Comments

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.

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.
    - The approach to uncertainty quantification is promising; however, clearer guidance on interpreting prediction intervals in practical applications would enhance user comprehension.
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.
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.

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.

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.

6. **Model decision**

Despite high accuracy, it reduces interpretability for policymakers or non-expert 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.

🧭 Minor Suggestions

- **Heavy Reliance on Legacy Data -**Despite harmonization efforts, relying heavily on such datasets can propagate uncertainties, especially in dynamic time-series analyses
- **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.
- **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.
- **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.
- **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. With improved methodological clarity and deeper contextual framing, the paper could serve as a benchmark for future soil informatics efforts.