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

The topic of the paper is relevant to the ongoing evaluation of the sign, magnitude, spatial variability, and potential future trajectory of land ecosystem feedbacks among increasing CO2, increasing N deposition, physical climate variables, and the global scale cycling of carbon and nutrients.
Some results presented here could help to inform the evaluation of existing models or the development of new modeling approaches, in particular the results summarized in Figure 8.
On the whole, the results are presented in the form of assertions without adequate quantitative support, and without sufficient process-level elucidation of either the behavior of individual models, the differences between models, or the relationship between models and observation-based datasets. The manuscript overall suffers for having too much description and not enough explanation.
We aimed to provide an honest appraisal of the models and their performance, without affectations. 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 have comprehensively reworked the paper with additional analysis and added quantitative references in convenient locations for the reader.

L53. Either "an important source", or "the main source".
Changed to "an important source".

L54. Meaning not clear
We have changed the phrasing to "overall performance" which we hope will be easier to comprehend.

L56. CLM?
We presume the reviewer means to point out that CLM is not European, and we absolutely did not mean to imply that. We were careful in our phrasing of "European ESMs" (not LSMs). We have clarified this to "of ESMs used in European Earth System modelling centres". We further clarify this in section 2.1: CLM4.5 is used within CMCC-CM2, and CLM5 is used within NorESM. The reason for specifying European ESMs is to give some rationale for the choice of LSMs. This project is centred on the EU Horizon 2020 project CRESCENDO so this is useful in leading the reader to the reason why other models are not included.

L69. Add citations to Thornton et al. 2007, 2009
Citations added.

L74 Does not appear in bibliography.
In the version reviewed this reference is at line 613 of the reference list (there is no bibliography). In the revised version this reference can be found between Sellar and Smith.

L87 land cover
Changed.

L96 Can you use the current results to test that assumption?
Unfortunately not, as that would require a whole new set of simulations (with identical initial states) that we do not have. The model protocol was carefully set up and was designed to test hypotheses

about nitrogen parts of the model and we believe it to be robust. While the question of this assumption could be interesting, it is beyond the scope of this particular project.

L116 Does this approach account for retranslocated N that is taken up once, then stored in the plant and reused for multiple growing seasons?
Note that this is a diagnostic calculation and is based on the modelled rates of NPP and Nuptake. Therefore, it does implicitly account for retranslocated N as dealt with by the individual models, as the rates of retranslocated N in year x affects N uptake in year x+1. Retranslocated N is considered to be within the plant and thus utilised (even if that utilisation is as labile N stores). Incorporating retranslocation into analysis would be challenging, as the models deal with retranslocation in a variety of ways that are not easily comparable.

L123 The meaning here is unclear. If allocation patterns at the plant level shift between tissues with different C:N, wouldn't NUE also shift?
We have revised this section to aid comprehension.

L153 There are many different kinds of model deficiencies that could lead to such mismatches. There are also potential deficiencies in the data that could explain the differences you observe. What can you say, mechanistically, about the models or about the data that might shed some light on the possibilities?
L156 A couple of issues with this paragraph. First, looking at the Fig 1a and 1b results, the differences between the two models outside the high latitudes seem even more pronounced - for example over the African tropics, Indonesia, southeast Asia, and the southeast of North America. Second, this paragraph simply lists a range of differences between the models, but doesn't do anything to try to explain why those differences might be related to differences in GPP in particular regions, or how any of that might relate to differences in connections between nitrogen cycling and GPP.
L165 I don't understand why the model results in Fig 2 are normalized as anomalies with respect to their own 1901-1910 period. The GCP assessments shown in black in Fig 2 do not include that same kind of anomaly calculation. The models should be showing losses of carbon from land during the early stages of the simulations, due to the effects of land use and land cover change. So they are probably all at rather different values of NEP in the 1901-1910 period (although hopefully these raw values would mostly be on the negative side). In other words, It is the raw NEP that should be compared to the GCP numbers.
L168 This paragraph is mainly composed of unsupported assertions. Yes, the two CLM variants are the lowest of the five models for the (anomaly-adjusted) NEP. See above comment for why the anomaly adjustment makes even this apparently simple assertion difficult to assess. The assertion that differences in GPP are larger than differences in NEP, and that spatial patterns should be considered, is not supported by any quantitative assessment, and since there is no spatial map of NEP it's not possible to assess that part of the comparison. Finally, the assertion that the supposed similarity in NEP in contrast to differences in GPP are caused by similar N effects on the respiration terms is pure speculation, with no supporting analysis.
L175 This paragraph summarizes a few rank relationships, but offers no mechanistic explanation, other than an unsupported claim that considering more processes increases uncertainty.
Upon reflection of the comments of reviewer 1 we have removed this section and included the figure on GPP in the SI, as the stronger part of the paper is the +N and +CO2 experiments and that ought to be the focus. The reviewer is correct that the model NEP ought not to be normalised to 1901-1910. This part of the paper was written before the publication of the latest GCB paper (Friedlingstein et al. 2019), but since the publication of the final version of this paper we have decided that this figure is defunct and have removed it altogether.

