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. Please see Page 7, lines 205-210.

" In this version of the model, the fire module is activated by default. Erosion module is not activated. We assume soil C, N, and P cycling can take place to the 3.8m depth as the assumption in CLM4.5 (Koven et al., 2013). We also provide the key model parameters in Table S1 (PFT specific) and Table S2 (soil order specific). We note that only leaf parameters vary with PFT, but we include all other tissues in Table S1 to provide all parameters in the consistent format."

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. Please see page 20, lines 598-613.

"ELMv1-CNP is not always better than ELMv1-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 ELMv1-CN model performs better than ELMv1-CNP for ecosystem respiration when comparing the Fluxnet metric, but ELMv1-CNP does better than ELMv1-CN for the GBAF metric on the same output variable. In the case of GPP and NEE, although ELMv1-CN is performing better or the same as ELMv1-CNP for both Fluxnet and GBAF metrics, the overall better scores of the ELMv1-CNP model for the relationship metrics connected to GPP give us more confidence that ELMv1-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."

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. Please see page 12, line 359:

"……..between GPP and precipitation and the relationship between annual mean LAI and precipitation"

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. Please see page 13, lines 374-378.

"ELMv1-CNP simulated NUE is higher in temperate and boreal forests and lower in tropical grassland and tundra, which is consistent with GOLUM-CNP (Fig. 5a). Temperate grassland NUE and PUE in ELMv1-CNP 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. Please see page 14, lines 424-425:

"…….which could be due to the reason that ELMv1-CNP still yet to include an explicit representation of peatland carbon dynamics."

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. Please see page 14, line 426:

"...model simulated values of 1134.41 Pg C are within….."

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. Please see page 14, lines 426-427:

"estimate from the Harmonized World Soil Database (HWSD) (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) as reported"

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. Please see page 15, line 434:

"....and 73±6 Pg C (Pan et al., 2011)."

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. Please see page 15, line 450.

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". Please see page 49, lines 1441-1443:

"Definition of colimitation is subjective here, but 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. Please see page 19, lines 582-583:

"decreasing west-east gradient in productivity is mostly related to total soil P across the Amazon basin."

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. Please see page 22, lines 659-660:

"...lower N uptake in the tropical forests,…"

l. 593: Also for tundra?

Response: We agree for tundra the model is overestimating NPP. We will make it clear in the revised manuscript. Please see page 22, lines 663-664:

“...matches well with NPP from GOLUM-CNP except for Tundra…”

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. Please see page 7, lines 207-210:

“We also provide the key model parameters in Table S1 (PFT specific) and Table S2 (soil order specific). We note that only leaf parameters vary with PFT, but we include all other tissues in Table S1 to provide all parameters in the consistent format.”

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. Please see page 41, lines 1343-1345:

“. The datasets that are in green boxes are either carbon pools or fluxes while the datasets in orange boxes are relationships between carbon pools/fluxes and environmental variables such as precipitation or temperature.”

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. Please see page 47, lines 1405-1406:

“There are two observations of NSC shown here, one is for sugar with a mean value of 1.3 and the other is for starch with a mean value of 1.8…”

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. We've added a few sentences in the revised manuscript regarding this point. Please see page 23, lines 722-725:*

" However, large uncertainties remain regarding the turnover rate of the NSC pool. Further synthesis of field measurements on NSC in CO_2 enrichment experiments are needed to evaluate and constrain the representation of NSC in models."

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 9b has been revised to exclude grid cells with GPP less than 100gC/m2/yr.

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. All units in Fig 11 (now figure 10 in the revised manuscript) have been clearly labeled. Please see page 50, lines 1457-1461.

Technical comments, typos

-co2 missing subscript in several places

Response: We will have them corrected in the revised manuscript. All missing subscript has been corrected.

l. 442: Missing unit here.

Response: We will add units here in the revised manuscript. Please see page 16 line 475.

l. 637: eCO2 has not been introduced.

Response: We will add the definition in the revised manuscript. Please see page 23 line 719.

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

Response: Thanks. We will have it corrected in the revised manuscript. Please see page 42, line 1348.

Fig 5: denote subplots

Response: We will denote subplots in the revised manuscript. Please see page 45, Figure 5.

