Comment on https://doi.org/10.5194/essd-2025-336 "OpenLandMap-soildb: global soil information at 30 m spatial resolution for 2000–2022+ based on spatiotemporal Machine Learning and harmonized legacy soil samples and observations". Validation of soil organic carbon and bulk density predictions at the national scale of Mexico.

Carlos Arroyo, Viviana Varon, and Mario Guevara

Geosciences Institute, National Autonomous University of Mexico, Campus Juriquilla, Queretaro, Mexico.

The authors present an interesting spatial and temporal digital soil mapping effort to predict soil key variables at the global scale. Among other variables, soil organic carbon and bulk density are critical to understand soil responses to environmental change and land use. The authors increase the global availability of these variables with unprecedented spatial resolution for its use by multiple users across a large diversity of applications. There is a high scientific merit behind this effort and we hope to see the final version published soon.

However, important implications exist in the misuse of model derived products, because they are not error free and they include intrinsic and multisource uncertainty. In a revised version, the narrative could better prevent the misuse of soil model derived products across high uncertainty dominated areas. While the authors report relatively high accuracy in model predictions from cross validation, we hypothesize that such accuracy will drop-down significantly when compared with fully independent datasets, e.g., leave one dataset out cross validation, because each dataset is collected for a different purpose. Our overarching goal is to increase interoperability of digital soil mapping efforts from the plot, to the global scale. Therefore the objectives of this comment are a) to highlight the existence of fully independent national databases in Mexico that can be used to improve model accuracy of global soil predictions, or to calibrate country specific estimates, and b), to compare country-specific values of soil organic carbon and bulk density from fully independent datasets, with values derived from the new global soil variability models across 30m grids.

We use two fully independent datasets to validate global soil predictions at the national scale in Mexico. The first dataset was collected and analyzed by our National Institute of Geostatistics and Geography-INEGI in the year 2008 to assess soil erosion at the national scale, considering multiple land covers (INEGI, 2014). The second database was collected and analyzed by the former Ministry of Agriculture (now SADER) with support from FAO in 2012, considering only agricultural land (Arroyo et al, 2025). While the dataset from INEGI is representative of the topsoil, from the mineral surface to a maximum of 30 to 40 cm of soil depth, the agriculture soil dataset is representative of the first 30cm of soil depth. The INEGI dataset is available here: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825004223 and the SADER dataset is described and available here: https://bsssjournals.onlinelibrary.wiley.com/doi/10.1111/ejss.70116. Note that INEGI metadata is available in Spanish only (please let us know of any required assistance). We first download the soil datasets and model predictions from OpenLandMap-soildb, and then we compute the R2 between soil carbon and bulk density values from global predictions and the datasets. Because the agriculture dataset reports organic matter values rather than organic carbon, we use the conventional 0.58 factor as explained by (Van Bemmelen, 1897).

We observe, as expected, relatively low correlation compared to that reported in the paper, when comparing predictions against fully independent datasets (Fig. 1). Comparing global models with fully independent datasets is appealing to identify the main drivers of soil research across countries and identify the capacity of a global model to reproduce nationwide information.

Comparing all land uses, the correlation between the Openlandmap derived soil carbon values and the INEGI 2008 dataset increases significantly from R2 0.06 to R2 0.34 when transforming their values to a natural log scale. The Openlandmap soil carbon predictions and the soil carbon values in the dataset described in Arroyo et al, (2025) from 2012 across agricultural land only shows an R2 value of 0.23 that, interestingly, was not sensitive to the logarithmic transformation. Bulk density in the Mexican datasets is also different from that reported in the Openlandmap products (Fig. 1).

