Response to Reviewers' Comments

COMMENTS TO THE AUTHORS

Dear authors,

Thank you for submitting your manuscript to SOIL. This paper addresses an important and timely problem in digital soil mapping---the correction of biased prior soil property distributions using iterative, probabilistic residual correction---and the core methodological concept has genuine merit. However, after careful reading, I have identified a number of substantive concerns that must be addressed before the manuscript can be published. The most critical of these relate to the validation strategy, the interpretation of performance gains, the calibration of uncertainty estimates, and the depth of engagement with existing literature. I detail these concerns below and ask that the authors address these before we can move to publication.

We sincerely thank the executive editor (Prof. Raphael Viscarra Rossel), topic editor (Prof. David Rossiter), and three anonymous reviewers for their thorough reading of our manuscript and for providing detailed, constructive comments. The reviewer's insightful critique and suggestions have helped us to substantially improve the manuscript. Below, we provide a point-by-point response to the technical corrections raised. All changes in the revised manuscript are highlighted in *red text* for convenience.

Here is a summary for changed made in this iteration.

Comment	**Location in revised manuscript**
I-1: Research gap not explicit	Introduction, para 4-5
I-2: Novelty undifferentiated	Introduction, para 6
I-3: No structured aims	Introduction, para 6
M-1: Framework/implementation conflation	Sec. 2.1.2 and 2.2.2
M-2A: Random depth order unjustified	Sec. 2.1
M-2B: Convergence threshold inverted logic	Sec. 2.1.2 and 2.2.2
M-2C: No spatial cross-validation	Sec. 3.2, Sec. 4.4
M-2D: IRC inherits prior structural biases	Sec. 4.2, Fig. 10
M-2E: Compositional closure mechanics unclear	Sec. 2.1.3
M-3: OOB insufficient; field data not held out	Sec. 3.2, Sec. 4.2
R-1/R-2: OOB optimistic; possible overfitting	Sec. 3.2, Conclusion
R-3: No CI on RMSE differences	Fig. 7
R-4: Poor priors inflate improvement; single-pass comparison needed	Sec. 3.2, Sec. 3.4, Fig. 13
R-5: Same samples for prior/posterior comparison?	Sec. 3.1
R-6: Spurious SOM values unresolved	Sec. 3.2, Sec. 4.2

R-7: Semi-variogram analysis missing	Fig. 9, Sec. 3.2
R-8: Uncertainty calibration analysis absent	Fig. 12, Sec. 3.3
D-1: Discussion lacks hypothesis statement	Sec. 4.1
D-2: No engagement with literature in Discussion	Sec. 4.1, 4.5
D-3: Conclusion restates methods/results	Sec. 5

INTRODUCTION

Comment I-1m

The introduction should better articulate the research gap that this manuscript addresses. The gap is implied rather than explicitly stated. A reader must infer it from the accumulated list of limitations, rather than encountering a clear, declarative statement. The transition between the bias/uncertainty discussion and the description of the proposed method is abrupt. The gap as implied conflates several distinct problems. The claim that kriging-based post-processing fails in areas of high spatial heterogeneity is valid but is presented too briefly.

Response: We thank the reviewer for this constructive suggestion. In the revised manuscript, we have restructured the final two paragraphs of the Introduction. The third paragraph now explains the kriging limitation in terms of second-order stationarity assumptions and the resulting spatial non-stationarity artifacts ('bull's eye' artifacts around isolated observations), providing a more rigorous justification for a non-parametric alternative. A new dedicated gap paragraph (paragraph 5 of the Introduction) now opens with: *'Taken together, these limitations point to a research gap. Existing DSM products contain biases that are difficult to correct with current methods.'* It then introduces kriging stationarity constraints, QRF operating on numerical inputs only and being limited to single-pass simulation, and Bayesian updating requiring distributional assumptions, and closes with statement: *'There is a need for a framework that integrates legacy soil survey... and georeferenced pedons to update posterior distributions and improve soil mapping performance, without using synthetic sampling, distributional assumptions, or site-specific parameterization.'* This delineates the primary problem (the absence of a non-parametric, distribution-free framework for posterior updating that jointly leverages survey and pedon data) from secondary contributions (improved spatial heterogeneity representation and better-calibrated uncertainty).

Comment I-2

The proposed combination of pHRF with IRC is potentially innovative, but the novelty is poorly differentiated from prior work. The authors correctly acknowledge Sylvain et al. (2021) and Zhang & Lu (2012) as conceptual predecessors, which is commendable. However, the incremental contribution over these works is stated vaguely. Phrases like "extends these concepts by probabilistically updating posterior distributions" and "dynamically optimizing the feature space" are not sufficiently explained. A reviewer or reader cannot judge the degree

of innovation without understanding how these mechanisms differ mechanistically from existing iterative or ensemble correction approaches.

Three specific novelty claims deserve sharper articulation: (a) iterative residual correction until convergence; (b) layer-by-layer vertical correlation preservation; (c) non-parametric probabilistic posterior updating vs. QRF or Bayesian approaches.

