https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Open Access Earth System Science Data Discussions

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

Tomislav Hengl[1], Davide Consoli[1], Xuemeng Tian[1], Travis W. Nauman[2], Madlene Nussbaum[3], Mustafa Serkan Isik[1], Leandro Parente[1], Yu-Feng Ho[1], Rolf Simoes[1], Surya Gupta[4], Alessandro Samuel-Rosa[5], Taciara Zborowski Horst[6], José L. Safanelli[7], and Nancy Harris[8]

[1]OpenGeoHub Foundation, Doorwerth, The Netherlands
[2]Moab, UT, USA
[3]University of Utrecht, Utrecht, the Netherlands
[4]Department of Environmental Sciences, University of Basel, Basel 4056, Switzerland
[5]Universidade Tecnológica Federal do Paraná, Santa Helena, Paraná, Brazil
[6]Universidade Tecnológica Federal do Paraná, Dois Vizinhos, Paraná, Brazil
[7]Woodwell Climate Research Center, Falmouth, MA, USA
[8]World Resources Institute, Washington DC, USA

Correspondence: Davide Consoli (davide.consoli@opengeohub.org)

Abstract. There is increasing interest in global dynamic soil information with changes in soil properties mapped over time and at high spatial resolution. Thanks to long-term, multi-temporal, and fine- and medium-resolution satellite missions such as Landsat, MODIS, Copernicus Sentinel and similar, it is possible to produce globally consistent predictions of key soil variables that match other 10–30 m spatial resolution global data sets. This paper describes data preparation, modeling, and production of OpenLandMap-soildb: global dynamic predictions of soil organic carbon content, soil organic carbon density, bulk density, soil pH in H_2O, soil texture fractions (clay, sand and slit) and USDA subgroup soil types (USDA soil taxonomy subgroups) at 30 m spatial resolution based on spatiotemporal Machine Learning (Quantile Regression Random Forest with output predictions showing the mean plus the lower and upper prediction intervals of 68% probability). To train the models, a large compilation of soil samples imported from legacy soil projects was used: 216,000 soil samples with soil carbon density (kg m^{-3}), 408,000 soil samples with soil carbon content (g kg^{-1}), 272,000 samples with soil pH in H_2O, 363,000 samples with clay, silt, and sand (%), and 134,000 samples with bulk density oven dry (t m^{-3}). Soil carbon and soil pH were mapped with 5–year time-intervals; soil texture fractions, bulk density, and soil types were mapped for recent years only. The cross-validation results indicate RMSE of 17.7 (kg m^{-3}; 0.486 in log-scale) and CCC of 0.88 for SOC density, RMSE of 51.3 (g kg^{-1}; 0.574 in log-scale) and CCC of 0.87 for SOC content, RMSE of 0.15 (t m^{-3}) and CCC of 0.92 for bulk density of fine-earth, RMSE of 0.51 and CCC of 0.91 for soil pH, RMSE of 8.4% and CCC of 0.87 for soil clay content, and RMSE of 12.6% and CCC of 0.84 for soil sand content respectively. The most important variables for predicting soil organic carbon density (kg m^{-3}) were: soil depth, Landsat-based uncalibrated Gross Primary Productivity (GPP), Normalized Difference Vegetation Index (NDVI) and CHELSA bioclimatic indices. The global distribution of soil pH can be primarily explained by the CHELSA Aridity Index (long-term), annual precipitation, and salinity grade. The global stocks for 2020–2022+ period for 0–30 cm depth interval are

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Open Access Earth System Science Data Discussions

estimated at 461 Pg (Peta grams); the results further indicate that, in the last 25 years, the world has lost at least 11 Pg of SOC in the top soil. Suggestions are made on how to set up global permanent monitoring stations to accurately track land degradation and enable land restoration projects. The training dataset is available at https://doi.org/10.5281/zenodo.4748499 (Hengl and Gupta, 2025), while the resulting data products can be accessed at https://doi.org/10.5281/zenodo.15470431 (Consoli et al., 2025). Both datasets are released under a *CC-BY* license.

1 Introduction

Soils symbolize fertility and are the foundation of our civilization; one of the most undervalued natural resources. Changing that perspective is a mission worth dedicating a career. Common modern threats to soil health include the loss of organic matter, the loss of biodiversity, soil pollution, soil salinization, and soil erosion. There is an increasing focus on soils due to their importance for ecosystem services: from growing crops, to filtering water, and providing building material (Smith et al., 2020). Soils are also one of the potential carbon pools that could significantly help decrease greenhouse gas (GHG) emissions in the atmosphere. Unsustainable land use and population pressure are the main drivers of soil degradation (Montgomery, 2007; Borrelli et al., 2017; Kraamwinkel et al., 2021). We are at a crossroads in history in our attempt to preserve soil resources before we completely lose them.