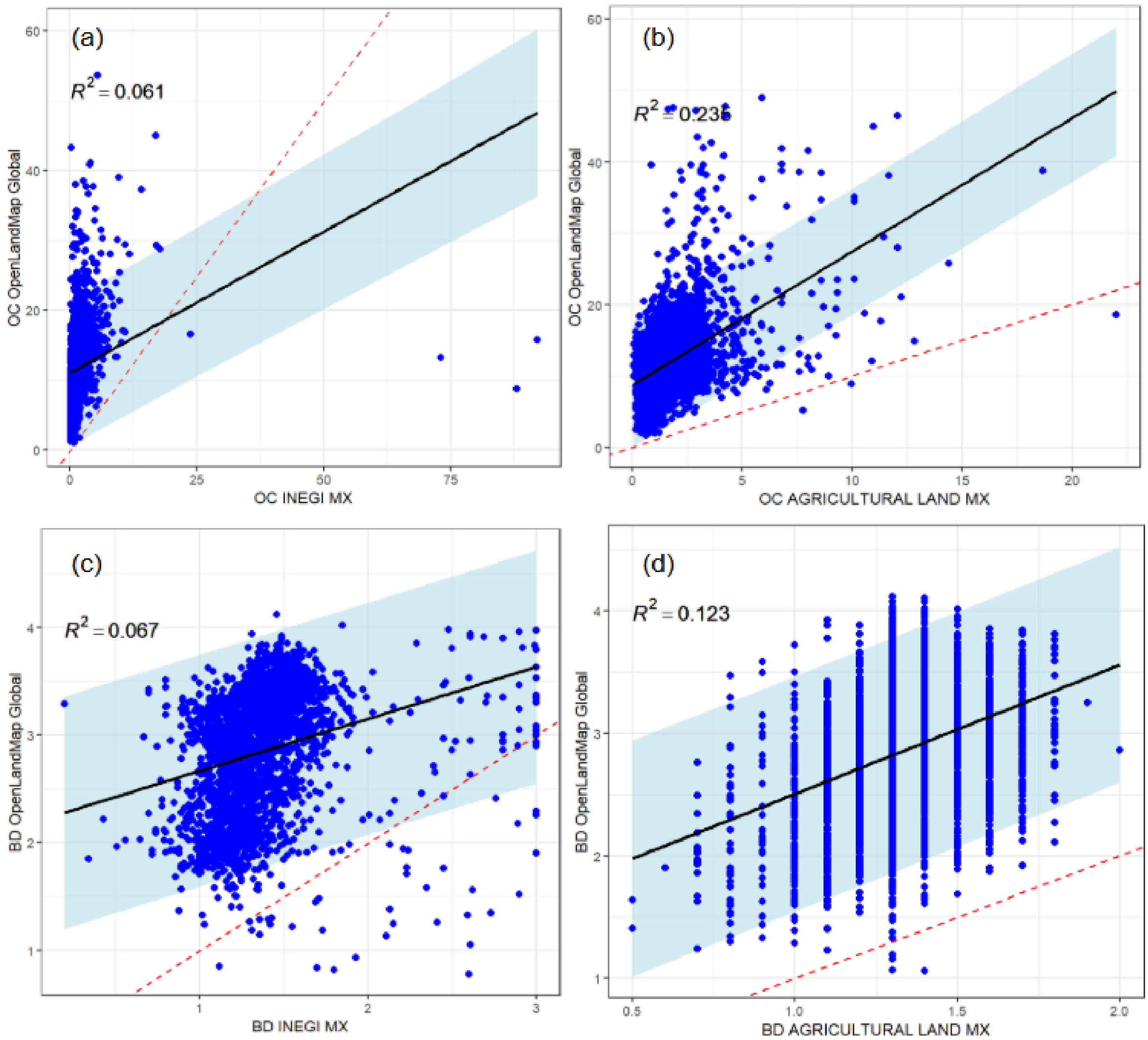

Fig 1 Scatterplots of soil organic carbon and bulk density values from the Openlandmap project compared with independent soil datasets across Mexico. Soil organic carbon carbon

across all land uses considering a national dataset representing the year 2008 showa lowest correlation against the Openlandmap product (a). Considering a national dataset collected in 2012, only agricultural land, the correlation is slightly higher (b). Bulk density across all land uses (c) and across agricultural land (d) show even lower correlation values.

It is clear that, based on R2 metrics and an independent dataset, the validation at the national scale is different from that reported in Openlandmap products. Due to this kind of global soil map being commonly used in governmental institutions for decision making, overall in countries with a lack of soil information. Therefore we propose that it would be interesting to report a country-based validation; for example, leaving- one-country-out validation. Maps of R2 variation across the world help users (i.e., public institutions, universities) to understand the specific limitations of global products in their countries.

The authors present an unprecedented opportunity to increase soil data quantity, quality and accessibility by combining local to national datasets into global soil variability models. The synergy between regional to global soil variability models brings positive implications towards more robust soil estimates (Zhang et al., 2025). We hope that the authors find the highlighted datasets useful for their global soil mapping efforts towards an increased interoperability among national to global soil mapping groups. We believe that highlighting all possible sources of uncertainty and clarifying the scope of the new information would help to promote the responsible use of global soil variability models. In conclusion, our comment enriches the ongoing discussion around global soil mapping by grounding it in real-world national data and offering constructive pathways for improvement. It’s the kind of feedback that can elevate both the scientific robustness and practical relevance of large-scale environmental models.

References:

Arroyo-Cruz, C.E., Prado, B., Kolb, M., Mora-Palomino, L.N., Todd-Brown, K. and Guevara, M. (2025), Synthesis of a National Soil Dataset Across Productive Land in Mexico: The Importance of Making Existing Data Accessible. Eur J Soil Sci, 76: e70116. https://doi.org/10.1111/ejss.70116

INERGI 2014, Conjunto de Datos de Erosión del Suelo, Escala 1: 250 000 Serie I Continuo Nacional https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825004223 (last accessed 07/08/2025).

Lei Zhang, Lin Yang, Yuxin Ma, A-Xing Zhu, Ren Wei, Jie Liu, Mogens H. Greve, Chenghu Zhou, Regional-scale soil carbon predictions can be enhanced by transferring global-scale soil–environment relationships, Geoderma, Volume 461, 2025,117466,ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2025.117466.

Van Bemmelen, J.M., 1897. Die Absorption. Das Wasser in den Kolloiden, besonders in dem Gel der Kieselsäure. Z. Anorg. Chem. 13 (1), 233–356.