Response: Thank you for giving us the constructive suggestions. We have revised the Introduction to address each of the three sub-points: **(1) Convergence:** The revised text now explicitly defines convergence as the stopping criterion is declared when residuals stabilise across iterations, yielding diminishing gains in predictive accuracy; the threshold is a customizable parameter. **(2) Layer-by-layer vertical correlation:** This is now stated explicitly as: *'after correcting each depth layer, its updated soil property values replace the original feature column for that layer when correcting the next layer down. This preserves inter-layer correlations while dynamically updating the feature space at each step.'* **(3) Non-parametric probabilistic updating:** The revised text contrasts this explicitly with QRF (which generates quantiles from decision tree outputs but cannot incorporate categorical soil survey information and is computationally prohibitive as an iterative method) and Bayesian updating (which requires specifying prior distributional forms and likelihood parameters). The IRC approach requires neither. It operates on the prior distribution which integrated soil taxa as input data and adjusts residuals non-parametrically using a Random Forest regressor.

Comment I-3

The aims are not explicitly stated. The final sentence is far too generic and could apply to virtually any DSM paper. There is no structured list of objectives, no defined scope, and no indication of how success will be evaluated. A well-structured aims section should state:

1. The specific objectives of the study

2. The hypotheses being tested (if applicable)

3. The geographic/dataset scope

4. How the study advances beyond the cited state-of-the-art

Response: Thank you for your insightful suggestions. We have replaced the closing sentence with a structured aims section (paragraph 6 of the Introduction) that lists four numbered objectives: *(1) To develop the IRC framework as a non-parametric method for updating posterior distributions of soil properties by iteratively correcting residuals between previous predictions and georeferenced soil pedons, without synthetic sampling or distributional assumptions; (2) To implement and evaluate IRC using six soil properties (sand, silt, clay, pH, bulk density, and SOM) across California, assessing predictive performance using RMSE and R2 and uncertainty quantification; (3) To assess the added value of iterative*

correction by comparing model performance relative to both the pHRF prior and a single-pass residual correction baseline; (4) To study the spatial structure of posterior predictions by comparing semi-variograms and predicted soil profiles from prior and posterior maps. The geographic scope (California, 1-km resolution) and the potential for CONUS extension and influence on the DSM community are also explicitly stated.

METHODS

Comment M-1

The methods section conflates framework description with implementation details in places, making it harder to distinguish what is universally applicable versus specific to the California case study. For instance, the convergence threshold (5th percentile of value changes) is introduced generically in Section 2.1.2 but only contextualised in Section 2.2.2, creating unnecessary back-and-forth for the reader. A clearer separation between the general framework and its specific instantiation would improve readability considerably. Some terminology is used inconsistently. "Top-probable values" and "representative values" are introduced without a formal probabilistic definition early enough, readers unfamiliar with the pHRF prior framework (Xu et al., 2025) will struggle to follow the feature space construction.

Response: Thank you for your suggestion. We have revised Section 2.1.2 to describe the convergence criterion at a general framework level: the stopping criterion is presented as a customizable parameter without specifying the California-specific threshold. The California implementation details (5th percentile of residual distribution changes; requiring convergence to hold across three consecutive iterations) are now presented in Section 2.2.2, which is clearly labelled as the case-study implementation. We have also added an earlier definition and equations of 'top-probable values' and ‘representative values’ in the worked example (Section 2.1.1) to ensure readers unfamiliar with the pHRF framework can follow the feature space construction without needing to consult Xu et al. (2025).

- *Representative soil property values (1 dimension): The expected value (weighted mean) of the soil property at each pixel in the modeling layer, representing the current best estimate. This is computed as the weighted sum of top-probable values.*
  Let $\{v_i\}_{i=1}^{k}$ *denote a set of candidate soil property values with associated normalized weights* $\{w_i\}_{i=1}^{k}$, *where* $\sum_{i=1}^{k} w_i = 1$. *Then a representative value* $\hat{v}$ *is computed as:*

$$\hat{v} = \sum_{i=1}^{k} w_i \, v_i$$

- *Top-probable values refer to current estimates of soil property values with top-k ranked highest weights, and k is 12 in this example.*

At the current iteration, candidate values are updated as the sum of predicted residuals and the prior estimates of soil property values. Let $\{v_i\}_{i=1}^{k}$ denote a set of candidate values based on their associated weights. Specifically, the k candidates with the highest weights w_i are retained as the "top-probable" values. The weights are normalized such that $\sum_{i=1}^{k} w_i = 1$, defining a discrete approximation to the updated distribution:

$$P(V = v_i) = w_i, \quad i = 1, \ldots, k$$

where V is the soil property at a given pixel and soil layer. The parameter k controls the level of truncation of the candidate set.

Comment M-2

Several choices raise questions that I do not think are adequately addressed. For instance, (A) The rationale for randomly selecting which depth layer to model per iteration is not explained. Does this improve convergence stability? No sensitivity analysis is provided.