It is, in fact, a striking paradox that on the one hand, soils are one of the most promising solutions for mitigating greenhouse gas emissions, while, on the other hand, 60–70% of soils are currently unhealthy (Panagos et al., 2022). In the last 150 years, half of the topsoil on the planet has been degraded due to erosion, compaction, desertification, acidification, and loss of organic carbon and primary nutrients; mostly due to changes in global land use and climate. Hou et al. (2025) estimate that 14–17% of all croplands are polluted with toxic metals exceeding agricultural thresholds. Moreover, soil erosion could increase up to 60% in the next 30 years (Borrelli et al., 2017). For instance, the Continental United States alone may lose 1.8 Pg (petagrams) of soil organic carbon under climate change (Gautam et al., 2022). Padarian et al. (2022a) estimates that agricultural land could lose approximately 14% of the carbon sequestration potential of soil by 2040 due to climate change. Meanwhile, some recent estimates by Sasmito et al. (2025) indicate that half of the land use carbon emissions in Southeast Asia can be mitigated through the peat swamp forest and mangrove conservation and restoration. Padarian et al. (2022a) estimates the additional SOC storage potential in the topsoil of global croplands to be between 29 to 65 Pg C.

The ability to measure and evaluate progress towards maintaining or restoring healthy soils will be critical to the success of improved land management promoted by stakeholders and policy makers. Today, every land manager should have easy access to verified GHG emissions and removal data at the parcel level, and carbon farming must support the achievement of the proposed net removal targets — for example, 310 Mt CO2eq in the land sector in the EU until 2030 (Searchinger et al., 2022). However, the production of reliable estimates of global SOC stocks and SOC carbon sequestration has proven complex (Scharlemann et al., 2014; Minasny et al., 2017). The uncertainty in the estimates of the total organic carbon stocks in the soil of our planet for the 0–1 m depth interval is large (Scharlemann et al., 2014; Tifafi et al., 2018; Feeney et al., 2022; Lin et al., 2022), leading to problems of general credibility of these maps.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Direct measurement of soil properties from space is cumbersome (van Wesemael et al., 2024; Broeg et al., 2024; Li et al., 2024). Soils are often hidden below the surface under dense vegetation, and most EO systems do not penetrate the soil. Saha et al. (2024) reviewed the direct use of EO products and systems to monitor SOC from space and concluded that direct SOC detection is limited due to the low signal-to-noise ratio and low spectral resolution: most predictive mapping models have a limited R^2 between 0.3 and 0.7. Even bare surface spectra can be used to represent only the first few centimeters of topsoil, while, on the other hand, many studies often ignore soil management practices such as crop rotation, conservation tillage practices, fertilization level, plow depth, addition of manure to soil, and similar (Saha et al., 2024).

The uncertainty about how much organic carbon is in the soil and how much could potentially be sequestered appears to be high, especially for northern latitudes, tropical peatlands / wetlands and semi-arid areas (Crowther et al., 2016; Lin et al., 2022). The most up-to-date point data from Canada and the Russian Federation now indicate that large pools of soil organic matter in tundra and taiga-like biomes have probably been underestimated in previous global maps (Shaw et al., 2018; Wagner et al., 2023). Global warming and rising temperatures are likely to perpetuate the release of soil carbon in high-latitude areas dominated by permafrost (Crowther et al., 2016; Van Gestel et al., 2018). Therefore, accurate estimates of the carbon budget beyond 60° north, including the distribution of peatland soils (covering only 2–3% of the total area, but representing probably 40–50% of total stocks), are increasingly important. In tropical areas, Xu et al. (2018) and Gumbricht et al. (2017) have estimated that the extent of peatlands is somewhat larger than expected (currently estimated to be 2.8% of the total land mask), and there appear to be still many unmapped bogs of peat and organic material, especially in Latin America (Gumbricht et al., 2017), Africa (Fatoyinbo, 2017), and mangrove forests (Atwood et al., 2017). Deforestation and degradation of tropical forests appear to also perpetuate the loss of SOC (Drake et al., 2019).

