General comments

In this study X. Yang and co-authors show global evaluation of ELMv1-CNP-model, which has fully prognostic carbon (C), nitrogen (N) and phosphorus (P) cycles. The model evaluation is done using the ILAMB benchmarking system, GOLUM-CNP data derived product and meta-analysis from elevated co2 manipulation experiments. Also the published values in literature are used in evaluation of simulated global C, N and P pools and fluxes. After the model has been proven to have good performance against evaluation metrics, influence of the P cycle on historical carbon balance is discussed. The main findings here are widespread co-limitation of N and P as well as a prominent influence of P on the historical C balance. The authors also go in detail to the causes of differences between GOLUM-CNP and model results and they also discuss the downsides of the benchmarking data used in ILAMB. They also cover the development needs that their model has.

The paper is well-written and figures are clear and demonstrative. The topic is wide and many of the global/biome scale evaluation results shown are not discussed in detail, which is justified, as a lot of things are covered. However, now in some occasions it's mentioned that the model is not doing so good in some respect e.g. in one biome and this fact is not re-visited later in the text. It would be interesting for other modellers working with these issues to hear more insight from the authors what they think the reasons are. But the model performance overall is good, so this is just a suggestion.

Overall the paper is of high quality and I recommend its publication in Biogeosciences after some comments below (mainly very minor) are addressed.

Response: We thank the reviewer for the positive comments and suggestions. We carefully addressed each comment as shown below and will revise our manuscript accordingly.

Specific comments

Model overview & simulations: It was not clear for me that the fire module was activated before the discussion. Maybe this could be mentioned here already. Few points that would also be interesting (also in the light of rest of the manuscript) to mention, what was the soil depth and if fixed stochiometry was used and if the leaves were the only pool where PFT-specific stochiometric ratios were used.

Response: The fire module is activated by default in ELMv1. Soil depth is 3.8m in ELMv1. Fixed stoichiometry is used for livewood, deadwood, and coarse root and fine root. Leaves are the only pool that has PFT-specific values. These have been described in Yang et al. (2019). We will provide brief descriptions about these and relevant references.

l. 299: Do any other of the models shown in Fig. 1 have CNP-cycles enabled?

Response: To our knowledge, none of the other models in Fig.1 have phosphorus cycle enabled.

l. 314: Is CNP-version always better than CN? If I'm not mistaken, the FLUXNET (for NEE, respiration & GPP) is better captured by CN-version? Also Precipitation/GPCP2?

Response: You are correct, the CNP-version is not always better than CN from the benchmarks in the current ILAMB system. One of the benefits of a multi-metric analysis package like ILAMB is that we can compare performance at different levels of granularity, and it is rare that any one model has uniformly improved performance over any other single model on every fine-grained metric. By having multiple data sources for a given metric we can often see improvement against one data source and degradation compared to another for the same model output. For example, the CN model performs better than CNP for ecosystem respiration when comparing the Fluxnet metric (as you point out), but CNP does better than CN for the GBAF metric on the same output variable. In the case of GPP and NEE the CN model is performing better or the same as CNP for both Fluxnet and GBAF metrics, which could point toward a concerning bias. If other related metrics all tended to favor the CN model, we would have less confidence. In this case, the overall better scores of the CNP model for the relationship metrics connected to GPP (tan shaded rows in Figure 1) give us more confidence that CNP is actually an improvement. Each metric has its own advantages and disadvantages, and there is still considerable subjectivity in how to interpret the multi-metric collection. For example, the site-level evaluations in iLAMB do not take into account site-specific disturbance histories, which can be an important driver of NEE variability over time at a given site. We will add text to the discussion to highlight these points.

l. 337: Are these annual mean LAI values or mean LAI values for the time when there are leaves?

Response: These are annual mean LAI values. We will clarify this in the revised manuscript.

l. 352: Are you referring here to both temperal and tropical grasslands or also tundra? The TEG seems to have high NUE and PUE values in the distribution. Why do you think that occurs?

Response: Both ELMv1 and GOLUM estimated NUE and PUE is higher in boreal and temperate forests and lower in tropical grassland and tundra. Temperate grassland NUE and PUE in ELMv1 are higher in distribution because of the higher variation in NPP allocation to non-structural carbon pools.

l. 394: Are you also simulating peatlands in your model?