Response: We thank the reviewer for your suggestion and raising this point. We did not apply sensitivity analysis on the selection order. The random ordering of depth layers per iteration serves two purposes: (1) it was designed to ensure model robustness. Soil properties often exhibit vertical structure, for example, soil organic matter is typically higher at the surface. If the model consistently followed a fixed top-down or bottom-up sequence, there is a risk that the residual corrections would "over-fit" the pattern of a specific layer early in the process; (2) random selection adds a form of stochasticity to the simulation. It improves convergence stability by ensuring that the feature space update path varies across iterations, analogous to stochastic gradient descent methods that avoid "local minima" by introducing randomness into the update order. We have added a brief justification of this design choice to Section 2.1 of the revised manuscript.

(B) The convergence threshold at the 5th percentile of value changes is unusual and seems inverted from conventional convergence logic.

Response: We thank the reviewer for suggestion and this careful reading. To clarify: convergence is declared when the median of residual changes across all pixels drops below a threshold. In the California implementation, that threshold is set to the 5th percentile of the initial residual distribution (i.e., a small value representing the lower tail of meaningful residuals), not the 5th percentile of current iteration changes. This means convergence is reached when the median update becomes negligibly small relative to the range of initial residuals. We have rewritten Section 2.1.2 and the California-specific description in Section 2.2.2 to make this logic explicit and unambiguous, and confirmed that the criterion follows conventional convergence logic (iterations stop when the median change is small). In addition, soil properties like pH (ranging from 3 to 10) and sand content (ranging from 0 to

100) differ by an order of magnitude. A single fixed constant would be too restrictive for some variables while being too lenient for others.

(C) Simply removing co-located samples does not fully address spatial autocorrelation in model evaluation. No spatial cross-validation scheme is implemented or discussed.

Response: Thank you for your insightful suggestion. We agree that OOB evaluation does not constitute spatially independent validation and that this is a meaningful limitation of our evaluation framework. We have added an explicit acknowledgement of this limitation in two places: (1) in Section 3.2 of the Results, where we note that *'While the OOB evaluation can be subject to spatial autocorrelation, we interpret these gains not as evidence of broad spatial extrapolation into unsampled regions, but more as the framework's improved capacity for data assimilation'*; (2) in Section 4.4 of the Discussion, which now discusses two concrete paths toward spatially robust evaluation: spatial cross-validation with a minimum distance constraint between training and evaluation sets, and independent evaluation using CONUS-wide networks such as the U.S. Climate Reference Network. We acknowledge that implementing a full spatial blocking cross-validation is a priority for future work.

A primary reason for not implementing spatial blocking in California (CA) is that soil properties in CA is highly spatially heterogeneous. If a block lacking a particular landscape type were used for evaluation while that type was present in training, the model would likely yield overly pessimistic. Therefore, the spatial cross-validation is more meaningful when conducted on a CONUS-wide training set that includes representative examples of each landscape type, allowing evaluation blocks to test the model's ability to generalize to unseen but ecologically similar conditions.

(D) The IRC method inherits all structural assumptions of the pHRF prior. If the prior is systematically biased in certain landscape types, the IRC correction may be insufficient. This vulnerability is not discussed.

Response: Thank you for providing this suggestion. We now address explicitly in Section 4.2 (Limitations in Soil Profile Data). We note that in data-sparse regions such as California's desert areas, the residual correction is constrained by the number of available profiles, and where the prior is systematically biased and local observations are few, the IRC correction may be insufficient to overcome the inherited prior error. The land-use-stratified vertical profile analysis (Figure 10) provides evidence of this: bulk density in cultivated areas shows larger posterior discrepancies than in forest or wetland settings, which we interpret as a case where agricultural management dynamics create a prior bias that is not fully captured by the available correction profiles.

We also want to clarity that a source of improvement in the IRC come from the integration of more availability of soil data. The prior estimates rely on a Harmonized Soil Property

Database built by depth-harmonizing and aggregating pedons from the National Cooperative Soil Survey (NCSS) Soil Characterization Database and map-unit components from the Soil Survey Geographic Database (SSURGO) into profiles for over 20,000 distinct soil series. However, this method oversimplifies soil variation and often smooths out soil variability within soil map units and within soil components. In contrast, the IRC method integrates georeferenced soil profiles, directly providing observed values of soil properties. By leveraging a large availability of new soil data, the posterior results can recover detailed spatial variations that were previously "lost" in the prior.

(E) The mechanics of ensuring sand + silt + clay = 100% when three texture fractions are corrected independently is not clearly described. Are the three texture fractions corrected jointly or sequentially? How is the closure constraint enforced?