Some of the most recent global maps of SOC at 1 km and 250 m are provided by FAO (2022) and Poggio et al. (2021). At the continental level, Yigini and Panagos (2016) produced detailed SOC maps for Europe; Liang et al. (2019) for China; Hengl et al. (2021) for Africa; Guevara et al. (2018) for South America; Ramcharan et al. (2018) and Nauman et al. (2024) for the United States. Beyond mapping the general spatial distribution of SOC, there is also an increasing interest in mapping changes in soil properties over time, with a special focus on soil carbon, soil nitrogen, pH, and other soil nutrients that are more dynamic and prone to land management changes (National Academies of Sciences, Engineering, and Medicine and others, 2021; Broeg et al., 2024; Li et al., 2024). Although soils change gradually, often on a scale of a few hundred years, locally there can be drastic effects, especially as a result of land degradation or sudden change of land use. In general, current systems in place to monitor soil properties (physical, chemical, and biological characteristics) together with soil loss and soil degradation measures do not provide sufficient information to accurately quantify changes in soil resources over time (National Academies of Sciences, Engineering, and Medicine and others, 2021).

The three most common groups of soil properties of interest for dynamic mapping are soil organic carbon stocks, soil nutrients (Chen et al., 2022), and soil hydrological properties such as available soil water (López-Ballesteros et al., 2023) and soil moisture content. Guo and Gifford (2002); Stockmann et al. (2015), and Stumpf et al. (2018) focused on modeling changes in SOC primarily as an effect of changes in land use and/or land cover over decades. The second most important soil-forming or controlling factor for predicting SOC changes at large scales is climate. Jones et al. (2005) and Gottschalk et al. (2012),

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

for example, provide estimates of changes in SOC due to climate change, with a special focus on predicting potential SOC losses in the future. Padarian et al. (2022b) proposed a two-step semi-mechanistic framework to model SOC over time: first, the baseline of the SOC stock is estimated using predictive mapping (in this case the baseline is the year 2001), and second, the SOC values are then propagated year by year over time by incorporating changes in land cover. Padarian et al. (2022a) uses a similar data set to estimate the SOC sequestration potential for agricultural land. Heuvelink et al. (2021) mapped the SOC dynamics of Argentina at 250 m spatial resolution using a time series of NDVI images for 1982–2017 and Random Forest. Their results indicate that, in fact, bio-climatic variables are somewhat more important than NDVI images for modeling SOC. Ugbemuna Ugbaje et al. (2024) developed spacetime predictions of SOC stocks for Australia at a 90 m spatial resolution covering 1990 and 2018. Venter et al. (2021) produced three decades of predictions of top-soil stocks for South Africa at 30 m spatial resolution; based on the time-series of predictions authors also provide estimates of soil carbon change in kg m^{-2} (for 0–30 cm depth interval). van Wesemael et al. (2024) produced triannual predictions (2018–2020, 2019–2021 and 2020–2022) of top-soil SOC (in %) for European Union, using a combination of spectral models for croplands (bare surface soil spectra) and the digital soil mapping approach for forest and grasslands.

Currently, the most referenced global soil data set with prediction intervals per pixel is SoilGrids V2.0 available at 250 m spatial resolution (Poggio et al., 2021). In addition, the FAO has recently updated the Harmonized World Soil Database (HWSDV), produced at 1 km spatial resolution (FAO & IIASA, 2023) and is also maintaining the Global Soil Partnership's GSOCmap (FAO, 2022). In practice, all three (SoilGrids V2.0, GSOCmap, and HWSDB) are lagging behind in spatial resolution with comparable global vegetation data sets, now usually focusing at 30 m or even 10 m, e.g., representing land cover dynamics (Potapov et al., 2020), crop classification (Van Tricht et al., 2023), forest canopy parameters (Turubanova et al., 2023), and similar. In addition, updating global soil maps for shorter periods, such as 1–2 times a year, has never materialized.

In this paper, we describe a fully documented open framework for producing predictions of primary dynamic soil properties at 30 m spatial resolution for the period 2000–2022+ (5–year composites), in addition to the spatial distribution of soil types. We focus on the following four research questions:

- R1: Do Landsat 30 m resolution images help improve the accuracy of predictions? If so, which Landsat-derived biophysical indices are the key for soil mapping?
- R2: How well do predictions from global models compare to observed values at locations not used in the map calibration/training, i.e., what is the expected prediction error at unvisited locations?
- R3: What are the key drivers leading to changes in SOC? How, for example, does conversion of tropical forests to croplands and pasturelands impact SOC and pH on a scale of 25+ years?
- R4: What are the world's remaining hotspots of SOC stocks?

