treatment of the carbon cycle, making simple overarching conclusions inappropriate, as they would be misleading. Therefore, more nuanced conclusions, whilst not "clever", are more scientifically robust. An alternative approach, to implement model assumptions into one common framework (e.g. Meyerholt et al., (2020)), results in a cleaner identification of process importance, but can also rightly be criticised as not representing the effects simulated by the actual CMIP6 model.

Note that the ability of models to simulate recent trends in GPP and NBP is not due to N cycle. Models without N cycle can achieve this too as is seen in the TRENDY intercomparison which contributes results to annual Global Carbon Project studies (Le Quéré et al., 2018).

We appreciate the reviewer's sense of humour in pointing out TRENDY simulations when one of the authors of this paper is the lead author for GCP 2019. We included this section because what is notable about these models is that despite incorporating a major new model component to models which simulate GPP/NBP well the models are *still* able to simulate recent trends. As the reviewer is no doubt aware, it is not a given that a model will (continue to) perform well after a major change has been made. Examples exist (e.g. Koven et al., (2013)) where the original inclusion of a N-cycle representation limited the ability of the model to reproduce the contemporary carbon balance. However, upon reflection we agree with the reviewer's implied point that this case is made sufficiently in other places already. Therefore, we have moved the first two figures to the SI, removed the associated text, and incorporated brief references to the figures into the main text regarding the N budget.

Other comments

Page 1, abstract, lines 26-28. Upon reading these lines it is clear that 200 ppm CO2 and 50 Kg N/hectare.year N deposition increase are both hypothetical. But as a reader I was wondering what observations are used. At this point in the abstract the reader is not aware that model results are being compared to results from meta-analyses later in the manuscript.

Obviously the observations used are important, but we do not feel it is appropriate to detail in the abstract all the different sources of observational data. Since meta-analyses are observation-based, we feel this is a fair brief description for the abstract of a paper that focuses on the models.

Page 3, line 85. Please consider rewording "All models ran a global spin-up for all ecosystem pools up to the year 1860" to "All models pools were spun up to equilibrium using climate and other forcings corresponding to year 1860".

We have changed this paragraph to make this clearer.

Page 4, Section 2.2 and 2.3. Please consider summarizing in a table the runs performed. After the pre-industrial spin up, it seems, three runs have been performed – a 1861-2015 historical simulation, a +CO2 simulation for the period 1996-2015, and a +N simulation for the period 1996-2015.

We have inserted a table of this information to the SI.

Page 4, equation (1). Please use C:N ratio in this equation as opposed to NUE.

This equation follows the precedent set in Zaehle et al. (2014a), and we regret that it would not be appropriate to change it and therefore make it inconsistent with previous work on the same topic by (some of) the same authors.

Page 4, equation (2). This equation is incorrect. Change in NPP cannot be simply determined by multiplying the changes in NUE and N uptake. Please see https://en.wikipedia.org/wiki/Product_rule which explains the product rule of differentiation.

We apologise for this mistake and have revised this section accordingly and removed this equation.

Page 4, equation (3). Please define delta N (which implies N balance, I think) properly in words. It seems it is the change in total amount of N in the land (Tg N). But the right hand side terms of the equation are all fluxes which implies the units of N should be Tg N/year. I am confused. The term "N balance" is used throughout the manuscript. It is an important term and yet in the absence of clear worded definition and units it is difficult to follow the context in the rest of the manuscript where this term is used.

Thank you, we recognise this was not sufficiently clear and have amended the line before equation 3 (now equation 2) appropriately.

Page 5, lines 136-137 reads "This generation of N models are generally consistent within observational constraints, showing an improvement compared to CMIP5 N models". However, nowhere in the manuscript have model results from CMIP5 models been shown so how can one conclude CMIP6 models are better than CMIP5 models. Please reword this sentence.

This sentence has been deleted, as it evidently caused more confusion than the benefit of the point which was being made was worth.

Page 6, line 168. Please consider replacing "non-N model structure" with "C cycle related processes".

Changed as suggested.