Response: Thank you for your constructive suggestion. The three texture fractions are corrected independently (sequentially, one property at a time), and a compositional closure step is applied after all iterations are completed by redistributing any deviation from the unit sum proportionally across the three fractions at each pixel. We have expanded Section 2.1.3 in the revised manuscript to describe this procedure explicitly, step by step. We also now explicitly acknowledge the limitation of this approach: *since the fractions are not modelled as a joint compositional vector (e.g., within a log-ratio geometry such as an isometric log-ratio transform), the independent correction followed by proportional rescaling does not fully capture the inherent inter-dependencies and non-linear constraints between soil textures*. This is identified as an area for methodological improvement in future work.

Comment M-3

OOB sampling is insufficient as the primary validation strategy for a spatially explicit mapping method. OOB samples in Random Forest are withheld from individual trees but are still drawn from the same statistical population as training data. In a spatial context, this means OOB samples can be geographically proximate to training samples, and Random Forest predictions at those locations benefit from spatial autocorrelation with nearby training points. The result is typically an optimistic estimate of predictive performance, particularly for interpolation across data-sparse regions. The field measurement data from the Salinas Valley (Section 2.2.1.3) appears to be used for model training. It is unclear whether any of these samples are reserved for independent validation. Given that these are the only truly independent field measurements in the study, their exclusive use for training would be a missed validation opportunity.

Response: Thank you for the suggestion. We acknowledge that OOB validation is not spatially independent and that the performance metrics are likely optimistic, particularly in interpolation across data-sparse regions. This caveat is now stated explicitly in Section 3.2: *the OOB-based gains should be interpreted as evidence of improved data assimilation capacity*

rather than of generalisation to entirely unsampled regions. Regarding the Salinas Valley field measurements: these samples (Section 2.2.1.3) are used as part of the residual correction training set. We acknowledge that leveraging external dataset as independent evaluation would strengthen the study (while holdout a subset is not independent validation); this was not done in the current study because the sample size was limited and prioritised for model training. We identify this as a limitation in Section 4.2 and note that future work with CONUS-wide application will benefit from larger independent validation datasets such as USCRN soil observations.

RESULTS

The results are largely descriptive rather than analytical, the validation approach carries over the problems identified in the methods, and the uncertainty analysis is notably superficial for a study that foregrounds probabilistic outputs as a core contribution.

Comment R-1 & R2

OOB samples are not spatially independent. The metrics reported are likely to be overoptimistic. This needs to be acknowledged and discussed.

The magnitude of improvement from prior to posterior is remarkably large, for instance, sand R2 increasing from 0.35 to 0.84, and prior SOM going from a negative R2 to 0.61. While this could reflect genuine improvement, it also raises a methodological concern: the IRC correction model is trained and evaluated on overlapping sample sets (via OOB rather than a fully independent holdout). The dramatic improvements may partly reflect the IRC model simply learning the training data distribution rather than generalising spatially. This possibility must be acknowledged and discussed.

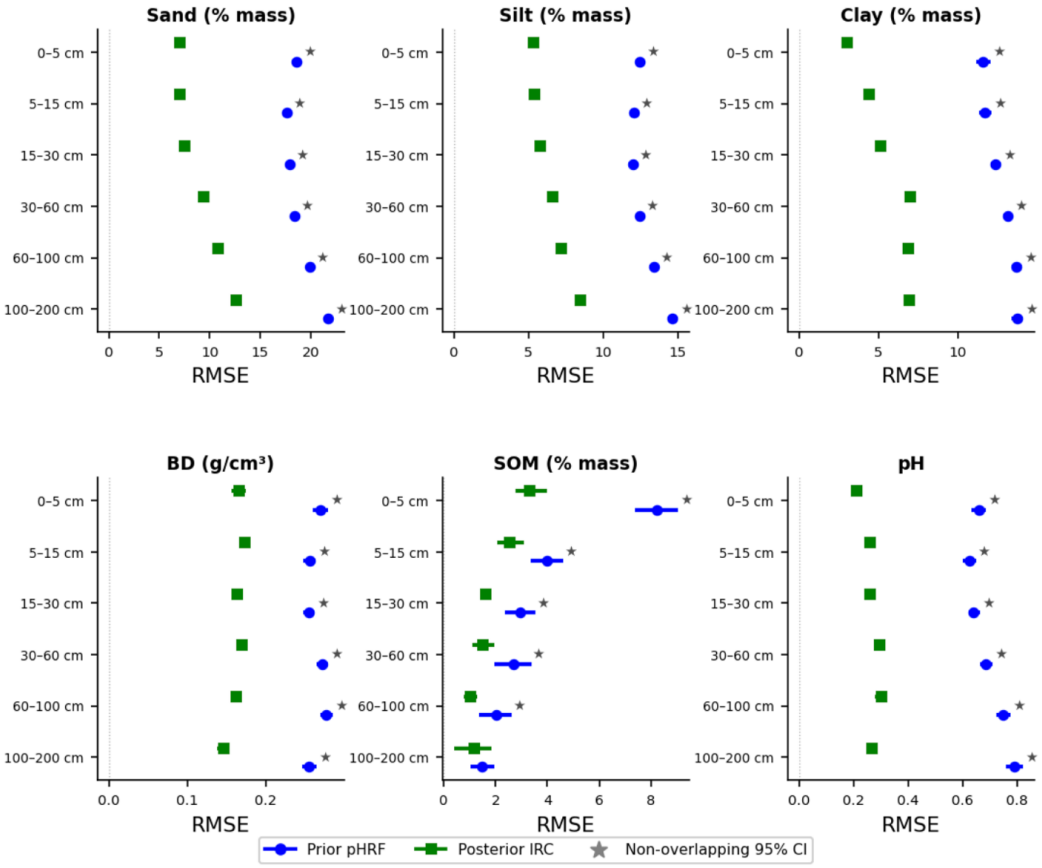
Response: We thank the reviewer for your suggestions. We agree and have addressed this in two ways. First, we added an explicit statement in Section 3.2: *'While the OOB evaluation can be subject to spatial autocorrelation, we interpret these gains not as evidence of broad spatial extrapolation into unsampled regions, but more as the framework's improved capacity for data assimilation.*'

