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 #1

The authors propose in this manuscript a statistical approach allowing to convert Olsen P data (OlsP) into Mehlich 3 P (M3P) data and reciprocally, taking into account important soil properties at sample level. This is by itself a good idea as more and more scientists working on reviews and on models need such conversion functions to transform often heterogeneous databases on soil available P in a single usable indicator. Using data collected in two databases (Legacy P project and ALP) from North America the authors show that the ratio OlsP to M3P strongly differ among sites, so the use of a single ratio would lead to wrong conversions. They also show that this ratio can be successfully predicted when integrating appropriate soil properties into their models, the best being the GLM M3P model. This manuscript is interesting but still needs to undergo major revisions before it can be accepted.

Thank you for your critical review and helpful comments.

Major comments:

This paper implicitly makes the hypothesis that each soil will have a single OlsP or M3P value that will be independent from the laboratory where it will be analyzed. However, that might not be always the case. A study by Lischer and Neyroud (2003, journal of plant nutrition and soil science 166, 422-431) showed that OlsP measured on same 135 soils sent to various European laboratories could in average vary by a factor of 2. The conclusion of this paper was that if the protocols are not detailed enough and/or not strictly followed, then the risk that different labs using theoretically the same protocol would obtain different results would be important. Was this risk considered in this paper submitted to SOIL? How many laboratories were involved in total? Was at least a representative subset of samples sent to each of these labs and were the results compared? If this has been done for the data used in this paper, this needs to be said (the reviewer did not look at the Simpson et al 2025 paper mentioned at line 103). If it has not been done, the uncertainties linked to the laboratories need to be addressed explicitly in the discussion. The other soil analysis methods, their justification for this specific work, and the references where the protocols can be found should also be mentioned for the sake of transparency.

The reviewer raises a critical point: the measurement of STP is rife with error unless handled carefully. We originally omitted thorough discussion on measurement uncertainty except for a brief mention (L464) but were well aware of the issue. E.g., the cited study of Jordan-Meille et al. (2012) included a brief inter-lab comparison (two soils); most labs were within a few mg/kg but one lab differed drastically. We thank the reviewer for the Neyroud & Lischer (2003) reference, which is a great analysis (while worrisome for interpreting soil fertility laboratory results of that time) and which we will cite. [It's unfortunate the data are not plotted or made available, as it's difficult to interpret; e.g. Neyroud & Lischer say "Three laboratories *using the Olsen-P method were in very good agreement" yet certain lab-to-lab comparisons were as bad or worse than the reviewer points out.]*

Lab error was given due consideration here. For discussion, we describe below what that entailed for the two soils datasets. We propose revising Table 1 to include the method reference and related details (where called for), while directing the reader to complete information that is available in the open-access article by Simpson et al. 2025. Our revised methods text will reflect the lab QA/QC detailed below.

Before describing our own measurement error, we note that the reviewer's point speaks more broadly to soil testing as done in practice. That is, we are much less worried about measurement error in our own dataset (thanks to QA/QC, and the novel inter-lab data in ALP) than in the soil testing reports farmers, agronomists, and researchers regularly receive. In our revision, we have highlighted this more in Discussion as no conversion will improve on poorly measured STP.

Also, lack of method reference was an oversight on our part (L103) and we will now include the references in Table 1. Detailed laboratory protocols are easily gathered from the Simpson et al. 2025 paper.

For the Legacy P dataset, all soils were analyzed by one central, non-commercial lab (to reduce such errors the reviewer points out) which was well-equipped and rigorously operated – a deliberate project choice. Analyses of extraction solutions included calibration with matrix-matched solutions, method test blanks, calibration blank checks, and quality control standards. Aside from the analysis via ICP-OES (recall L108), the Olsen protocol resembles that of more-detailed protocols like Rubæk and Kristensen (2017)*, e.g., due care for soil:solution, standard lab temperature (20 C), immediate filtering, and <u>very careful</u> execution of the 30 minute extraction time. Likewise, Mehlich-3 analyses largely followed Mehlich* (1984) *with exceptions by Wolf and Beegle* (2011)*, notably the use of mass basis for the soil-to-solution ratio. Duplicate samples were included in each batch (ca. 30 samples), which had ≤1% replication error (whole method), with further replicate solution analyses for assessing instrument precision (within 0.25%). In addition, each batch included two to three soils from the North American Proficiency Testing (NAPT) program (https://www.soils.org/napt-program); these analyses were well within NAPT's interlaboratory median value ± 2.5 times the interlaboratory median absolute deviation. For the Legacy P dataset, we ensured the lab reproduced the consensus value for standardized soil samples derived by many independent labs by including NAPT samples in every batch.*

