Title: Open LandMap-soildb: Enabling High-Resolution Soil Intelligence for Climate, Land Restoration, and Agricultural Policy

Soil degradation is a growing global crisis, threatening food security, carbon sequestration potential, and ecological resilience. To tackle this, precise, spatially explicit, and temporally consistent soil information is essential. The newly developed OpenLandMap-soildb offers an unprecedented advancement in this space, providing global soil data at a fine spatial resolution (30 m) across two decades (2000–2022), using a spatiotemporal machine learning framework and harmonized legacy datasets.

This initiative delivers dynamic predictions for key soil properties including soil organic carbon (SOC) content and density, bulk density, soil pH, and USDA soil types. These outputs are based on over 1 million quality-controlled and harmonized soil samples, combined with Earth Observation (EO) satellite data, terrain models, and climatic indicators. Notably, the study estimates that the planet has lost more than 11 petagrams (Pg) of SOC in the top 30 cm of soil over the last 25 years, a signal of worsening land degradation and a missed opportunity for carbon sequestration.

This manuscript presents an ambitious and technically compelling global soil dataset spanning over two decades at high spatial resolution. The integration of legacy soil samples with modern satellite-derived covariates via machine learning methods is a noteworthy advancement for soil science and spatial ecology. However, certain methodological and interpretative aspects warrant clarification and refinement before publication.

RE: We thank our colleague for providing feedback on the article. Even though it appears that a large part of reviewer's notes have been generated using a LLM, we address some of the issues raised below.

Major Concerns

1. Model Transparency and Reproducibility
   The use of Quantile Regression Random Forests is appropriate, but the manuscript lacks sufficient detail regarding hyperparameter optimization, feature selection criteria, and potential overfitting mitigation strategies. o The approach to uncertainty quantification is promising; however, clearer guidance on interpreting prediction intervals in practical applications would enhance user comprehension.

RE: The paper is already 60+ pages with 18+ figures and explanation of steps is extensive. We would appreciate it if the reviewer would provide more detail about where exactly we *"lack sufficient detail regarding hyperparameter optimization"* etc. The scikit-learn framework we use is among the most used and most developed Machine Learning frameworks and the documentation is extensive including all exact steps how we fitted the models and fine-tuned hyperparameters (https://github.com/openlandmap/soildb/tree/main/modeling_steps).

2. Temporal Granularity
   Five-year intervals may oversimplify dynamic changes due to land use transitions or climate events. The authors should discuss how these limitations affect the detection of soil change patterns.

   RE: The rationale for 5-year intervals is discussed on the P43L15. In a nutshell, due to high data volumes and high costs of computing, but also due to limited accuracy / limited numeric resolution (Fig. 20; also discussed in https://doi.org/10.1016/j.jag.2012.02.005) we had to limit predictions to granularity that allows detecting significant changes in soil properties.

3. Spatial Validation Design
   There is limited description of spatial cross-validation strategies. It's essential to confirm the use of geographically independent test sets to avoid inflating predictive performance due to spatial autocorrelation.

   RE: Spatial and spatiotemporal CV strategies used have been discussed in detail on P22–23.

4. Legacy Data Harmonization
   While the dataset is impressively large, the harmonization process of legacy samples (e.g., sampling depths, analytical methods, and metadata consistency) needs greater transparency. Including a harmonization workflow or uncertainty estimates tied to legacy data variability would be beneficial.

   RE: All import, standardization and harmonization steps are discussed in detail and are documented using computational notebooks at: https://soildb.openlandmap.org/025-import_chemical_data.html.

5. Spatial Data Bias
   Over-representation of North America and Europe; sparse coverage in Asia, Russia, and Africa. This introduces spatial bias, which may influence the global model predictions unfairly, especially for underrepresented biomes and land-use systems.

RE: This is a correct point. Currently the biggest gap in geographical representation is in fact the Russian federation. Unfortunately we do not have any simple solution to this issue (except to motivate countries to collect new samples and share laboratory results openly). The Fig. 6, nevertheless, shows that at least all continents and all climate zones are represented for training of models.

6. Model decision
   Despite high accuracy, it reduces interpretability for policymakers or nonexpert stakeholders. More explainability or uncertainty quantification per region would improve utility. Inclusion of SHAP (Shapley Additive Explanations) or permutation importance at regional levels will improve the same. Offer uncertainty maps with visual warnings in extrapolated areas.

   RE: Prediction intervals (prediction intervals per pixels i.e. maps) are provided for all predicted variables. We also provide a step-by-step tutorial on how to visualize uncertainty:
   https://github.com/openlandmap/soildb/blob/main/OpenLandMap_soildb_tutorial.ipynb

Minor Suggestions

1. Heavy Reliance on Legacy Data
   Despite harmonization efforts, relying heavily on such datasets can propagate uncertainties, especially in dynamic time-series analyses
2. Soil Classification Framework
   The choice of USDA soil taxonomy over other globally recognized systems (e.g., WRB) should be contextualized, especially given the international scope of the dataset.

   RE: We are working on WRB predictive mapping framework. This should be available in early 2026.

3. Data Accessibility
   The use of Google Earth Engine and Cloud-Optimized GeoTIFFs makes the product accessible, but a brief tutorial or reference to documentation could help less-experienced users navigate it.

   RE: A tutorial on how to access data is available at https://github.com/openlandmap/soildb?tab=readme-ov-file#layers-available and https://github.com/openlandmap/soildb/blob/main/OpenLandMap_soildb_tutorial.ipynb
   All layers are also listed with metadata via:
   https://github.com/openlandmap/soildb

4. Environmental Covariates

Some satellite-derived indices (NDVI, GPP) may reflect transient vegetation conditions unrelated to underlying soil properties. A short discussion on how such confounding effects are addressed or minimized would be valuable.

RE: For this reason we use more complex modeling frameworks e.g. Random Forest combined with feature selection that tries to locally (RF = ensemble of regression trees) adjust and account for multitude of environmental soil forming factors.

5. Pseudo-Observations and Expert Knowledge Integration
   While this is a practical necessity, it can create artificial patterns in data that may not reflect on-ground conditions. This must be presented more cautiously in terms of predictive confidence. This is a highly promising contribution to digital soil mapping and global environmental monitoring.

   RE: We agree. Pseudo-observations should be added only when and where necessary, and need to be based on robust and reliable expert knowledge (so should be as least speculative as possible). We have documented all pseudo-observations in detail: https://soildb.openlandmap.org/025-import_chemical_data.html#glance-pseudo-samples; we use the GLANCE data set for pseudo-observations, which is based on the photo-interpretation using 30cm resolution VHR images and has been documented in detail in https://doi.org/10.1038/s41597-023-02798-5.

With improved methodological clarity and deeper contextual framing, the paper could serve as a benchmark for future soil informatics efforts.