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. Please page 48, lines 1417-1418.

Fig. 10. Could you add the units?

Response: We will add units in Fig. 10 (now Fig. 11 in the revised manuscript). Please see page 51, line 1465.

Table 1: Typo in "transient" (LULCC column)

Response: Thanks. We will have it corrected in the revised manuscript. Please see page 52, Table 1.

Fig. S2: denote subplots

Response: We will denote subplots in the revised manuscript. The subplots are denoted. Please see the supplementary material.

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

Response: It will be removed in the revised manuscript. Please see supplementary material.

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. Subplots are added.

-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. The order of these plots has been changed.

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. Lat and lon are clearly defined for each site in the figure caption. Grey shading in the seasonal cycle plots (left) is meant to show the magnitude of seasonality. Grey shading in the time series plots (right) is to show the decades (e.g. 1990 -2000). Please see the revised supplementary material.

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. Corrected.

Supplement – references: Richardson paper not in alphabetical order.

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

Revision of “Global evaluation of ELMv1-CNP and the role of the phosphorus cycle in the historical terrestrial carbon balance” by Yang et al.

In this study, the authors evaluated the global application of the ELM-CNP model and used different data to evaluate model simulations. They compare the model performance against CN version as well as several models from CMIP6. Moreover, they compared their results against a data-driven model GOLUM-CNP. I am familiar with this model. Thus, it was interesting to see the global application of this model. While I appreciate the work, several points in the model codes, outputs, and manuscript need further clarification to make this work merit publication in GMD journal.

Response: We thank the reviewer for the time and efforts in reviewing our manuscript. We carefully addressed each comment as shown below and will revise our manuscript accordingly.

Model codes:
The simulation description states that the simulations were first spun up to bring C, N, and P pools to equilibrium. I believe this is not entirely correct. In your codes in PhosphorusStateUpdate3Mod.F90: you are ignoring the phosphorus pools during spinup but estimating only the fluxes. I see your comment in the codes that the rationale is not ending up with depleted pool size during the transient run, but how are you reaching a steady state from your spinup runs while ignoring the changes in pools?

Response: There are two versions of nutrient competition scheme in ELM, one is the relative demand scheme (RD) and the other one is the Equilibrium Chemistry Approximation (ECA). The default version of ELMv1 uses the RD nutrient scheme as described in this manuscript. The part of code that ignores the phosphorus pools during spinup only applies to ECA scheme (in the code: if ((nu_com .ne. 'RD') .and. ECA_Pconst_RGspin)). For the RD scheme, all the P pools are being updated during spinup.

Also, in SoilLittDecompMod, you introduced a ‘new’ C:P decomposition. Is this rate fixed across all soil types/biomes, or is it changing (similar to the plant stoichiometries)?
Other parameters are also fixed for other processes. For instance, in your ErosionMod the eroded phosphorus (pp2poc) is estimated using a fixed value from outdated reference (Meybeck (1982)) across all the soil types and biomes. What is the rationale behind this? And couldn’t you use the updated reference studies and different values per soil type?

