This paper by Yu et al. describes a new algorithm to better represent soil phosphorus sorption dynamics in a terrestrial biosphere model – QUINCY. The authors proposed the use of a double-surface Langmuir isotherm to better capture the non-linear relationships between solution P and labile P pools in the soil. They performed a review on both published data and model assumptions on P sorption. They then compared their simulation against data at a range of P availability sites, performed sensitivity on parameters and compared simulations for CO2 and P enrichment scenarios. They argued that the double-surface Langmuir isotherm is a better modeling scheme because it simulated observed pattern of soil organic pools well, it maintained a relatively stable solution P pool to act as a buffer against instability, which then led to less P limitation at the P-poor site, and it led to improved simulation of folia N and P concentration.

Overall, this is a clearly-written manuscript. The rationale and objectives are crystal clear. The discussion is also well written. My comments mostly focus on two aspects of the results that I want to discuss with the authors and receive their clarifications:

- Did it indeed lead to improved estimate relative to the conventional single surface approach? All models performed well for reproducing the measured SOC etc. as reported in figure 3. The novelty of the double-surface scheme, as the authors argued, is that it better reproduced the ratio between Plab and exchangeable Pi (L190-191; Table 3). But looking at Table 3, the statistical significance is relatively weak (p =
0.014 for lab-to-exchangeable P ratio, and 0.044 for SIP). At the same time, I wonder if the new scheme actually increase model complexity or not. May be the authors should make a paragraph discussing whether the gained benefits in terms of improved simulation accuracy is worth the added complexity, if there’s indeed additional complexity associated with the new scheme. In particular, does it require additional parameters relative to the conventional approach? And, if we want to constrain the parameters in the new model scheme, what data collection should we make? If it doesn’t involve additional complexity, I think it’s very useful to highlight.

- What does it mean for the land C sink estimates under future rising CO2 if the model simulated a less P limitation at the P-limited site. As the authors introduced, there has been a lot of model development to add P-cycle into models. The relative magnitude of the P limitation is obviously different, but one of the crucial argument for the inclusion of P-cycle in models is that they would impose additional processes to constrain ecosystem productivity for P-poor regions of the world. The new scheme seemed to alleviate the extent of P limitation, and therefore I wonder how does it compare to a simulation without the P-cycle turned on. Do you obtain similar CO2 responses for the P-limited site? Obviously the CN-only simulation does not have the capacity to accurate reflect the processes limiting CO2 responses at the P-poor site, but it would be interesting to see if there’s indeed difference between the two approaches.

Lastly, one question I have, which isn’t a criticism per se, is that why the authors didn’t use the Jena Soil Model (that they developed) to investigate the effects of different soil P sorption functions. It appears to be a soil question and therefore I wonder is there any particular reason that deemed JSM unsuitable for this work?

Specific comments:

Title: replace P with phosphorus.
Figures: Figure presentation needs to improve quite significantly. Figures are currently in low quality resolution. Units and variable names in the figure legend/axis need to be properly labeled.

For Figure 3, the authors may need to think of a better way to show the results, as it’s not very clear to see differences in panel a and b (but maybe because they are similar and therefore it’s not important to show the differences?). But still, the 4pool and control color are too hard to differentiate from each other.

Table 3. One could argue the statistical significance is very weak.