Page 6, line 174 reads "Across the ensemble there is a slight correlation between the global GPP total and NEP". Please note that for the pre-industrial spin up models' NEP is zero since the model has been spun up to equilibrium. This implies for the pre-industrial state there is no correlation between GPP and NEP. Over the historical period, there is no reason to expect a strong correlation between absolute GPP values and NEP. What is expected is a strong correlation between rate of increase of GPP and NEP since it is the rate at which GPP increases that determines the land C sink.

Page 6, lines 175-176 are unclear.

This paragraph (lines 174- 177) has been removed as it appears to be confusing rather than enhancing the overall message of the paper.

Page 6, line 186 reads "BNF on the other hand has a wider observed range . . .". For a reader it is unclear where the observed range of BNF comes from.

Reference added in the text (reference already in Figure 3, referred to in the previous sentence).

Page 7, lines 202-204 read "Looking at inputs and losses excluding anthropogenic N addition (BNF + N Deposition – N Loss), all the models have a surplus of N and could be said to be 'open' systems with regard to N balance". I am not sure what this means. Recall that after the pre-industrial spin up the sum of all model input N fluxes should ideally be the same as output model fluxes. Was this evaluated? During the transient simulation additional N deposition and fertilizer input leads to increased gaseous losses of N, perhaps increased leaching, and accumulation of N in organic and inorganic pools. I am unsure what 'open' and 'surplus of N' means– does it mean all the additional N input is lost as gaseous fluxes and to leaching. We all know BNF (especially due to increase in crop area) and N deposition increase over the historical period so N balance, as defined in the manuscript, will always be +ve. What's more important here is where does this additional N ends up?

The issue of whether there is an "open system" with regards to N is a debate popular in some parts of the N community. This comment has made us reconsider whether this is an aspect is worth highlighting and therefore we have removed this sentence rather than enlarge and explain a point that is, perhaps, esoteric.

With regard to the spin-up, as mentioned already in the text of the methods section 2.2, the models were spun-up to equilibrium (as specified by the protocol). The model groups were responsible for ensuring their model was appropriately spun-up and are experienced with running simulations such as these with their models.

Page 7, line 206 reads " . . . Soil+Litter C is generally low, compared to observational estimates . . .". Does the observation-based estimates contain C from peatlands? The CMIP6 models, I suppose, do not account for C in peatlands and perennially frozen C in permafrost. Could this be the reason for low model estimates.

Page 7, line 209. "Comparing the C:N of Soil+Litter global total weight the ratios are similar across models . . .". This sentence doesn't read properly. Also, this section reports the reason for higher C:N ratio of the soil organic matter in the JSBACH model as "The higher ratio for JSBACH is due to the 10:1 ratio for slowly decomposing soil carbon (humus) and larger ratio for litter". I cannot follow what this sentence is trying to imply. This is true for all models. Soil C always decomposes slowly than litter.

Given the reviewer comments we have given a lot of thought as to which parts of this paper are providing the most pertinent discourse. We decided this paragraph on soil and litter C is not sufficiently useful, given the uncertainties the reviewer mentioned, to warrant it remaining. Therefore, we have removed it.

Page 7, Section 3.1. In Figure 4, sub panel, it seems the global model response is compared to observation-based estimate from Baig et al. 2015. Is the Baig et al. 2015 average representative of the whole globe or weighted heavily towards certain geographic regions.

The Baig values are taken from Table 3 in that paper and refers to the % eCa effect on total biomass. It is based on 82 observations of woody plant responses in a variety of locations. The value comes from an equal weighting of each value, regardless of geographical location. We have replaced the reference to Baig et al. (2015) with that from Song et al. (2019) which has more observations and is not limited to woody plants.

Page 8, lines 226-227 read "Therefore, although the models reach a majority consensus on +CO2 NPP effects overall, the important regional details are still contradictory". Does OVERALL in this sentence means globally? When the manuscript says the "important regional details are still contradictory", I think, it is meant that regional response to +CO2 do not agree amongst models. I think, it doesn't mean they the models contradict some observations because there aren't any regionally aggregated +CO2 observation-based responses. Please reword this sentence.