L183 Consistency with what?
Clarified as “the internal consistency of this schematic”.

L190 One useful contribution could be to try to quantify these differences, by showing a few details of the implementation, together with differences indriving variables (NPP or ET) among models.
A “few details of the implementation” of the BNF are available in Table 1. The driving variables (NPP or ET) are only relevant to three of the five models and GPP is already available (in what was figure 1 and is now in the Supplementary Information). The papers by Wieder et al. (2015) and Meyerholt et al. (2016) that we cite already offer the useful contribution of quantifying the differences between different BNF representations.

L197 So, are you suggesting that the comparison to observed values would be even worse if the observed values had not been corrected for root respiration? I'm not clear on the meaning of "however" in this sentence.
We appreciate the reviewer drawing our attention to this not being phrased as clearly as it ought to have been and have adjusted the text accordingly.

L204 It isn't clear what the significane of being an "open" system is. The models are accumulating carbon on land in the period summarized in Fig 3. They are also accumulating N.
This point was not clear to the other reviewer either, so in the interests of focusing on the most important areas of the work we have removed this point.

L209 This comparison could be more informative if there was a table of C:N by model and pool, and the accompanying C totals py model and pool.
We have added the C:N ratios to figure 1 (previously figure 3).

L211 It is a stretch to make this statement based on the analyses presented. There are lots of additional sources of observations for both fluxes and pools.
This sentence has been removed.

L218 The cited study looked at woody species only. Are the model results being weighted somehow in this analysis to emphasize the influence of woody biomes?
We have replaced the comparison with Baig et al. (2015) with a comparison with Song et al. (2019). The latter is not vegetation type specific.

L226 It would be useful and interesting to examine why these differences occur. A mechanistic investigation of the differences in model structure or parameterization. Some relationship to nutrient cycling, perhaps?
At this point it is not clear why this occurs, and we have added a statement to that effect to the text.

L230 It could be interesting to contrast these models in terms of their development of N limitation over time. That would mean including some time series outputs.
This would indeed be very interesting, but we note that N limitation is an emergent outcome of the model simulation and not readily quantified by any modelled quantify that can be easily compared across models. Using reference simulations with a C-cycle only representative would also be a plausible mean for any one given model, however, also here the across model comparison would not be straight forward.

L238 Not clear what this means, or why it is significant. The spatial patterns for +CO2 and +N for JSBACH seem quite different. Is that a dichotomy?
We have added a new figure and results section that develops this point. You will see from that figure that while the effect in JSBACH is less pronounced, the homogeneous distribution of +CO2 and +N results that "quite different" spatial patterns would give, is absent.

L243 It would be better to have a table where the results by model are summarized for different biomes. Then the last sentence of this paragraph would have some quantitative basis. As it is, several of the models appear to have high +N effects in grassland regions, but that is just my crude analysis "by eye" of the results shown in Fig 5. Surely you can do better, having access to all the model outputs.
A table of these values, by biome, has now been included in the SI.

L253 I am surprised that such a cursory sort of analysis is included as a result. If you want to assert a correlation, even a weak one, go ahead and plot the data or show us a regression result. I think you would find that the correlation is very weak indeed, in this case, if you made that effort.
This sentence has been removed.

L257 This statement is confusing. Looking at Fig 6, it looks to me like the CLM4.5 response for BNF and N balance are the smallest of all models, not middle of the range.
Apologies, this has been corrected to gaseous loss and leaching and uptake.

L260 So are you saying that this is just a difference in accounting? If so, can't that be corrected for in the analysis to make a more useful comparison?
Done.

