Reviewer comments are in plain font. *Responses by the authors are in italics below each comment. Proposed changes to text in* "quoted plain font".

Referee #2

Simpson et al. present a rather technical but useful study on conversion equations between the two common soil P tests: Olsen P and Mehlich-3 P. The manuscript raises some interesting points and the conversion equations will surely be useful for future large-scale studies needing to harmonize soil P test data. However, in my opinion scientific depth (contextualization and discussion) and quality of writing should be improved in several places.

Authors: Thank you for your critical review and helpful comments.

Major comments

I recommend to rename the datasets for this paper. The names legacy P and ALP are confusing, because legacy P is of course its own concept, but not really relevant for this paper. What about something simple like: pan_USA and case study datasets, or similar? Of course in the methods you can make the link to the projects.

We see the reviewer's concern regarding 'legacy P' (the concept) vs. 'Legacy P' (the USDA project). We respectfully disagree that much (or any) confusion arises from the naming conventions as each is explicitly defined. 'Legacy P', always with upper case, is a defined project (full name: USDA Legacy Phosphorus Assessment Project). As the reviewer noted, the 'legacy P' concept is only mentioned in passing (L78, 512), always with lower case, and far apart from any specific mention of the 'Legacy P' dataset. Note that 'Legacy P' is very often followed by 'dataset' or 'site' or is used in some context where there's little potential for confusion. ALP is a commercial lab accreditation service who provided data for the study. We prefer names that clearly indicate the origin of the data and are relatively short. Based on this reasoning, we have retained our naming structure throughout the text.

Please start with a clear definition of labile P if that is going to be the word you use. Then be consistent in the wording.

Please see our response to Reviewer #1 who made a similar comment.

Authors do not mention the broader limitations of STP in general: extracting P with an alkaline or acid solution is not process-based. Indeed for local agronomic applications yield response relationships are quantified. But an Olsen value of 15 mg/kg has a very different environmental / agronomic meaning in one soil than another. This is touched on a bit in the discussion but not enough and it should come already earlier as well. See also crop response e.g. (Hirte et al., 2021; Morel et al., 2021).

We agree with the reviewer on the issues with STP. We struggled with delineating how far our Discussion should wade into that territory, but did refer to it often (e.g. opening sentence of 4.3). We also note that Steinfurth et al. (2022) *established that Olsen P of about 15 mg kg*$^{-1}$ *is a somewhat consistent threshold for crop response to P fertilizer across many areas. For environmental impacts: agreed, it does depend on context.*

We propose to briefly mention how agronomic soil tests are not process-based (process-inspired, perhaps) in the Introduction, likely around L37 in first paragraph. This may also be the best spot to introduce *what is* *process-based, alongside soil P 'lability'.*

Why are sediments included? I don't think M3 and Olsen P are used for sediments…

We discussed this internally early on. Pilot analyses showed that the distinction between soil and sediment was immaterial to the models and so it was simple enough to include them. (But note that there are of course differences between them, see e.g. Fig 2 in Simpson et al. 2025.) Some of the 'sediments' here are more like upland soils than typical streambed substrate (e.g., ditch bank material) so it's difficult to draw the line. We agree that agronomic tests have little meaning for sediments but, as for the key reason 'why', the inclusion of such samples broadens the biogeochemical characteristics to test/validate our models on.

Fig 1: This figure could be improved. Why are ALP sites not plotted as points? What do the colors mean for the legacy P sites? Why not make color or point size as a function of sample size?

We are happy to improve the map figure, where possible and when called for, but don't currently see how.

Answering the questions in turn: (1) Because these include private land and we do not, and cannot, have the specific coordinates; (2) the colors are explained in the legend; (3) we do make color a function of sample size for ALP but don't see the same need to do so for Legacy P soils (where sample size was fairly consistent between sites** and another legend item makes the map visually cluttered). *Note that, this being the first figure of the manuscript which serves to generally orient the reader with where the samples originate, we also use the opportunity to establish the color scale used for the rest of the figures. **The 7 sample sizes were 64, 84, 85, 91, 99, 100, 100 – so average of 89 and sd of 13, we'll add this summary stat to the caption.*

Data coverage: some areas are very sparsely covered, while the legacy P sites are overrepresented with up to 100 samples per site. How did you account for this?