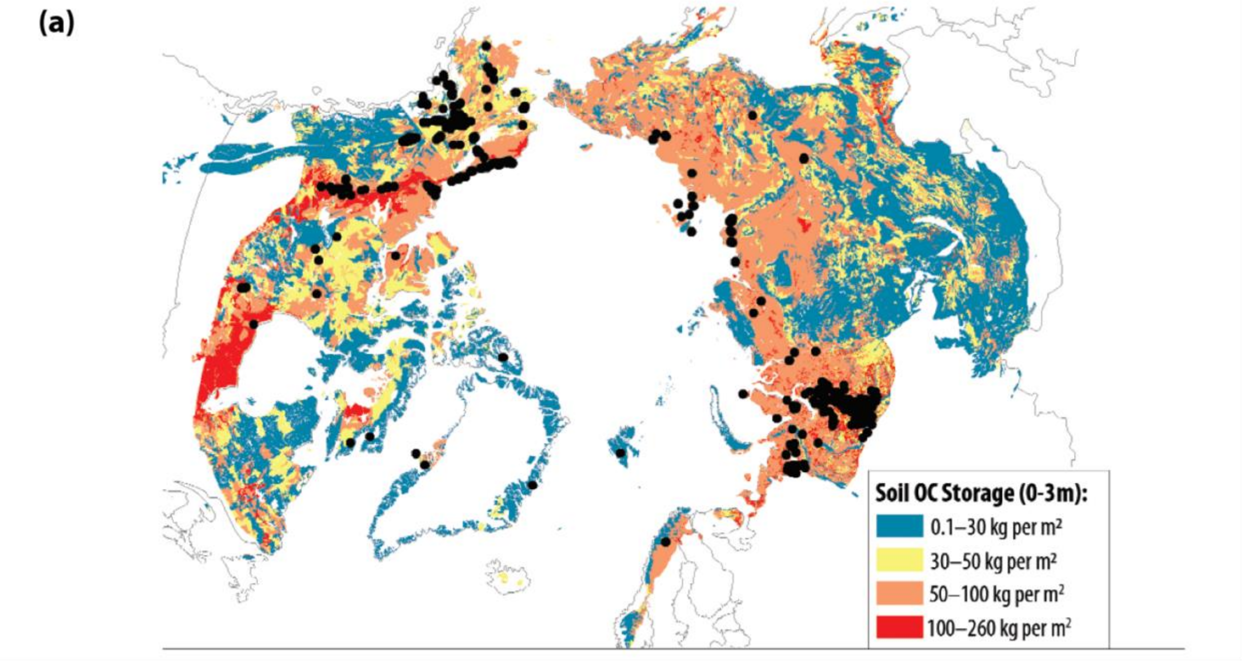
Response: The “new” C:P decomposition rate is fixed for all soil types/biomes. Erosion module is not active in the default ELMv1 model, therefore it is outside of the scope of this study. We will clarify this in the revised manuscript. Please see page 7, lines 205-206:

“The soil erosion module is not activated.”

Model outputs:
I processed one of your output files as an example. In your runs using the CNP model (ALL),

there are some extreme values (for instance, in total soil organic matter carbon (TOTSOMC) +240 kg C m-2) (the following figure). What is the reason for such unrealistic values in model outputs? Have you tried to detect these and understand the underlying issues? Were your final reported values excluding these extreme values?

Response: TOTSOMC is for total soil organic carbon for the whole soil column – up to 3.8m depth. There are a few grid cells in the arctic region that have very high soil carbon, as high as 240 kg C m-2. Total soil carbon to 3m depth in the permafrost region can be up to 260 kg C/m2 based on Schuur et al. (2015). Figure below is Figure 1a in Schuur et al. (2015) which shows the SOC in the interval of 0 to 3m depth of the northern circumpolar permafrost zone. The high soil carbon in these regions is mainly due to cold anoxic conditions. We did not exclude these high values in any of our analysis.



Model results:

Your model results show significantly underestimated leaching of P compared to Wang et al., 2018. Considering this underestimated leached fraction of soil P, your uptake is overestimated consequently (Table 2). Therefore, the available P for plant demand is overestimated as well. This could be the reason that despite the global application of P into the ELM model, still, you under/overestimate productivity similar to the pattern produced by the CN version of the model (figure 3) and overestimate the land sinking capacity (figure 2). This is in contrast with the objective of your study to quantify P limitation over land C sinking capacity. Moreover, in your code PhosphorusDynamicsMod (lines 334-439) you estimate the leaching only from sub-surface drainage flux. Do you have advection of soil P between soil layers? If not, why you do not estimate the leached fraction from each layer using the runoff/soil moisture (total water)?

Response: Our estimated leaching flux of P is lower than that of Wang et al., 2018, which could contribute to the underestimate of P uptake and overestimate of land carbon sink. We will add some statements in the discussion to discuss this uncertainty. Advection of soil mineral P and runoff was not included in this version of ELM, which could contribute to the lower P leaching. We are working on improving the representation of leaching fluxes through the ongoing efforts of coupling ELM with MOSART river biogeochemistry, which will be released in a future version of ELM. Please see page 22, lines 668-671 and page 29, lines 891-894:

". Soil P availability might be overestimated considering ELMv1-CNP estimated P leaching is much lower than the estimate of Wang et al. (2018), therefore leading to relatively higher P uptake in ELMv1-CNP. "

*"*ELMv1-CNP is likely underestimating P leaching, in comparison to the estimate of Wang et al. (2018), which could contribute to the underestimate of P uptake and overestimate of land carbon sink. We will further improve the representation of P leaching in ELMv1."