For the ALP dataset, these n=275 values are themselves the results of many (usually >30) accredited labs analyzing the same sample. The median (consensus) value of these labs is what's used in our study. We note that the inter-lab error in ALP resembled the Horwitz function, whereby the inter-lab median absolute deviation (MAD) relative to the median concentration, expressed as %, <u>declines with the median concentration</u>, which is a common phenomenon for inter-lab error (Hanley, 2016; Horwitz, 1982)*. For example with Olsen P, this uncertainty is ~30% at Olsen P of ~8 mg kg^{-1} (i.e. MAD = ~2 mg kg^{-1}), but declines to ~16% by ~30 mg kg^{-1} (R.O. Miller, pers. comm.). To give a further idea, the plot below shows the results from 7 soils spanning an order of magnitude in STP concentration.*

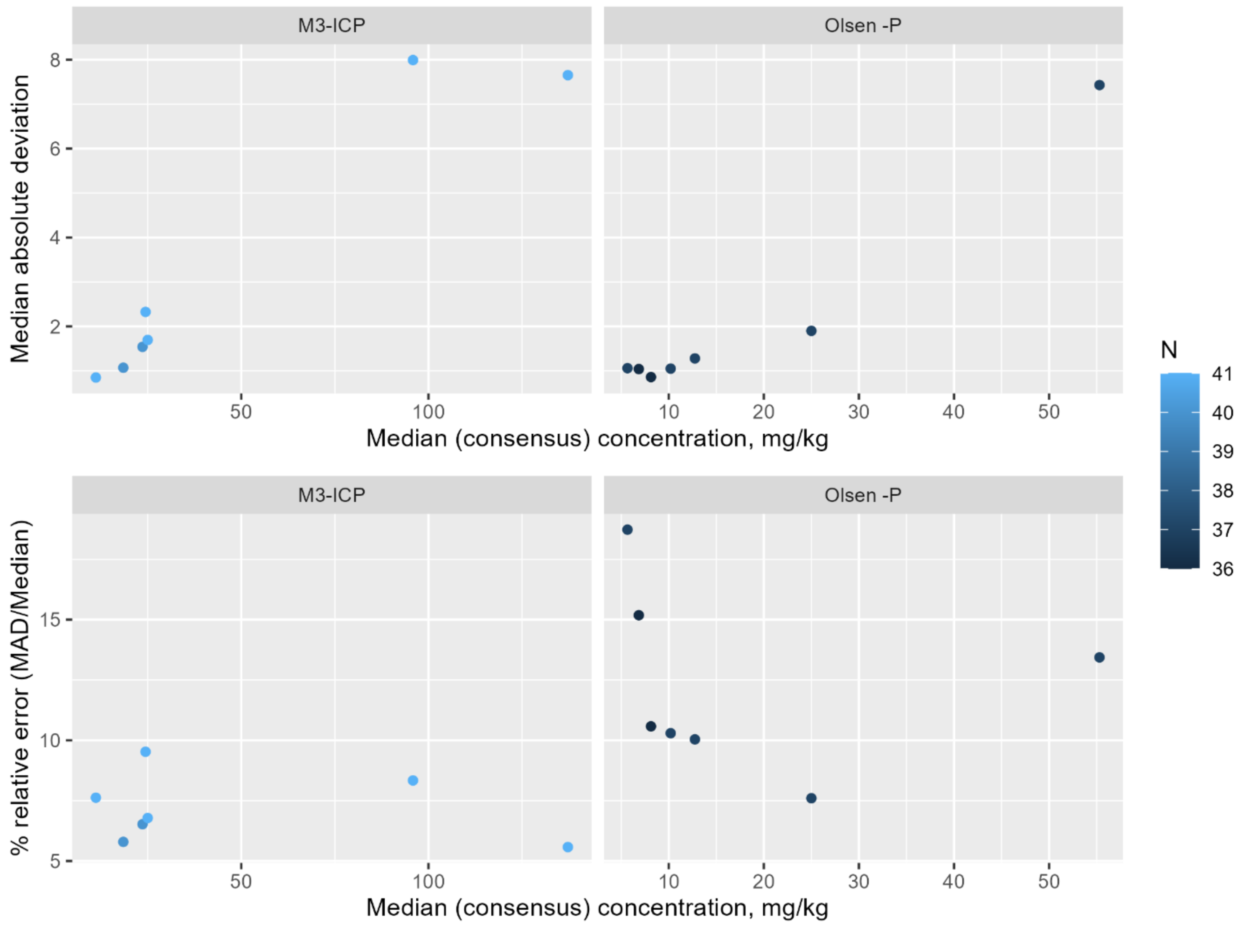

Figure 1. Select results from ALP inter-laboratory comparison. N is the number of reporting labs. The top row shows inter-lab variation as the MAD while the bottom row expresses it as a coefficient of variation (relative %). Results for either soil test are the same 7 soils.

We propose to distil (greatly!) the above details into brief mentions in sections 2.2 and 2.3, to convey the robustness of the datasets.

The authors mention often the concept of intensity, capacity and quantity. I reckon that it is an old albeit useful concept, but if this is important to the authors, these terms should be clearly defined and related to the soil P parameters considered here. Furthermore, the authors use very often the terms labile and lability. These terms are often used in literature but unfortunately with unclear and variable meanings. I suggest that these terms are removed from the text and replaced by well defined terms.

Agreed that we should clarify our terms. Our revised text will be much more careful here. Below, for discussion only, we expand on our language and the thinking behind it. (Our revised text will succinctly state definitions early on.) Unfortunately, unlike for related terminology (Haygarth and Sharpley, 2000)*, our literature lacks a good, single reference for certain terms (e.g. 'soil labile P'), hence the discussion.*