Response: We don't have a peatland model in the current version of ELMv1. A peatland-enabled version of ELMv1 has been developed which captured peatland specific hydrology, plant function types and biogeochemistry (Ricciuto et al., 2020; Shi et al., 2015). This could help to explain why our soil carbon estimates are lower than some observation-based estimates. We will add related text to the discussion.

l. 397: Sorry, what was the value for your top 1 m soil carbon? (I found it it Table 2, but that had not been referenced yet here.)

Response: soil carbon to the 1m depth 1134.41 Pg C. We will add this value here in the revised manuscript.

l. 398: The estimation you refer to from Todd-Brown is originally from the HWSD? Would it be fair to mention also that source?

Response: Thanks for pointing this out. We will add the original reference (HWSD) in the revised version.

l. 401 & 403: Are the estimates from Pan (2011) really exactly the same for litter C and CWD?

Response: Thanks for pointing this out. Litter C is 43±3 Pg C, but coarse woody debris C stock should be 73±6 Pg C in Pan et al., 2011. We will correct it in the revised manuscript.

l. 419: Is the Xu and Prentice paper having two different estimates for vegetation N, the other one agreeing exactly with the Zaehle et al. estimate?

Response: 3.8Pg N is the estimate from Zaehle et al. (2010) and 5.3Pg N is the estimate from Xu and Prentice (2008). We will revise accordingly.

l. 452: In the caption of Fig. 9b you say that values close to 1 show co-limitation. Could you be more specific and say how close to one the values need to be that co-limitation is prominent?

Response: By plotting the ratio f_N/f_P in Figure 9b, we are able to show degrees of co-limitation that are relative to the mean limitation of the two nutrients. For example, for any given value of f_N, if we assume a value for f_P that is 10% higher then the ratio will be 0.91, whereas if we assume a value for f_P that is 10% lower the ratio will be 1.11. There is not a strict definition of co-limitation that would allow us to say that a specific deviation of this ratio away from 1 would no longer be considered a co-limiting condition. Instead, we suggest adding text to the caption indicating that the definition of co-limitation is subjective, and that "a difference of 10% or less between the values for f_N and f_P would lead to a range of about 0.9 to 1.1 in the plotted ratio".

l. 527-528: Should you clarify here, that the W-E gradient is referring to Amazon? Another point: I didn't find Quesada -paper in your references.

Response: Yes, we will clarify that the W-E gradient is referring to Amazon in the revised manuscript. We will add Quesada et. al.(2012) in the reference.

l. 589-590: Is this true also for tropical grasslands, or only tropical forests?

Response: This statement is true for tropical forest, but not for tropical grassland. We will revise the statement accordingly.

l. 593: Also for tundra?

Response: We agree for tundra the model is overestimating NPP. We will make it clear in the revised manuscript.

l. 596: Have you mentioned earlier, how you defined you stochiometric ratios? To my understanding they were PFT-specific, but that was only for the leaves?

Response: As shown in responses to earlier comments, stoichiometry for livewood, deadwood, and coarse root and fine root are fixed for all PFTs. Leaves are the only pool that has PFT-specific stoichiometric values. These have been described in Yang et al. (2019). We will provide brief descriptions about these and relevant references in the revised manuscript.

Fig. 1. You have some datasets in green boxes, some in orange boxes. Is there a difference between these datasets/variables?

Response: The datasets that are in green boxes are either pools or fluxes while the datasets in orange boxes are relationships between pools/fluxes and other environmental variables such as precipitation and temperature. We will add text to the figure caption to make this clear.

Fig. 7: Unclear, which is model and which is observation, since in the caption only circles are mentioned, but also triangles are shown. Since there are two observations for NSC (these should be also clearly denoted, which one is which), it's clear that the green triangles are from the model. The NPP vs. GPP response in the simulations shows quite similar response (or NPP response is larger), whereas in the observations the NPP response seems to be lower compared to GPP response. Would you like to comment on that? In Fig. 6b there is more pronounced co2 effect seen in NPP than GPP in Central Canada. Would you like to explain a bit more what is happening there?