We first present in detail all the data preparation, modeling, and prediction steps and how accuracy was assessed using robust procedures. In the results section, we report results of standardization, accuracy assessment, and change-analysis. We also provide visual evidence of patterns in the predictions and zoom in on the potential drivers of change in soil properties.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025


Open Access Earth System Science Data Discussions

The data and code used to produce the results and instructions on how to access the data are publicly available through https://github.com/openlandmap/soildb.

2 Materials and methods

In the following sections, we explain in detail how the point (training) data were prepared, how the covariate layers were selected and prepared for analysis, how and why we inserted pseudo-observations, and why we have made some design choices. In addition, we explain how we conducted cross-validation and how the prediction intervals were derived (per pixel). We run extensive tests to check predictive performance and then report results in both original and transformed spaces, which is especially important for log-normal and composite variables.

2.1 Spatiotemporal Machine Learning

We developed a fully automated global soil mapping framework based on a large stack of covariate layers representing the standard soil-forming and controlling factors (relief, climate, parent material, living ecosystem, and human impact) (Jenny, 1994) and an optimized machine learning pipeline as implemented in the scikit-map library for Python. The general soil mapping framework is illustrated in Fig. 1 and has been used to predict continuous dynamic soil variables and static soil properties, i.e., soil types and physical soil properties. We refer to the mapping framework as the *"EO-SoilMapper"* because the most important covariate data are the Earth Observation (EO) time series of images. We are able to produce predictions at 30 m and for a period of almost 25 years, mainly because we use the complete and cloud-free Landsat Archive previously prepared by Consoli et al. (2024), and the global digital terrain model (DTM) and its multiscale variables produced by Ho et al. (2025).

Spatio-temporal Machine Learning (ML) implies (Hackländer et al., 2024; Tian et al., 2024):

1. *Spatio-temporal overlay*: observations & measurements (O&M) are overlaid with covariate layers by matching both the geographic location and the start / end time period. In this paper, we only match observations by year, although some soil properties, such as soil moisture, would also require refined temporal identification.

2. *Strictly defined time-period of interest*: covariate layers need to match the distribution of O&M's in the time domain, i.e., there needs to be enough training points spread across the period of interest (in this case 2000–2022+).

3. *Spatio-temporal cross-validation*: for accuracy assessment, we report both spatial blocking cross-validation and leave-one-year-out (LOYO) cross-validation to prevent producing over-optimistic validation results for densely sampled/clustered points, due to e.g. strong spatial auto-correlation.

4. *Predictions in spacetime using spacetime blocks*: predictions are strictly spatio-temporal, i.e., they are connected with certain begin/end time periods. We refer to the spacetime prediction reference as *"spacetime blocks"*.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