Second, we reframe the large prior-to-posterior R^2 gains in context: part of the apparent improvement stems from the very poor prior baseline (discussed below in response to R-4), and part from the IRC model learning the local training distribution. We have added language in Section 3.2 and the Conclusion explicitly acknowledging that independent or spatially blocked evaluation is needed to confirm the degree of genuine spatial generalisation.

Comment R-3

Given that improvements are claimed across all six properties and all depths, even a simple paired comparison or confidence interval on RMSE differences would strengthen the conclusions.

Response: Thank you for your insightful suggestion. We have added Figure 7 (forest plots of RMSE with bootstrap-derived 95% confidence intervals, based on 1,000 bootstrap resamples) to Section 3.2 of the revised manuscript. These plots show prior and posterior RMSE with non-overlapping confidence intervals (CIs) for nearly all depth-property combinations, providing that the improvements are not due to sampling variability. The one exception SOM at 100-200 cm, where CIs overlap is explicitly discussed in Section 3.4 and attributed to the skewed distribution and small sample size in the extreme tail at that depth, rather than a pure failure of the IRC framework.

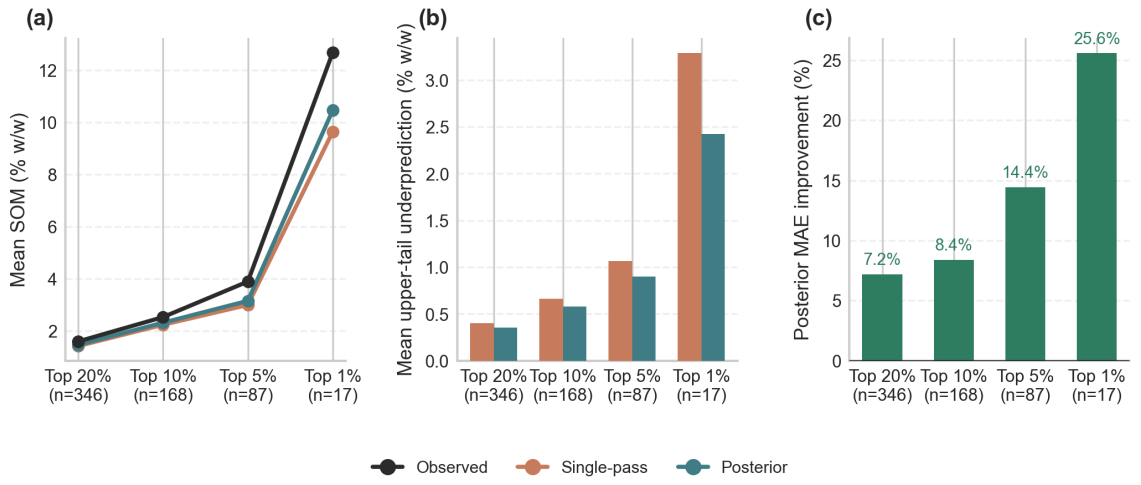


Comment R-4

The prior R2 values are very low: sand at 0.35, silt at 0.19, clay at 0.25, bulk density at 0.16, and SOM near zero or negative. These are very poor priors, which naturally makes any correction look impressive by comparison. The authors should acknowledge that the magnitude of apparent improvement is partly a function of poor prior performance, not solely a reflection of IRC's capability. A comparison against a simpler single-pass correction would

contextualise how much of this gain is attributable to the iterative framework specifically, versus residual correction in general.