Other comments:
Line 137: Prior to this paragraph, give a brief explanation on what are the P cycle interaction with C-N components (for instance P availability impact plant productivity (Vicca et al., 2012; Wang et al., 2010) or NPP (Aragão et al., 2009)).

Response: Thanks for the suggestion. We will add a brief explanation as suggested. Please see page 5, lines 137-139:

*"*Field and modeling studies have shown that forest productivity tends to increase with increasing soil phosphorus availability (Vicca et al., 2012; Aragão et al., 2009; Wang et al., 2010)."

Line 206: I do not think this is correct. As explained in my comment on the model codes. Furthermore, it will be helpful to include the spinup results at the equilibrium in the supporting documents.

Response: As explained in response to the comments regarding model code, the part of code concerned was a special configuration (ECA) in ELMv1 and was not used in the default ELMv1 simulations as described in this study. The spinup results for all the P pools are included in the supplementary material (Fig. S1)

Line 219: Why by factor 10? Is there any reference for this value in accelerated spin-up for these pools? Did you test a range of factors to increase the turnover of this pool?

Response: The default mortality rate for the dead wood pools is 0.02/yr (a 50-year turnover time). We accelerate this mortality by a factor of 10, creating an effective turnover time of 5 years. The goal of *accelerated spinup is to obtain steady state pool sizes quickly but as close as possible to those when the model is run with default turnover times. Koven et al. (2013) found that when accelerating decomposition pools in the soil too quickly, there were strong effects on the seasonal cycle that affected the steady state values - thus, there is a trade-off between faster acceleration and the disequilibrium between accelerated and non-accelerated steady states that requires a longer "final" spinup. The same effect occurs when accelerating vegetation mortality, and we found that accelerating to a 5-year turnover gave us a good balance between these factors. Please see page 8, lines 226-229:*

“. The factor of 10 was chosen to have a good balance between faster acceleration and the disequilibrium between accelerated and non-accelerated steady states that requires a longer regular spinup following Koven et al. (2013).”

Line 223: How did you deal with the Gelisol, Histosol, and Andisol which were not included in Yang et al 2013 but included in this study?

Response: For the grid cells that don’t have values based on Yang et al.(2013), we applied the nearest neighbor extrapolation method to estimate the values. Please see page 8, lines 232-234:

*“*For the grid cells that don’t have values in Yang et al. (2013), we applied the nearest neighbor method to estimate the values.”

Line 224: The rationale behind using a developed P map for initialization is not clear to me. I believe that the model should be able to reproduce the P-related dynamics from bare to aged soil without using the initialized map.

Response: Soil P transformations occur on geological time scales. It is not realistic to run a land surface model like ELM for millions of years. An approach that appropriately estimates P status for model initialization is more efficient than modeling P processes at these timescales to arrive at present day conditions. More details about the rationale of developing P maps for model initialization can be found in Yang et al. (2013). We will add a few statements in the text to emphasize this point. Please see page 8, lines 234-238:

*“*Since the P cycle involves both biological and geochemical processes that occur on geological time scales, the initialization of P pools provides some reasonable estimates of soil P pools without running the model for millions of simulated years. More details regarding the rationale of using the developed P maps for initialization can be found in Yang et al. (2013). “

Line 226: What is the period for this spin-up?

Response: As stated in the manuscript, we ran the normal spinup for 600 years. Please see page 8, line 238-239.

Line 231: It is strange to see a very small variation in the labile pool. I understand that the occluded and parent material pools (due to very small rates used in the model) should not change much, but for the labile and adsorbed pools, it should not be the case. Is this because of the initialization of these pools using a global map?