L264 This is a good example of a process-based description of the differences between models.
We are glad you appreciate this.

L270 The assertions that N availability is primarily controlled by BNF, and that N demand is primarily driven by NUE are unsubstantiated here. One could just as easily say that availability is controlled more by mineralization rate, or by low levels of loss - without some quantitative assessment to back this up it is just a conjecture. Similarly, why not say that N demand is driven mainly by GPP? Why tie the causality statement for N demand to a term that has N uptake as its denominator?
This paragraph was intended to be a helpful introduction to the section on BNF and NUE, giving an overview of why we focus on those two metrics, without overburdening the reader with extensive background, references, and explanations as why these two metrics are the focus. Since this short paragraph is evidently causing confusion rather than smoothing the transition of topic, we have opted to remove it.

L285 It would be helpful to say something here about the observations: how representative do you expect they would be of a true global mean response?
There is a new section in the methods detailing the observations used and extensive discussion of the robustness (or not) of the observations in the discussion section. While we understand that the uncertainties about the observations are important, we prefer for the results section to primarily contain the results of our work.

L287 Saying it is counter-intuitive is different than saying it is counter-factual. Here is an example: in a shrub tundra community dominated by alder (an N fixer), there may still be significant N limitation. By

fertilizing the site, it is possible that the alder could expand its occupancy, with an overall increase in NPP. Assuming a constant NUE (for the sake of argument), a higher NPP would mean a higher N demand. It could well be that with higher N input and higher N demand, the community might still experience significant N limitation. In that case increase of BNF with increase in NPP under increased N input seems a plausible outcome.
The phrase "counter-intuitive" has been removed and the sentence rephrased. As the reviewer rightly says, there are a small number of situations where fixation might continue in the presence of increased N supply. The reviewer gives the example of alder, which is an obligate N fixer. However, N fixers can also be facultative, and thus able to slow or stop BNF where increased N supply makes it not energetically worthwhile (Menge et al., 2009). In the figure caption we cite observations (Zheng et al., 2019) at the larger biome scale, that show BNF reduces with increased N deposition. We have added a citation to the main text so the reader can more easily see the evidence that it is "counter-factual" that at the macro scale models increase BNF with increased N supply.

L291 It is in line the current theory of Walker et al., 2015.
Changed as suggested.

L293 Again, couldn't you address this in the analysis to make for a more internally consistent comparison?
We have added BNF to Nup for CLM5 as suggested. In the process of amending the code we found a small mistake in the JULES-ES results which we have now corrected. This has prompted us to review all the other figures and we found no further issues.

L299 It would be better to state this as a hypothesis, and provide an explanation for why you think this is the likely behavior of real ecosystems.
As suggested, we have rephrased this as a hypothesis and enlarged on this point.

L302 These results are your most interesting material.
Thank you for this considerate comment. Upon reflection we agree and have increased this section correspondingly and removed some other extraneous material.

L330 Has anyone tried to disprove this? It seems like the kind of model-generated hypothesis that would be testable in laboratory conditions using pre-industrial CO2 concentrations and N fertilization. Some actual discussion of this would be interesting.
So far as we know this is an ongoing debate in the N modelling community. We have added to the discussion of this point by providing each side of the discussion.

L345 You should at least try to connect the sign of the supposed bias (overestimation of NPP response to +N) to the factors you've listed. Otherwise, you seem to be saying that the response is overestimated because the system is complex. Why not underestimated?
We have rephrased this paragraph to enhance ease of comprehension.

L356 Is the NUE increase coming from shifting allocation, or shifting C:N? and is this a statement of what the models are doing, or how the real ecosystems respond? Couldn't +N also cause an increase in production without increasing LAI, for example through more fine root allocation?
This sentence has been rephrased to aid comprehension.

L381 I agree.

We’re glad this statement meets with your approval.

References

Menge, D. N. L., Levin, S. A. and Hedin, L. O.: Facultative versus Obligate Nitrogen Fixation Strategies and Their Ecosystem Consequences., The American Naturalist, 174(4), 465–477, doi:10.1086/605377, 2009.

Zheng, M., Zhou, Z., Luo, Y., Zhao, P. and Mo, J.: Global pattern and controls of biological nitrogen fixation under nutrient enrichment: A meta-analysis, Global Change Biology, 25(9), 3018–3030, doi:10.1111/gcb.14705, 2019.