Response: Thank you for your constructive suggestion. We agree and have addressed both parts of this comment. First, we explicitly acknowledge in Section 3.2 that the magnitude of improvement is partly a consequence of the prior's limited performance, which itself stems from the harmonized soil property database that aggregates pedon and map-unit components, inherently smoothing out intra-unit soil variability and intra-soil component variability. Second, we have added a dedicated Section 3.4 and Figure 13 that directly compares single-pass and iterative (IRC) residual correction for the case (SOM at 100-200 cm depth). The analysis focuses on the upper tail of the distribution (the segment where single-pass averaging-based corrections show CIs overlap) and demonstrates that the iterative posterior reduces MAE by up to 25.6% relative to the single-pass model for the top 1% subset, with improvement growing monotonically towards the extreme tail. This isolates the specific contribution of the iterative mechanism beyond simple residual correction.



Comment R-5

The metrics are not computed on the same samples for prior and posterior. It is implied but not explicitly confirmed that prior and posterior metrics are evaluated on the same samples. If they are not, the comparison is invalid. This needs clarifications.

Response: Thank you for your suggestion. We confirm that all prior and posterior performance metrics reported in Table 1 and Figure 6 are computed on the same set of co-located observations. Specifically, the OOB samples generated by the same Random Forest regressors used for residual correction. For the prior evaluation, the expected values (means) of the prior pixel distributions at co-located OOB sample locations are extracted and compared against the observed values using the same metric formulas. For the posterior, the expected values of the updated distributions at the same locations are used. We have added a clarifying sentence to Section 3.1 to make this explicit: *'All prior and*

posterior metrics are computed on the same co-located OOB sample set, using the expected value of the respective distributions at each sample location.'

Comment R-6

Unrealistically low SOM values generated during correction, attributed to van Bemmelen factor conversion is a methodological error that is disclosed but not resolved. If the conversion factor is introducing spurious values, this should be corrected before final mapping, not simply noted as an observation in the results. It undermines confidence in SOM posterior maps specifically.

Response: Thank you for your suggestions. We agree and would like to explain both why a full fix is not straightforward at the California scale, and a path forward is discussed in Section 4.2. The root issue is that the conversion factor between SOC and SOM is not a universal constant. Pribyl (2010) demonstrated through a comprehensive review that empirical conversion factors range from 1.4 to 2.5, with medians of 1.9 (empirical) and 2.0 (theoretical) being more defensible than the conventional 1.724. More importantly, recent work shows that the appropriate factor is modulated by soil properties: clay content and texture (Jensen et al., 2018; Hoogsteen et al., 2015), ecosystem type (Sanderman et al., 2018, who found factors of 1.92 for saltmarsh and a non-linear relationship for mangrove sediments). This means a principled resolution would not simply swap 1.724 for another fixed constant but would assign profile-specific conversion factors based on measurable soil attributes.

In practice, implementing this requires sufficient samples within each environment (texture class, land use, depth horizon and so on) to empirically calibrate or validate the factor. At the California scale with our current profile pool, the sample size, particularly wetlands and organic-rich soils, where the factor deviates most from 1.724, is insufficient to make defensible specific assignments. Flagging profiles by conversion method can identify which observations are at higher risk of conversion artefacts and provide a confidence level for SOM estimates in those regions, but it cannot correct the residuals already learned from potentially biased SOC-derived SOM values.

We therefore adopt the following position in the revised manuscript: (1) we retain the flagging mechanism as a measure that users of the posterior maps can apply to assess local reliability; (2) we identify soil-property-guided conversion factor estimation as a targeted improvement for the forthcoming CONUS-wide application, where the much larger and more diverse profile pool will support robust stratification by texture, land use, ecosystem type, and depth (following the framework of Jensen et al. (2018) and Pribyl (2010)), allowing conversion-induced artefacts to be corrected before profiles enter the residual correction pipeline rather than flagged after the fact.

Reference

Hoogsteen, M. J. J., Lantinga, E. A., Bakker, E. J., Groot, J. C. J., and Tittonell, P. A.: Estimating soil organic carbon through loss on ignition: effects of ignition conditions and structural water loss, Eur. J. Soil Sci., 66, 320–328, 2015.

Jensen, J. L., Schjønning, P., Watts, C. W., Christensen, B. T., Peltre, C., and Munkholm, L. J.: Converting loss-on-ignition to organic carbon content in arable topsoil: pitfalls and proposed procedure, Eur. J. Soil Sci., 69, 2018.