Response: There are very small variations in the labile P because labile P is constantly interacting with other pools in the system. When the system reaches equilibrium, the inputs to and outputs from the labile P pool are balanced out so the pool size itself does not change much. Likewise, when the inputs to and outputs from adsorbed pools are balanced out during spinup, there are very little changes in the pool size itself.

Line 239: Name the environmental effects that you wanted to study, e.g. CO2/land use/climate impact or something else

Response: The environmental factors included in this study are CO2 forcing, land use and land cover change (LULCC), climate, and nitrogen deposition as summarized in Table 1. We will list the environmental factors in the revised manuscript. Please see page 9, lines 252-253:

"effects of changing environmental factors (atmospheric CO_2, land use and land cover change, climate, and nitrogen deposition, Table 1)."

Line 242: What was the rationale for bypassing the P limitation? Moreover, how did you prescribe enough P for each grid at each time step to exactly match the demand in the system?

Response: Here the idea is to run a simulation in which there is no P limitation on productivity and decomposition, in other words, the CN configuration. At each time step, we calculate the demand for P and the supply of P and supplement the difference between supply and demand. Please see page 9, lines 254-257:

"we also performed historical transient and single-factor simulations with P limitation switched off (supplementing P availability to fully meet demand at each grid cell and for each timestep so there is no P limitation on productivity and decomposition)."

Line 239-245: There is a repetition of the methodology here. If you have run with enough P that ignores the excess C as a result of P limitation, this is equivalent to the CN version run. My suggestion is to make these lines shorter and clearer.

Response: We will remove the repetition here as suggested. Please see page 9, lines 258-259.

Line 253: Firstly, this table can move to supporting document. Also, in most modeling papers in order to study different environmental factors' impact on the changes, there is one run with all the changes then the other factors attribute would be the run excluding it minus the run with all the changes. This way you keep the consistency between runs. What is the rationale behind your configuration with recycling all the other parameters except the study factor?

Response: We think it would be useful to have Table 1 in the main text. This table summarizes all the simulations in this study and the readers can reference back to the table when reading the results section.

Regarding the use of differencing to estimate single-factor effects in the model: There are two schools of thought on how to configure a multi-factor modeling experiment, and the Reviewer has described one while we prefer the other. The difference between them lies in how the interaction effects among the multiple factors are accounted for in the differencing of simulations. In our approach, we find the single factor effect by differencing a single factor experiment against a control with no factors varying. In that case there are no interaction effects among experimental factors mixed into the single-factor result. In the approach suggested by the Reviewer, where an all-factor experiment is differenced against an all-but-one-factor experiment, not only the single factor but also all of the two-way (or three-way, etc) interactions among factors are aliased into the single-factor result. Both are valid and useful as long as the differences between them are taken into consideration, but we prefer the simpler approach which avoids aliasing of the interaction effects, because it is easier to understand mechanisms related to differences. If needed, the total interaction effects are still quantifiable in the simpler approach as long as there is also a single all-factor run, as here.

2.3. ILAMB: These lines are too long and exhausting. You can summarize it in a few lines.
Line 277: which ones are CN/CNP models?

Response: We will shorten the description of ILAMB as suggested. ELMv1 CN/CNP models are not part of LS3MIP. Please see page 9-10, lines 271-288.

Line 284: The comparison against the steady-state model like GOLUM-CNP is not clear to me. If this is an intermodal comparison, firstly, it does not add any value to this study as the author stated as well as the uncertainty in equilibrium estimation by the GOLUM-CNP model (Wang et al., 2018). Secondly, if the intermodal comparison was the objective, why authors did not evaluate against a similar global process-based P-enabled model to the ELM-CNP such as ORCHIDEE (Sun et al., 2021)?

Response: For a nutrient-enabled model like ELM, it is important to also evaluate its performance on simulating nutrient pools and fluxes in addition to the evaluation of carbon pools and fluxes. The global land model benchmarking system ILAMB, however, does not include any observational dataset on nitrogen and phosphorus pools and fluxes. This is one of the major limitations of ILAMB and ongoing efforts are undertaken to address this limitation. GOLUM-CNP, unlike process-based models, provides estimates of present day nitrogen pools and fluxes by integrating observation-based estimates of C, N, and P pools and fluxes in terrestrial ecosystems into a diagnostic model framework. This observational-based dataset provided the next best thing for nutrient cycling evaluation. GOLUM-CNP has also been used in the evaluation of other land surface models (Sun et al., 2021). Please see page 10, lines 308-315:

"Despite large uncertainties and the steady-state assumptions, GOLUM-CNP provides a global data-driven product that can be used to test nutrient cycles in land surface models. GOLUM-CNP has also been used in the evaluation of other land surface models (Sun et al., 2021)."