Reworded to: "Therefore, although the models reach a majority consensus on +CO2 NPP effects globally, models show contradictory responses for some important regions."

Page 8, lines 235-239. I wonder, if there a way to quantify or plot this dichotomy between +N and +CO2 responses.

We have added a new plot and results paragraphs to show this.

Page 8, lines 255-256. "The largest responses to +N and +CO2 of input and loss do not necessarily correlate with either N uptake or changes to productivity". I am not sure what this sentence means.

We have revised and hopefully clarified this sentence and the rest of the paragraph.

Page 9, line 267. "In contrast, JSBACH has less than half the increase in loss of JULES in the +N simulation". By the time, a reader reaches this sentences he/she may forget what quantity is being referred to. Does this sentence refers to plant N uptake?

We hope that the clarification of "N loss" rather than "loss" will help aid the reviewer's understanding. The entire context helps, as the previous sentence discusses N loss, and since the sentence begins, “in contrast”, we hope it will now be clear: “In common with all the models, in JULES the N loss term is a fixed fraction of the mineralisation flux and the soil N pool size. In contrast, JSBACH has less than half the increase in N loss of JULES in the +N simulation (Fig. 6c) and almost no change in NUE (Fig. 7d).”

Page 9, lines 270-271. “Two of the most important factors for plants’ use of N are the availability and demand for N use. The variability of these processes is determined primarily by the BNF and NUE respectively, which are both known to be affected by increased CO2 and N”. This statement is not entirely correct. Variability in N demand is not primarily governed by C:N ratio (which is referred to as NUE in the manuscript). C:N ratio of plants changes gradually. The variability in N demand comes primarily from variability in NPP in response to interannual variability in climate. N availability on the other hand depends on pool sizes of ammonia and nitrate. While, BNF is the primary natural mechanism of inorganic input to soil the subtlety here is that pool sizes do not vary substantially from year to year while BNF does. So, I think, variability in N availability has to be very small. Plant N uptake on the other hand will likely be more variable because both passive and active N uptake depend on variability in climate. Please consider rewording this statement.

We have revised these sentences and much of the rest of the paragraph to enhance clarity.

Page 9, line 275. “The BNF responses to +CO2 of the models differ from the average response recorded in a global meta-analysis of CO2 manipulation (Liang et al., 2016)”. Here, Liang et al. is yet another meta-analysis that is being used to evaluate models without properly introducing it first.

As mentioned above, we have introduced a new methods sub-section in the methods to introduce all the observations used as comparisons.

Page 9, lines 279-284. This discussion about BNF is hard to follow.

We apologise that this section was hard to follow and have revised the text thoroughly.

Page 10, lines 300-301. “The large variations in signal and sign of BNF and NUE response between models suggests there is still progress to be made”. Perhaps reword this as “The large variations in the magnitude and sign of BNF and NUE responses to +N treatment between models suggests there is considerable uncertainty in our understanding”. There are now several meta-analyses (including that of Liang et al. 2016) that clearly show that elevated CO2 leads to increased BNF and studies that show elevated N input decreases BNF. This is also intuitively expected. So, I think, there is sufficient evidence to suggest a real world sign (+ or –) on the response of BNF to these two drivers (+CO2 and +N).

Text adjusted as suggested.

Page 11, line 343, reads “The models mostly represent high latitude northern hemisphere regions less well than other parts of the world, in part because of the unique challenges these areas set for models”. I am unsure how can it be concluded that high latitudes are represented “less well than other parts of the world”.

The evidence is given in the results section and this sentence has be reworded to refer the reader back to the appropriate section.

There are no gridded observations for +N experiment. Does this refer to the fact that the models do not agree at high latitudes. If yes, please say so explicitly.

While there are no gridded observations for +N, there are (as shown in Fig. 6) three separate estimates for +N for Tropical, Temperate, and Boreal regions. We concede that those biomes do not cover the whole globe, but think it is a fair statement. It is an attempt to draw a tentative "overarching conclusion" of which region is most challenging for all N models. The reviewer mentions in their first points about this paper that they would like more "overarching conclusions" so presumably would support this.