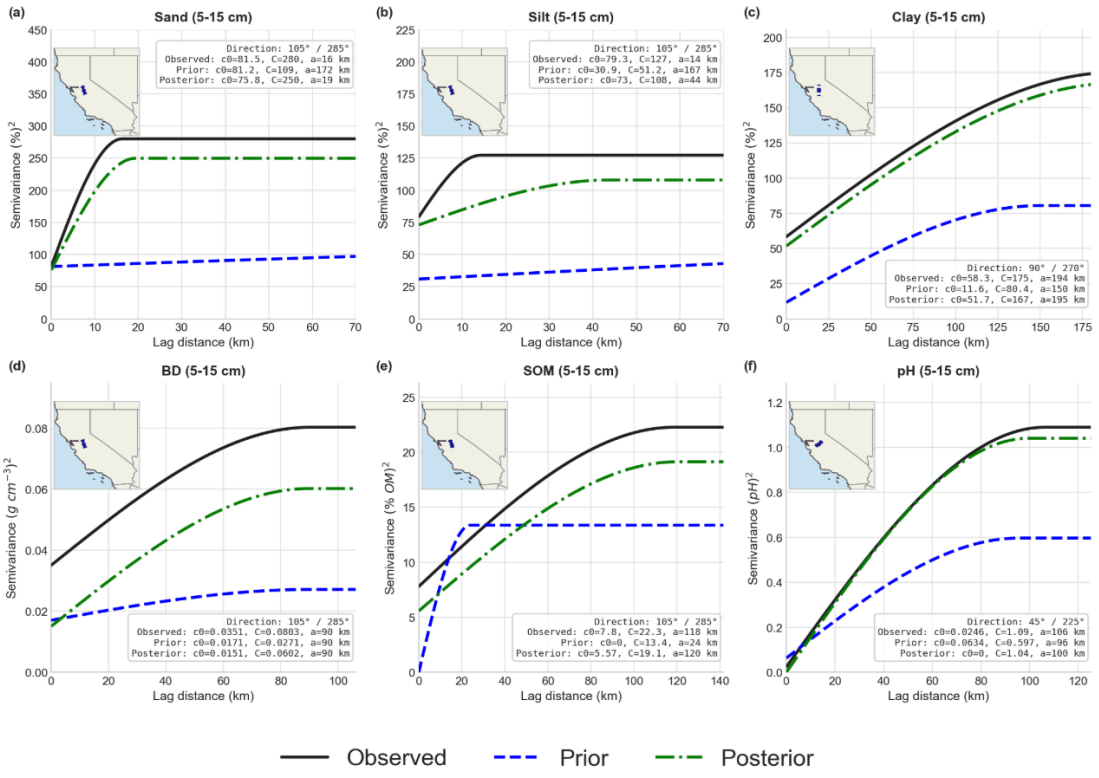
Pribyl, D. W.: A critical review of the conventional SOC to SOM conversion factor, Geoderma, 156, 75–83, 2010.

Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M. F., Benson, L., Bukoski, J. J., Carnell, P., Cifuentes-Jara, M., Donato, D., Duncan, C., Eid, E. M., Ermgassen, P., Lewis, C. J., Macreadie, P. I., Glass, L., Gress, S., Jardine, S. L., Jones, T. G., Landis, E., Lovelock, C. E., Waterhouse, J., Wollheim, W. M., and Landis, R.: Improved estimates on global carbon stock and carbon pools in tidal wetlands, Nat. Commun., 9, 4889, 2018.

Comment R-7

For a method that explicitly claims to improve spatial representation of soil properties, characterising whether the posterior maps have more realistic spatial autocorrelation structure would be a meaningful and expected result. This could be done with a semivariogram.

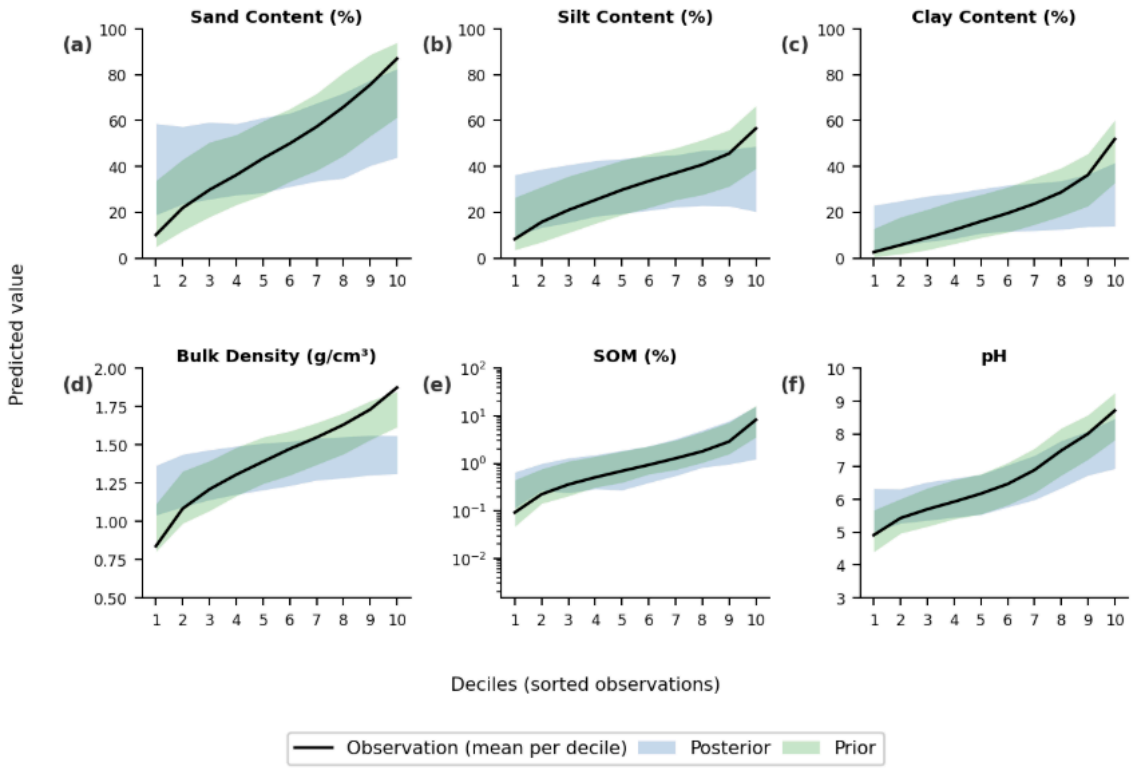
Response: Thank you for the constructive suggestion. We have added to Section 3.2 of the revised manuscript: a directional semi-variogram analysis for all six soil properties at the 5-15 cm depth over the Central Valley, California. Spherical models are fitted to the observed, prior, and posterior samples, and the nugget (C0), sill (C), and range (a) parameters are reported for each. The results show that the prior semi-variograms consistently underestimate the observed sill (reflecting the smoothing artefact of the prior) and misrepresent the effective correlation range. The IRC posteriors shift the semi-variogram towards the observed trend, recovering both the spatial variance structure and the correlation length.