First, by soil 'labile P' we mean: phosphate ions complexed by the soil surface which may readily exchange with phosphate ions in solution (Beckett and White, 1964; Larsen, 1967). *"Readily" must also be defined by choosing a timescale* (Larsen, 1967)*; following Sharpley et al.* (1984)*, we consider 24 hours as the appropriate cutoff. Note we measured this quantity with anion exchange membranes (a strong P sink) but this operation is unambiguous. We could refer to the measurement (not the concept) as AEM-*$P_{24\text{-}hour}$ *or similar to convey the operational definition (e.g. like 'Olsen P'). But, if we may, we'd prefer to use 'labile P' throughout for the measurement as well as the concept.*

Now, for 'quantity', 'intensity', and '(buffer) capacity', these are abstract nouns referring to, respectively: the mass quantity of 'labile' P complexed on the soil surface; the concentration in solution dictated by this quantity of soil labile P; and the relationship between the 'quantity' and 'intensity' (dQ/dI) (see also Bache and Williams, 1971). *All three aspects describe the ability of soil P to exchange with solution. This dynamic refers primarily to phosphate ions but certain organic P compounds can also be involved in such sorption-desorption reactions. The above derives from seminal works by Schofield* (1955) *(who started with questioning 'availability'), Beckett and White* (Beckett and White, 1964; White and Beckett, 1964)*, various workers studying soil P with ion exchange kinetics* (Fardeau, 1995)*, and many other scientists who came before* (Pierzynski et al., 2005 and references therein).

To more succinctly capture all three aspects of quantity-intensity-capacity, we took the liberty of using yet another abstract noun: 'lability'. (This was also a much-deliberated decision for Simpson et al. 2025.) Lability hence describes 'the ability of soil P to exchange with solution'. We recognize that this is a less commonly used term, and so will endeavor to establish it clearly in the revised text.

The discussion is too long (from line 300 to line 543 on a paper that has in total 562 lines). I suggest that the authors strongly decrease it. For instance, section 4.2 which attempts to use the results of the statistical analyses presented in this paper for a mechanistic discussion on soil inorganic P availability is really too long and do not bring any new information compared eg to the works of Barrow, Fardeau, Morel, van der Zee...

We have reviewed, thoroughly considered, and revised this section seeking an appropriate median as Reviewer 2 requested further and more detailed discussion.