Response: Mean observations are in open circles and mean model simulations are in green triangles. We will modify the figure captions to make it clear. There are two observations for the effect sizes of NSC, one is for sugar with a mean value of 1.3 and the other is for starch with mean value of 1.8. We will clarify this in the revised figure caption.

The increase of photosynthesis uptake under eCO2 and accompanying lower increase in NPP in the observations suggest that the additional uptake through photosynthesis is not necessarily translated into plant growth. This is mainly due to nutrient limitation. Our understanding of the fate of the additional fixed carbon is limited. In ELM, the additional carbon enters the non-structural carbon pool and then turnover at a specific rate. There are large uncertainties regarding this assumption and the parameter values for non-structural turnover. More observational data on the fate of carbon under eCO2 is needed to help resolve the discrepancies between observed and simulated GPP vs NPP responses to ECO2.

Fig 9b: In this case the values close to 1 are interesting, but unfortunately very similar in color to regions without vegetation. Whereas the extremes of the color scheme perhaps partly replicate the information already visible in 9a. Would it be possible to modify this figure to show the areas of co-limitation clearer?

Response: We agree that the current color scheme can be improved. We will revise the color scheme to show the difference between unvegetated regions and regions with co-limitation more clearly.

Fig. 11. The units are not now clear for me. They should be added. I'm pondering on the color scale for subplots b, d, and f. The below zero values show here the constraint caused by P to these variables, if I understood right. What is the unit on this effect? The color bars have exactly the same values for all the different variables, so has this effect been normalized? Why do the color bars stretch to the positive side? Are there any over zero values in these plots and if there are, how are those to be interpreted?

Response: We will label units more clearly for each figure. Units for a, c and e are g C/m2/yr, KgC/m2, and Kg/m2 respectively. Figures b,d, and f are unitless and have the same color bar values because they are normalized and calculated as percentage deviation between CNP and CN configurations. We will label all the units more clearly in the revised manuscript. As the reviewer mentioned, the negative values indicate the constraint caused by P. There are some grid cells that have the positive values in figure b, d, and f, which indicate that including P dynamics led to higher values of NPP, vegetation carbon, and soil carbon in 2001-2010. This could be due to N-P interactions that leads to overall less nutrient limitation on plant productivity.

Technical comments, typos

-co2 missing subscript in several places

Response: We will have them corrected in the revised manuscript.

l. 442: Missing unit here.

Response: We will add units here in the revised manuscript.

l. 637: eCO2 has not been introduced.

Response: We will add the definition in the revised manuscript.

Fig. 1. Caption, typo: JSBACH and MPI-ESM.

Response: Thanks. We will have it corrected in the revised manuscript.

Fig 5: denote subplots

Response: We will denote subplots in the revised manuscript.

Fig. 8: Could you explain in the caption the acronyms for the heterotrophic and autotrophic respirations?

Response: We will explain in the caption the acronyms for the heterotrophic and autotrophic respirations.

Fig. 10. Could you add the units?

Response: We will add units in Fig. 10.

Table 1: Typo in “transient” (LULCC column)

Response: Thanks. We will have it corrected in the revised manuscript.

Fig. S2: denote subplots

Response: We will denote subplots in the revised manuscript.

Fig. S7: Remove the subplot mark. (If ‘b)’ stands here for that...)

Response: It will be removed in the revised manuscript.

Fig. S8: Are the subplots marked? In S8a the highest latitude point for model is not seen.

Response: We will add labels for each subplots in the revised manuscript.

-It seems that S8 & S9 are referenced to before S7. Could the order of the plots be swapped?

Response: We will change the order of S7 and S8&S9 as suggested.

Fig. S9: Could you clearly denote the lat, lon -values for the sites? What are the grey boxes the cycle-plots? Denote the subplots. Include units. Replace 'var' with the variable name.

Response: We will make the lat and lon information more clearly defined in the revised manuscript.

Table S1: It's a bit mystery for me, why this table is called PFT-specific parameters, but only leaf parameters seem to be changing between PFTs...

Response: our thinking was to provide all parameters in the consistent format. Since leaf parameters does vary for different PFT, we use the same format for other parameters.

Table S4: typo 'Global net ecosystem'

Response: We will have it corrected in the revised manuscript.

Supplement – references: Richardson paper not in alphabetical order.

Response: We will have the references in the right order in the revised manuscript.