Comment R-8

The uncertainty section is underdeveloped relative to the study's stated aims, and represents the largest gap between what is claimed in the introduction and what is delivered in the results. Reporting differences in prediction interval width (PIW) only tells us whether intervals got wider or narrower. It says nothing about whether those intervals are well-calibrated — i.e., whether a 90% prediction interval actually contains the true value 90% of the time. A narrower interval is only an improvement if it remains reliably calibrated. The absence of any calibration analysis (PIT histograms, coverage probability plots, interval score) means the uncertainty results cannot be interpreted as genuine improvements in uncertainty quantification.

Response: Thank you for your constructive suggestion. We have added Figure 12 to Section 3.3: prediction band fan charts that visualise the coverage of the 90% prediction intervals across ten sorted observation deciles for all six soil properties. These fan charts show whether the prediction bands encompass the observed rank-ordered values across the full range of the distribution, a coverage-based diagnostic that complements the PIW analysis. The results show that the posterior 90% bands shift towards the observed trend, particularly in the extreme deciles where the prior was most overconfident.



DISCUSSION AND CONCLUSIONS

Comment D-1

The discussion is the weakest section of the paper. It is largely a catalogue of limitations and a brief product comparison, with almost no engagement with the study's core findings in relation to existing literature, no statement on whether the research hypothesis was supported, and no synthesis of what the results mean for the field. Above, I noted various aspects of the method and results that warrant discussion. The conclusion is a near-verbatim summary of the methods and results rather than a set of genuine conclusions, which is a fundamental structural problem. These need significant attention.

Response: Thank you for your constructive suggestions. We agree that the Discussion and Conclusion required restructuring and have substantially revised both sections. The key changes are as follows.

(1) New Section 4.1 — Performance of the IRC Method and Its Implications. We have added an opening Discussion section that states that all four research objectives were met and that the central hypothesis is supported. We document performance gains (R^2 more than doubling for texture fractions, pH posterior R^2 = 0.94, SOM and bulk density moving from near-zero to R^2 of 0.61 and 0.70 respectively) and identify two drivers: the iterative model architecture and the integration of additional georeferenced soil data. The section then discusses what these improvements mean for the community: recovery of spatial structure masked by prior smoothing (semi-variogram evidence), improved uncertainty

calibration for risk-sensitive applications, and improved vertical profile fidelity relevant to soil hydrological modelling.

(2) Expanded Section 4.5 — Comparison with Existing DSM Products. A new paragraph contextualises the IRC framework relative to existing products in terms of how each represents uncertainty: SSURGO/gNATSGO report only component-level low/representative/high summaries; POLARIS provides continuous pixel distributions but derived from synthetic sampling that propagates smoothing artefacts; IRC dynamically updates uncertainty by assimilating new observations. The section also explains the two-source design (prior from georeferenced soil taxa and harmonised survey; posterior adding georeferenced profiles) and its practical implication for CONUS-wide coverage. The benefit to land-surface model parameterization is also discussed.

(3) Rewritten Conclusion (Section 5). The Conclusion has been entirely rewritten. It no longer restates the methods. Instead, it: (a) explicitly confirms that the research hypothesis is supported; (b) identifies three contributions of the IRC framework beyond accuracy improvement: a scalable pathway for continuous map updating, physically constrained posteriors, and improved uncertainty calibration; (c) gives an honest and specific statement of the study's limitations; and (d) provides a forward-looking potential of the framework's role in next-generation digital soil mapping.

We hope that the revised manuscript now addresses the concerns raised. Thank you again for your time and constructive suggestions!

Sincerely,

Chengcheng Xu and co-authors