In our models (L165-190), we included site/dataset as either a discrete 'fixed effect' ("region-level intercepts") or as a varying intercept in all the rest of the models. For the latter, this was done precisely for the reason the reviewer mentions: observations coming from the same study location (a single Legacy P *site) have some degree of statistical dependence. Varying effects (aka random effects) are a natural solution to, in a sense, down-weight these observations* (McElreath, 2020). *Note also that our k-fold CV and leave-one-site-out CV exercises were partly meant as stress tests for the impact of within-site dependence.*

If possible, we'll try to briefly allude to this in the methods.

Please explain all the soil measurement methods in table 1, e.g. by including a reference for each method. What is EPC0? Is pH measured in water or KCl? How is total P measured? Clarity is important here if you want other people to be able to make use of your equations. Also it needs to be reproducible.

Please see also our response to reviewer 1 on methods text. We will add more detail here, while keeping as brief as possible and directing readers to the open-access article by Simpson et al. 2025 for complete laboratory protocols.

Authors should better explain the potential use of the harmonization between different soil P tests, either in the introduction and or the beginning of the discussion. Another important application is validation data for large global crop models or input data for digital soil mapping. Authors may consider also mentioning application for large scale/global soil property homogenization. See e.g. (McDowell et al., 2026, 2023), where soil test data from many different STPs was used.

Agreed, we are glad for this comment. It was challenging to adequately cover all the contexts where STP conversions come into play.

We propose to highlight digital soil mapping of soil P in the Discussion as another context (though this is usually in service of a subsequent task, such as some sort of synthesis/estimate, global crop model, etc.). It may be best to refine the text in first paragraph of 4.1, and bring in related mentions in either the 'Agronomic utility' or 'Environmental utility' sections.

We did discuss briefly another one of the uses of STP conversions for a global soil P synthesis by Prof. McDowell et al. (L63) and could also mention the recent Olsen P map (McDowell et al. 2026).

Authors need to discuss: the "unified" conversion equations are calibrated on USA and Canada soils. What is the usefulness for the rest of the world?

Fair enough, more continents would certainly bolster the models; we'll suggest that future work attempt this. In the meantime, we can better argue the usefulness in the Discussion instead of just highlighting the limitation (L523) that our dataset did not include Gelisols or Oxisols. So, some regions, particularly tropical ones, should be cautious if applying our models. Volcanic soils may also be a challenge as their soil P chemistry is quite unique and we had limited Andisols.

Further, we propose to add "Toward" at the beginning of our title to explicitly indicate room for further improvement in our 'unified' conversion equations.

The discussion of the underlying mechanisms (section 4.2) needs major revision to reach a deeper level of understanding P dynamics. Writing, structure, and depth need to be improved. The discussed mechanisms make sense for the acidic test M3, but not at all for Olsen P which has a pH of 8.5. The fact that both of these tests are assumed to extract "plant available P" already becomes interesting by just comparing the pH values of the extractants. Sure the difference in time scale is there, but the pH difference is more important since we're extracting through completely different mechanisms.

First, we note that Reviewer #1 asked for cutting text here (citing specifically 4.2). We will do our best to find an appropriate median.

We are happy to refine section 4.2. However, the reviewer's request was not very specific regarding 'writing, structure, and depth' outside of a concern for pH. (We only mentioned pH effects – i.e. protonation – as needed and indeed laid out the primary effect here in the 2nd paragraph.) Also, bear in mind that we may discuss a particular mechanism for Mehlich-3 (e.g. protonation, chelation) <u>as a distinction from the Olsen test</u>, but don't see the need to call this out for the reader every time; we'll do so at least once (e.g. for protonation).

We propose to revise the last sentence of the first paragraph to better lay out the structure of this (rather dense) section for the reader: "We discuss here these aspects of soil P lability and related soil P chemistry to ground our statistical models, suggest potential for further improved conversions, and highlight the limitations of soil tests and their conversions."

The following structure (each paragraph's central point) is what we'll refine and clarify:

- *Soil tests are loosely based on plant mechanisms, one of them being protonation; the pH difference in soil tests as a result of this emphasis explains much of the disparities.*
- *Soil tests are inextricable from soil P lability (as we define it); much of the variance in the conversion ratio is explained by lability*
- R *increased with buffer capacity conditional on the soil P quantity, which is not readily understood; we hypothesize that this is due to greater buffer capacity implying more soil P beyond the 'labile' threshold in conjunction with different kinetic constraints in the soil tests*
- *Chelation is an important yet often overlooked mechanism for P release in such soil tests.*

And we'll edit for brevity and clarity to the best of our ability.

