I get the awkward position of being a new reviewer introduced to a paper mid-review. Like being a step-parent trying to balance my way of doing things with the fact that the kids (authors) have already developed in another system.

Overall, this is a great paper. It is very challenging to take on a paper that not only deals with a lot of complexities and nuances within the models and observations, but also the fact that one can present the analyses/results in a gazillion different ways, making it hard for readers to absorb. The authors did an excellent job of distilling analyses, results, and interpretations, which make this paper a valuable contribution to the literature.

The biggest challenge is probably benchmarking N cycle impacts against a lot of C cycle measurements. Moreover, the authors do a lot of comparing model outputs to observational ranges; but, we know very well (and the authors discuss briefly in the Discussion), these magnitudes change with choice of forcing data (and other model run conditions). So, then how useful is it to make these direct comparisons? Is there not a different/better way of doing these evaluations that accounts for the fact that the end number changes so easily? The sensitivities and directions should mostly be the same no matter what forcing. I don't expect the authors to change their results at this point out of sheer exhaustion/frustration/workload related to this comment. Still, hopefully a next paper can consider this comment to advance the types of analyses done. That said, the evaluations/analyses done in this paper are much better than what is often done in other papers (e.g., let's just compare to LAI and say the difference is due to the one component that I developed in the model…).

- Abstract
    - Somewhere say that you ran the models offline with common spin-up and forcing protocols—this is very valuable for understanding model differences.
    - L26-28. Maybe put something quantitative to complement the qualitative sentences, something readers can grab as take-home stats.
    - L29-31. It would be amazing to add why…
    - L31. "better represented" is vague/unclear.
    - L33. Throw away sentence. Delete.
    - L34. "better understanding and more provision" is vague/should be more explicit.

- Introduction
    - L41. "allowable" is that the right word? More like nothing or everything is allowed. Projections are just whatever scenarios ESMs are presented with.
    - L59-61. Break up this long sentence?
    - L60. It doesn't totally make sense why this study is limited to European centres. You commented on that in response to one of the reviewers, but it doesn't make sense in the paper. The abstract/title and everything else up to this point seems like the paper is generalizable across the global modeling community. But, then this gets inserted that throws the direction off with a jolt.
    - L62-63. Cite.

- Methods
  - L87. Any update to Wiltshire et al forthcoming? How about a conference abstract?

- Results
  - Fig 1. Cool figure. I wonder if there is more room for artistry in it so that one can visualize the numbers and spread without having to do the math in one's head individually for each component. Could be quite powerful if you can figure it out (it's already quite powerful though, so don't get me wrong).
    - The arrows for higher/lower than obs are nice. BUT, when you have no arrow it means either that it's within range, or that there are no obs. So, you've got some confusion there in symbology.
    - Is there no uncertainty on Ndep obs?
    - It's weird that Ndep differs between models, when they were all forced with the same amount. I guess you explain it with differences in land fraction, but it's still weird.
    - What about having all the obs be a number plus/minus a number. Instead of having some be ranges. Or vice versa.
    - The yellow is hard/impossible to read. Pretty much leaves me "guessing" on those numbers… (sorry, just fishing for a comment on my humor, given that you were giving out those compliments to other reviewers…).
    - Why no model numbers for Nmin, Nup, and Soil Ninorganic?
  - L164. Perhaps a slight bit more elaboration on CLM5's BNF could be useful, as it does seem to be quite different than the other 4 models.
    - Maybe include discussion of CLM4.5 here too, given that you discussed all the models but CLM4.5?
  - L184. Guess→GUESS. Actually, there's inconsistency on this throughout the paper, so just do a find and replace and pick one.
  - Section 3.2-3. Are the Song et al numbers comparable in terms of global scale, temporal scale, CO_2, and climate? It seems from Song et al's Fig 1, the data are mostly geographically not where the models are being impacted most at the global scale (e.g., low for JSBACH/JULES-ES, or high for CLM's, LPJ-GUESS). If they're not comparable, then don't compare them. Throw Song et al in to the Discussion or something saying about what would be needed to make them useful.
  - Figure 4. Maybe put somewhere on the figure that we're looking at NPP (in addition to the caption)? Would be good to have this figure stand alone.
    - Maybe make the dots bigger? E.g., it's hard to see JSBACH and JULES-ES.
    - Is this plus/minus latitude? Or just N. Hemisphere? If it's plus/minus, then that really isn't clear in the figure.
  - Figure 5. The red/purple areas are hard to distinguish from one another. Same goes with the orange/yellow, though that's easier as they're more distinct geometrically.
    - Why is there no left purple solid line?

        - I’d consider ditching the dashed line altogether. It’s really just extra information that isn’t even used because the models mostly get nowhere even near the bar areas. The reader can assume the middle point.
    - Figure 6. Cool figure. I’m confused in b and c though. They appear to be showing the N response. But, the text in L262-271 refers to the NPP response.
    - Figure 7. You introduce Fig 7d first, then 7b, and never 7a.
        - Not sure if the publishing editors will pick this up, but sometimes you have a period after Fig, other times not.
        - L280-282. I’m not following this text as it relates to the Figure. The text refers to Fig 7b. It says that JULES-ES is within range of the obs (except boreal). When I look at 7b, I don’t see JULES-ES’s bar inside the gray bars. Am I interpreting this incorrectly? Same goes with the statements on CLM5. You say that it’s a clear outlier with a large increase in BNF. But, 7b shows a large decrease, plus it’s kind of similar to LPJ-GUESS. There is an increase in 7a, but one could also just say that all the models are outliers relative to the obs, *except* for CLM5 in the boreal, which it actually hits.
        - I know CLM5 best mostly because I know FUN. So, this is a question specifically from J. Fisher to R. Fisher: how much of the CLM5 N response is due to issues with CLM’s C-cycle, i.e., too much GPP/NSC/not enough Rh? It’s great to see that CLM5 is going in the right directions etc., but it also looks like the N cycle is hyped up on sugar, like a kid on Halloween. If you cut that GPP down, then you have less C to pay for BNF etc.
    - L300-302. Grammar edit.

Good work overall! I hope my comments are useful.

Josh Fisher