Regarding section 4.2, while we respectfully disagree with the reviewer, we hope our explanation and revisions clearly convey our intent: Our goal was not to use the stats to discuss soil inorganic P mechanisms, but rather, we sought to use the soil inorganic P mechanisms (established by the mentioned authors and others) <u>*to discuss our stats*</u>*. We believe statistical models like those we present, while still inherently wrong, are more robust and credible when grounded in process-based understanding (i.e. no black box, attempts to be causally consistent). We find some aspects of the models quite difficult to understand, even confusing. Rather than infer anything new about soil inorganic P availability, the Discussion in 4.2 is meant to give the predictive models a strong degree of explainability, unlike that seen for highly-predictive yet completely black-box models (i.e., much of modern machine learning). The transparent methodology we employed here* *could be developed further for similar cases (e.g., more STP conversion models) and so part of the motivation was to facilitate that work.*

Still, we take the reviewer's broader point and have trimmed throughout the Discussion where possible.

Specific comments

Line 83, what is P lability chemistry?

We will revise this mention in line with our discussion above on 'lability', including establishing the concept in the Intro.

Line 213, Le Suer is to be corrected to Le Sueur.

Thanks for this catch

Line 304, what is a fertility recommendation?

We will revise to 'nutrient management recommendations' to be more accurate

Line 356, what is P lability dynamics?

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

Line 408, what is a crop correlation?

The phrase 'crop correlation and calibration trials' is a mouthful and does read oddly to those less conversant in agronomy; this phrasing is rather common in fertilizer trial/rec literature. However, we will revise to "…local crop fertilizer response trials" to ensure clarity for the broadest audience.

Author references

Bache, B. and Williams, E.: A phosphate sorption index for soil, Eur. J. Soil Sci., 22, 289–301, https://doi.org/10.1111/j.1365-2389.1971.tb01617.x, 1971.

Beckett, P. H. T. and White, R. E.: Studies on the phosphate potentials of soils - Part III: The pool of labile inorganic phosphate, Plant Soil, 21, 253–282, https://doi.org/10.1007/BF01377744, 1964.

Fardeau, J. C.: Dynamics of phosphate in soils. An isotopic outlook, Fertil. Res., 45, 91–100, https://doi.org/10.1007/BF00790658, 1995.

Hanley, Q. S.: Chemical Measurement and Fluctuation Scaling, Anal. Chem., 88, 12036–12042, https://doi.org/10.1021/acs.analchem.6b02335, 2016.

Haygarth, P. M. and Sharpley, A. N.: Terminology for Phosphorus Transfer, J. Environ. Qual., 29, 10–15, https://doi.org/10.2134/jeq2000.00472425002900010002x, 2000.

Horwitz, W.: Evaluation of analytical methods used for regulation of foods and drugs, Anal. Chem., 54, 67–76, https://doi.org/10.1021/ac00238a002, 1982.

Larsen, S.: Soil Phosphorus, in: Advances in Agronomy, vol. 19, edited by: Norman, A. G., Academic Press, 151–210, https://doi.org/10.1016/S0065-2113(08)60735-X, 1967.

Mehlich, A.: Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant, Commun. Soil Sci. Plant Anal., 15, 1409–1416, https://doi.org/10.1080/00103628409367568, 1984.

Pierzynski, G. M., McDowell, R. W., and Thomas Sims, J.: Chemistry, Cycling, and Potential Movement of Inorganic Phosphorus in Soils, in: Phosphorus: Agriculture and the Environment, John Wiley & Sons, Ltd, 51–86, https://doi.org/10.2134/agronmonogr46.c3, 2005.

Rubæk, G. H. and Kristensen, K.: Protocol for bicarbonate extraction of inorganic phosphate from agricultural soils, Danish Centre for Food and Agriculture, Tjele, Denmark, 2017.

Schofield, R. K.: Can a precise meaning be given to available soil phosphorus?, Soils Fertil., 18, 373–375, 1955.

Sharpley, A. N., Jones, C. A., Gray, C., and Cole, C. V.: A Simplified Soil and Plant Phosphorus Model: II. Prediction of Labile, Organic, and Sorbed Phosphorus, Soil Sci. Soc. Am. J., 48, 805–809, https://doi.org/10.2136/sssaj1984.03615995004800040021x, 1984.

White, R. E. and Beckett, P. H. T.: Studies on the phosphate potentials of soils, Part I - The measurement of phosphate potential, Plant Soil, 20, 1–16, https://doi.org/10.1007/BF01378093, 1964.

Wolf, A. and Beegle, D. B.: Recommended Soil Tests for Macro and Micronutrients, in: Recommended Soil Testing Procedures For The Northeastern United States, NECC-1312, 2011.