Interesting discussion on implications for fertilizer recommendations (section 4.3) with tangible examples.

We're glad for the feedback that this section resonates. Seems the added figures were worth it.

Section 4.4. could start a bit broader, since hydrological modeling is an important but not the only environmental use of STPs.

We're not 100% sure what the reviewer has in mind but we'll try to revise so the section is more broadly applicable. By 'environmental', we were referring to ecological impacts (i.e. eutrophication); for that, it's the hydrological transport of P that is the concern.

For one example to broaden 4.4: Perhaps we can mention large-scale soil P estimates relying on such conversions (e.g., McDowell et al. 2026), which one could use to make large-scale estimates of P resource inventory (McDowell et al., 2024; Smil, 2000). *This may also tie in with the crop modeling literature provided by the reviewer (e.g. Hirte et al. 2021).*

Minor comments

l. 42 authors may consider broadening here to include another recent use of STP: input for large-scale modelling applications

We will add references to large-scale datasets (e.g., McDowell) to provide better a better indication of how these equations can be used for these types of studies. We originally cited the Muntwyler et al. (2023) *modeling study and briefly the subsequent large-scale application* (Muntwyler et al., 2024)*, but could expand on that mention, as it's a great example.*

l. 74-76 This is a non-exhaustive listing of places where these tests are used. To make the manuscript more interesting for a global readership, please list also other world regions where the tests are used, not only the anglo-saxon countries.

Not sure what the reviewer means with "anglo-saxon countries"... we didn't even mention Great Britain! :^)

That aside, it is not a trivial task to determine, exhaustively, the regions globally where either Olsen or Mehlich-3 are used (unless the reviewer has such a reference?). Hence we cited locales we were familiar with and could find solid references for. We will expand here to a few more locales where we can find somewhat solid evidence and default to McDowell et al. (2026) *for a general indication otherwise, where the collated datasets suggested the preference for Olsen/Mehlich-3:*

"In this work, we focus on conversions between Mehlich-3 P (P_{M3}) and Olsen P (P_{Ols}), the two most frequent STP methods used agronomically in the US (Lyons et al., 2023), Canada (typically Olsen, e.g., Grant and Flaten, 2019), several European jurisdictions (Czech Republic, Denmark, Estonia, France, Italy, Slovakia, Spain, and the UK; Jordan-Meille et al., 2012), New Zealand (Olsen; McDowell et al., 2020), China (Olsen; Li et al., 2015), India (typically Olsen; Pathak et al., 2003), and likely in many countries in South America, Africa, and elsewhere (McDowell et al., 2026)."

l. 83 “P lability chemistry”? This seems like strange wording. Perhaps: soil properties governing P exchange kinetics? This is a valid and important hypothesis, but comes a bit out of no-where. I think a few more sentences on P chemistry and exchangeability are warranted in the introduction, linking to latest state of the art understanding of P species and their turnover.

Agreed; in line with other comments on our use of ‘lability’, we’ll revise this mention and better introduce the concept beforehand.

l. 93 Not clear what ~90 per site refers to.

We will shorten here to simply say “In total, 622 samples were collected (around 90 per site).”

l. 139 in R? or python?

No change here as the interested reader can find out from the reference. We will ensure a citation of R (R Core Team) toward the end of the methods.

l. 149 is “awkward” the right word? Seems pretty colloquial

We’ll be more formal with “inappropriate”.

Fig. 2 “dataset” is cutoff on the right side

Thanks, we’ll fix this

l. 250 “ may speak to” please use precise language
We are deliberately meaning that the useful predictors are suggesting, but not confirming, soil P buffer capacity and so want the wording to convey that meaning.

l. 307-308 “in comparison to region-level regressions” this is one specific application. But I think in global studies they even use one global equation. Please check and discuss also there the impact. The good news here is really that these basic soil properties that lead to such drastic improvements in prediction are usually readily available

We agree, the use of one ‘global’ equation in various works is what we’ve observed as well (and partly motivated this study). We suggest this addition at the end of 2nd paragraph in 4.1:
“Instead of using one or a few regression equations implying a fixed ratio between soil tests, we implore workers to incorporate information at the sample level.”

And we will try to repeat the mention of “readily-available variables” (L546) elsewhere (perhaps abstract).