Line 309: In IPSL-CM6A-LR, ORCHIDEE version 2 was used which did not include the P cycling. The comparison rationale is unclear.

Response: Results from several other land models from LS3MIP archive in CMIP6 are used to provide a context in terms of model performance. Although these models don't have an active P cycle, these model outputs are from LS3MIP offline simulations using the CMIP6 protocols, which is consistent with the simulations used in this study. Please see page 11, lines 321-322:

", along with several other land models in CMIP6, which are provided to contextualize ILAMB scores for ELMv1-CNP."

Line 316: Using your tool (https://compy-dtn.pnl.gov/yang954/_build/), as an example of selecting the tropic zones, your estimated RMSE score for C pools and fluxes using ELMCNP has not improved much compared to the ELM-CN. This needs further explanation.

Response: We agree that while ELM-CNP performs better overall in simulating carbon pools and fluxes, for some statistical metrics, ELM-CNP has not improved much compared to ELM-CN. It is challenging to show model improvements for all the metrics in ILAMB for a complex land surface model like ELM. It is important to note that ELMv1-CNP produces higher ILAMB scores for the integrated benchmarks such as global net ecosystem carbon balance and carbon dioxide concentration. These two integrated metrics are most critical to a land model in ESMs as they are most relevant to the coupling between land ecosystems and radiatively-forced climate change. Please see page 18, lines 597-612:

"ELMv1-CNP is not always better than ELMv1-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 ELMv1-CN model performs better than ELMv1-CNP for ecosystem respiration when comparing the Fluxnet metric, but ELMv1-CNP does better than ELMv1-CN for the GBAF metric on the same output variable. In the case of GPP and NEE, although ELMv1-CN is performing better or the same as ELMv1-CNP for both Fluxnet and GBAF metrics, the overall better scores of the ELMv1-CNP model for the relationship metrics connected to GPP give us more confidence that ELMv1-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."

Line 334: Instead of an extra graph you could just report here the values for CN vs CNP.

Response: We prefer to keep the graph as it includes more information.

Line 345: Yet your error in estimated LAI is higher than other models in these regions (https://compy-dtn.pnl.gov/yang954/_build/)

Response: We agree there are models which have better LAI estimates in these regions. ELMv1-CNP simulated LAI in these regions are better than ELMv1-CN simulated LAI, which is the point we are trying to make here.

Line 360: As you state one of your biggest mismatches is in TRF, with overestimated P uptake (Figure S2) resulting in underestimated PUE (Figure 5). In the discussion, you state that this is mainly due to different plant stoichiometries between ELM-CNP and GOLUMCNP (line 600). Did you test the model using the same leaf/wood/root C:P ratios from GOLUM-CNP to show this?

Response: We hypothesized that the higher estimates of P uptake in ELMv1 compared to GOLUMCNP was mainly due to the lower wood C:P ratio used in ELMv1. We followed the reviewer's suggestion and ran an ELMv1 simulation using wood C:P ratio from GOLUM-CNP for tropical forests but found that ELMv1-CNP estimated P uptake in tropical forests is still higher than GOLUM-CNP. As pointed out by the reviewer in earlier comments, P leaching in ELMv1 might be underestimated and therefore P availability is overestimated, which could lead to higher P uptake in ELMv1. We will revise the discussions accordingly. Please see page 22, lines 667-670:

". Soil P availability might be overestimated considering ELMv1-CNP estimated P leaching is much lower than the estimate of Wang et al. (2018), therefore leading to relatively higher P uptake in ELMv1-CNP."

Line 441: How does ELM-CNP differ this much from (Yang et al., 2013), when you use its map for your initialization?