Page 11, lines 345-349. If I am following the manuscript as the authors intend, it seems the complex processes at high latitudes including potential for release in methane, albedo changes with vegetation expansion, and large amounts of C in soil are mentioned as why the +N response in this region is higher than the average seen in LeBauer and Treseder (2008) meta-analysis. I am not sure if I follow this reasoning because it hasn't been explained how these complex processes are linked to N cycle processes.

We apologise that this point was not as clear as it ought to have been, as the reviewer has misunderstood. We have rephrased this paragraph and hope the point is now clearer.

In addition, were any of the individual studies in the LeBauer and Treseder (2008) meta-analysis performed in the tundra region?

Yes, there is n=10 for Tundra in the LeBauer and Treseder (2008) study (see Table 1 in that paper).

If yes, what was their response to +N?

35% increase in aboveground net primary productivity (LeBauer and Treseder (2008), Table 1).

What is the northern most study in the LeBauer and Treseder (2008) metaanalysis?

78˚North, a Tundra site on Ellesmere Island (LeBauer and Treseder (2008), Fig. 1).

Page 11, lines 357-358 reads "For +CO2 there is the potential for increased NPP because the NUE increases, giving productivity increase without an increase in LAI". I am unable to follow this argument. Isn't is that the productivity increases in the +CO2 experiment simply because of the CO2 fertilization effect? The increase in NPP (due to CO2 fertilization effect) results in a higher C:N ratio of vegetation (which is referred to as NUE), and not caused by C:N ratio as this sentence seems to imply.

We have reworded this sentence to make this point clearer.

Figure 1. Please plot continental boundaries.

We have added continental boundaries to the map figures.

Figures 2 and 3. Using similar shades of green and blues for only 5 models is confusing. Please consider using other colours as well.

We are sorry the reviewer found the viridis colour palate (yellow, light green, dark green, blue, and purple) confusing. Viridis was designed to be comprehensible to a wide range of visual impairments and also converts well to grey scale to allow the paper to be printed black and white to save costs and environmental impact. See https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html for further information.

Figure 3. The arrow for heterotrophic respiration (rh) should come out of the SOM+Litter C pool not the vegetation pool.

This has now been corrected, thank you for noticing this mistake.

Figures 4 and 5. The ratio of small numbers are always misleading and not as meaningful. I am wondering if the geographical plots in Figures 4 and 5 would provide more information if plotted in gC/m2.year rather than percentage change. I realize that the observation-based estimate is in the percentage.

We have added the figures showing absolute amounts for the +N and +CO2 experiments as latitudinal averages to the SI and referred the reader to these plots in the main text. While we agree that the absolute numbers are useful, given that most observational comparisons are a percent change, this is most useful for the main figures.

Figure 7. The y-axis titles "BNF response" and "NUE response" are perhaps better written as "BNF change" and "NUE change", although please use C:N ratio instead of NUE .

"Response" has been changed to "change" in Fig. 7 (now Fig. 5). For consistency, as discussed above, we would prefer to leave NUE.

References other than those that are already in the discussion manuscript Fageria, N. K. and Baligar, V. C.: Enhancing Nitrogen Use Efficiency in Crop Plants, Adv. Agron., 88, 97–185, doi:10.1016/S0065-2113(05)88004-6, 2005. Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2019-513, 2020.

References mentioned in response:

Koven, C. D., Riley, W. J., Subin, Z. M., Tang, J. Y., Torn, M. S., Collins, W. D., Bonan, G. B., Lawrence, D. M. and Swenson, S. C.: The effect of vertically resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4, Biogeosciences, 10(11), 7109–7131, doi:https://doi.org/10.5194/bg-10-7109-2013, 2013.

Meyerholt, J., Sickel, K. and Zaehle, S.: Ensemble projections elucidate effects of uncertainty in terrestrial nitrogen limitation on future carbon uptake, Global Change Biology, n/a(n/a), doi:10.1111/gcb.15114, 2020.