And we suggest changing this line in Conclusions (L555):

"We advocate for preferring this method of converting soil P metrics to a common basis in works such as large-scale maps of soil P, bridging agronomic efforts between neighboring states/regions/countries, and meta-analyses or similar large-scale syntheses."

We go into some detail regarding the error involved in STP conversions, which we think the reader can propagate through an application of their interest. Our walk-through of the examples in Figures 5 and 6 should hopefully help as well.

l. 339 section title should be improved

Following from related discussion on 4.2, we propose a revised title: "4.2 Soil phosphorus chemistry guiding conversion equations and their limits"

l. 356 This sentence needs to be rewritten. No need to accuse Olsen and Mehlich of shortsightedness post-mortem ;)

We hold both Olsen and Mehlich in very high regard! To avoid giving the impression that we're speaking ill of the dead, we propose revising the sentence to: "While Olsen et al. (1954) and Mehlich (1984) gave less emphasis to the P lability dynamic in soils…"

l. 356 what is means with P lability dynamics? Please use precise language and aligned with state-of-the-art understanding of P cycling in soils.

We will better motivate the 'lability' concept before here as well as revise to avoid the potentially confusing 'lability dynamic' phrasing.

l. 362 I think here the word lability is used not for P exchangeability but for statistics? Please choose another word to avoid confusion

L362 has both "soil P lability" (not in quote) meaning the concept and "'lability'" (with 'quotes') meaning the particular model for conversion ratio. But we will be revising throughout to make clear what we mean with lability.

l. 366-367 Delete or improve this sentence

This sentence serves to set up the next paragraph. Continuity / logical cohesion would be lost without it.

l. 391 "were predictive"?

We suggest revising to "were similarly predictive" *to improve flow; we're saying that predictors related to chelation were another set of good predictors for the conversion ratio much like soil P quantity and buffer* *capacity predictors (discussed in previous paragraphs).*

l. 391 “power” is not the right word. Please use precise scientific language.

*Rephrased this sentence to: “*Unlike the Olsen extraction, Mehlich-3 contains strong chelating agents, acetate and EDTA.”

4.3 might want to mention large-scale / global P limitation modelling in crop models. For that you also often need to harmonize soil P test data from different tests.

Agreed, this section could touch on such crop models as well as general P resource/sustainability studies. We propose to discuss here some studies like (McDowell et al., 2024; Muntwyler et al., 2024; Ringeval et al., 2025) *to highlight how good, harmonized soil P data underpin these works.*

l. 472 “ germane”?
Yes, germane, meaning “most closely associated with, relevant”

l. 472-475. Why conditional tense? People are already working on this.

*Ah, we suggest adding “*reliably*” in front of “leverage”.*

l. 539 change “the success we found here” to precise language
*We propose changing to “*our own*” here.*

l. 545 I recommend to avoid abbreviations or re-introduce them in the conclusion. Some readers will only look at the conclusion.

Agreed, STP will be spelled out.

[R2’s References.] Note: references are to support my argumentation only. Not a recommendation of papers to cite.
We appreciate the suggestions, they’ve suggested directions for refining the text.

Hirte, J., Richner, W., Orth, B., Liebisch, F., Flisch, R., 2021. Yield response to soil test phosphorus in Switzerland: Pedoclimatic drivers of critical concentrations for optimal crop yields using multilevel modelling. Sci. Total Environ. 755, 143453. https://doi.org/10.1016/j.scitotenv.2020.143453
McDowell, R.W., Noble, A., Pletnyakov, P., Haygarth, P.M., 2023. A Global Database of Soil Plant Available Phosphorus. Sci. Data 10, 125. https://doi.org/10.1038/s41597-023-02022-4
McDowell, R.W., Noble, A.D.L., Doscher, C., Helfenstein, J., 2026. An improved global database and model of plant available soil phosphorus. Sci. Data 13, 880. https://doi.org/10.1038/s41597-026-07140-3
Morel, C., Plénet, D., Mollier, A., 2021. Calibration of maize phosphorus status by plant-available soil P assessed by common and process-based approaches. Is it soil-specific or not? Eur. J. Agron. 122, 126174. https://doi.org/10.1016/j.eja.2020.126174

Author references

Grant, C. A. and Flaten, D. N.: 4R Management of Phosphorus Fertilizer in the Northern Great Plains, J. Environ. Qual., 48, 1356–1369, https://doi.org/10.2134/jeq2019.02.0061, 2019.