Response: We only initialized soil inorganic P pools using maps from Yang et al., 2013 (lines 222-223). Soil organic P pools are from ad-spinup and then allowed to interact dynamically with vegetation and the initialized inorganic pools until all the pools reach dynamic equilibrium state.

Line 448-453: I suggest rewriting this part and instead of using "In many parts of the world", you report the relative N/P uptake in major biome classes.

Response: Thanks for the suggestion. We will revise the manuscript to report the relative NvsP limitation for major biomes. Please see page 16, lines 485-486:

“N and P are co-limiting productivity in tundra, boreal forests, and deserts.”

3.4. The effects of P limitation on the historical carbon cycle: Again, this whole paragraph is obscure. I suggest reporting the changes of P and C fluxes per major biomes, then the pools, and then. Reporting separately on environmental factors that impact these changes.
Additional note for figures: In some of the figures, units are missing. Please consider adding either on the plots or in the figure captions.

Response: One major component of this manuscript is to quantify the extent of P limitation on carbon pools and fluxes during historical time periods on the global scale. Therefore, in Fig 11, we show the ELMv1 estimated NPP, vegetation, biomass paired with the relative difference (in percentage) between CNP and CN. We will make sure the figure captions are detailed and easy to follow and units are provided for all the figures. Fig. 10 provides the estimates of cumulative global carbon fluxes due to each environmental factor and how those estimates are affected by including P cycling and therefore directly addresses one of the main questions in this manuscript. We feel like both figures are important to show and decide to keep both. We will switch the orders of Fig. 10 and Fig. 11 to make it easier to follow. Please see pages 16-17, lines 487-514. We will label units more clearly for each figure. Please see page 50, lines 1457-1461.

Reference:
Aragão, L. E. O. C., Malhi, Y., Metcalfe, D. B., Silva-Espejo, J. E., Jiménez, E., Navarrete, D., Almeida, S., Costa, A. C. L., Salinas, N., Phillips, O. L., Anderson, L. O. ., Baker, T. R., Goncalvez, P. H., Huamán-Ovalle, J., Mamani-Solórzano, M., Meir, P., Monteagudo, A., Peñuela, M. C., Prieto, A., Quesada, C. A., Rozas-Dávila, A., Rudas, A., Silva Junior, J. A., and Vásquez, R.: Above- and below-ground net primary productivity across ten Amazonian forests on contrasting soils, Biogeosciences Discuss., 6, 2441–2488, https://doi.org/10.5194/bgd-6-2441-2009, 2009.
Vicca, S., Luyssaert, S., Peñuelas, J., Campioli, M., Chapin, F. S., Ciais, P., Heinemeyer, A., Högberg, P., Kutsch, W. L., Law, B. E., Malhi, Y., Papale, D., Piao, S. L., Reichstein, M., Schulze, E. D., and Janssens, I. A.: Fertile forests produce biomass more efficiently, Ecol. Lett., 15, 520–526, https://doi.org/10.1111/j.1461-0248.2012.01775.x, 2012.
Wang, Y. P., Law, R. M., and Pak, B.: A global model of carbon, nitrogen and phosphorus cycles for the terrestrial biosphere, 7, 2261–2282, https://doi.org/10.5194/bg-7-2261-2010, 2010.
Yang, X., Post, W. M., Thornton, P. E., and Jain, A.: The distribution of soil phosphorus for global biogeochemical modeling, 10, 2525–2537, https://doi.org/10.5194/bg-10-2525-2013, 2013.
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, 7109–7131, https://doi.org/10.5194/bg-10-7109-2013, 2013.

Schuur, E., McGuire, A., Schädel, C. *et al.* Climate change and the permafrost carbon feedback. *Nature* **520**, 171–179 (2015). https://doi.org/10.1038/nature14338

Sun, Y., Goll, D. S., Chang, J., Ciais, P., Guenet, B., Helfenstein, J., Huang, Y., Lauerwald, R., Maignan, F., Naipal, V., Wang, Y., Yang, H., and Zhang, H.: Global evaluation of the nutrient-enabled version of the land surface model ORCHIDEE-CNP v1.2 (r5986), Geosci. Model Dev., 14, 1987–2010, https://doi.org/10.5194/gmd-14-1987-2021, 2021