Jordan-Meille, L., Rubæk, G. H., Ehlert, P. a. I., Genot, V., Hofman, G., Goulding, K., Recknagel, J., Provolo, G., and Barraclough, P.: An overview of fertilizer-P recommendations in Europe: soil testing, calibration and fertilizer recommendations, Soil Use Manag., 28, 419–435, https://doi.org/10.1111/j.1475-2743.2012.00453.x, 2012.

Li, H., Liu, J., Li, G., Shen, J., Bergström, L., and Zhang, F.: Past, present, and future use of phosphorus in Chinese agriculture and its influence on phosphorus losses, AMBIO, 44, 274–285, https://doi.org/10.1007/s13280-015-0633-0, 2015.

Lyons, S. E., Clark, J. D., Osmond, D. L., Parvej, M. R., Pearce, A. W., Slaton, N. A., and Spargo, J. T.: Current status of US soil test phosphorus and potassium recommendations and analytical methods, Soil Sci. Soc. Am. J., 87, 985–998, https://doi.org/10.1002/saj2.20536, 2023.

McDowell, R., Dodd, R., Pletnyakov, P., and Noble, A.: The Ability to Reduce Soil Legacy Phosphorus at a Country Scale, Front. Environ. Sci., 8, https://doi.org/10.3389/fenvs.2020.00006, 2020.

McDowell, R. W., Pletnyakov, P., and Haygarth, P. M.: Phosphorus applications adjusted to optimal crop yields can help sustain global phosphorus reserves, Nat. Food, 1–8, https://doi.org/10.1038/s43016-024-00952-9, 2024.

McDowell, R. W., Noble, A. D. L., Doscher, C., and Helfenstein, J.: An improved global database and model of plant available soil phosphorus, Sci. Data, https://doi.org/10.1038/s41597-026-07140-3, 2026.

McElreath, R.: Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd ed., Chapman and Hall/CRC, New York, 612 pp., https://doi.org/10.1201/9780429029608, 2020.

Muntwyler, A., Panagos, P., Morari, F., Berti, A., Jarosch, K. A., Mayer, J., and Lugato, E.: Modelling phosphorus dynamics in four European long-term experiments, Agric. Syst., 206, 103595, https://doi.org/10.1016/j.agsy.2022.103595, 2023.

Muntwyler, A., Panagos, P., Pfister, S., and Lugato, E.: Assessing the phosphorus cycle in European agricultural soils: Looking beyond current national phosphorus budgets, Sci. Total Environ., 906, 167143, https://doi.org/10.1016/j.scitotenv.2023.167143, 2024.

Pathak, H., Aggarwal, P. K., Roetter, R., Kalra, N., Bandyopadhaya, S. K., Prasad, S., and Van Keulen, H.: Modelling the quantitative evaluation of soil nutrient supply, nutrient use efficiency, and fertilizer requirements of wheat in India, Nutr. Cycl. Agroecosystems, 65, 105–113, https://doi.org/10.1023/A:1022177231332, 2003.

Ringeval, B., Demay, J., Helfenstein, J., Kvakić, M., Mollier, A., Seghouani, M., Nesme, T., Gerber, J. S., Mueller, N. D., and Pellerin, S.: Limitation of Maize Potential Yield by Phosphorus at the Global Scale, Glob. Change Biol., 31, e70485, https://doi.org/10.1111/gcb.70485, 2025.

Smil, V.: Phosphorus in the Environment: Natural Flows and Human Interferences, Annu. Rev. Energy Environ., 25, 53–88, https://doi.org/10.1146/annurev.energy.25.1.53, 2000.

Steinfurth, K., Börjesson, G., Denoroy, P., Eichler-Löbermann, B., Gans, W., Heyn, J., Hirte, J., Huyghebaert, B., Jouany, C., Koch, D., Merbach, I., Mokry, M., Mollier, A., Morel, C., Panten, K., Peiter, E., Poulton, P. R., Reitz, T., Rubæk, G. H., Spiegel, H., van Laak, M., von Tucher, S., and Buczko, U.: Thresholds of target phosphorus fertility classes in European fertilizer recommendations in relation to critical soil test phosphorus values derived from the analysis of 55 European long-term field experiments, Agric. Ecosyst. Environ., 332, 107926, https://doi.org/10.1016/j.agee.2022.107926, 2022.