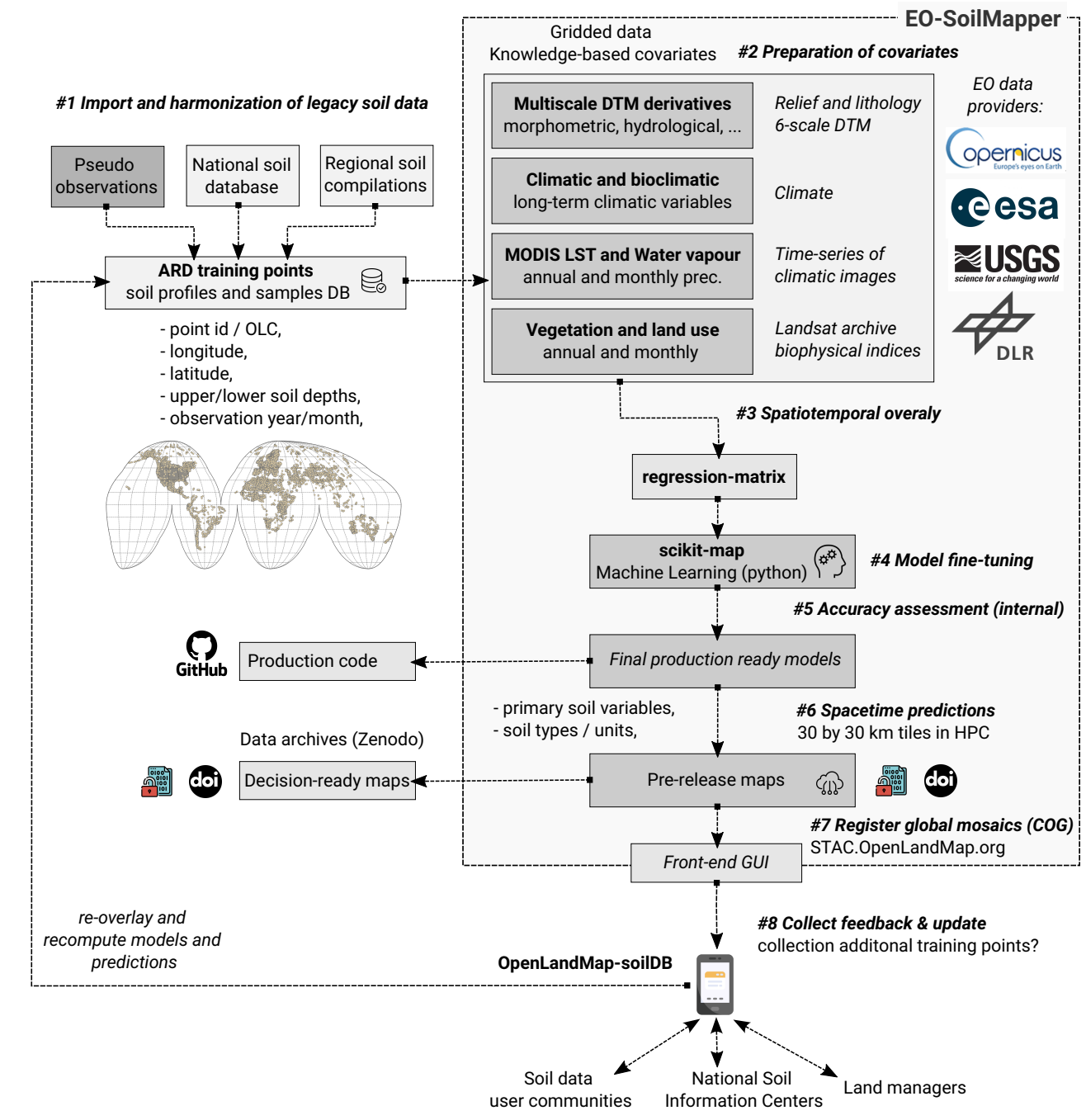

Figure 1. The general processing diagram of EO-SoilMapper with key steps. This is a modular system with four main components developed independently: (1) standardized soil samples, (2) covariate layers, (3) the computing engine, and (4) back-end/front-end infrastructure for serving seamless data. ARD = Analysis-Ready Data, OLC = Open Location Code, DOI = Digital Object Identifier, DLR = The German Aerospace Center. Automation of modeling, model fine-tuning and prediction is important as it allows for updating the predictions as more training data is added.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025


Open Access Earth System Science Data Discussions

2.2 Domain of interest: global land mask

We generate predictions for the global land mask at 30 m resolution withing the the 2000–2022+ period (with years 2023 and 2024 under production). To derive a consistent land mask, we used GDTM30 (global DTM at 30 m) (Ho et al., 2025) and time-series of land cover maps 2000–2022 (Zhang et al., 2021). We derived a long-term land mask based on a land-conservative assessment of the ocean mask for 2000–2022+, so that some pixels are potentially covered with water in more recent years.

We mask out the world's deserts and permanent ice to avoid predicting values or soil types for areas that are marginally *soil* (e.g. the Sahara desert) or are completely hidden. We recommend instead using standard values for shifting sand areas as follows:

- 0 value for soil carbon content/density, total N, P, and K;
- 100% for sand content;
- 0% for clay/silt content;
- 1.6 t m^{-3} for bulk density;

The 30 m resolution maps are about 70 times larger in size than 250 m resolution maps. The land mask at 30 m resolution in EPSG:4326 projection system (WGS84) contains about 210 billion pixels, while without deserts and permanent ice, about 190 billion pixels. Predictions of 1 soil variable for 5–year periods for 3 standard depths with lower and upper prediction intervals account for about 9 trillion pixels; as size on disk, this results in about 5–10TB of data (after compression). Because we also provide predictions for blocks of years, our outputs are even a few hundred times larger in size than long-term 250 m products.

2.3 Target soil variables of interest

As target variables of interest for dynamic soil mapping, we consider the list suggested by Chen et al. (2022), which is based on bibliometric analysis, and the variables listed in National Academies of Sciences, Engineering, and Medicine and others (2021). As Tier 1 variables of interest, we especially focus on soil organic carbon (SOC) content (g kg^{-1}), soil organic carbon density (kg m^{-3}), soil pH in H_2O, texture fractions (sand, silt, and clay) based on USDA system, bulk density (t m^{-3}), and soil types. We use USDA and/or ISO variables and laboratory standards as much as possible, as these are documented in the highest detail and are often used in international projects; for example, we use Dry Combustion for SOC and USDA soil taxonomy for soil types, which is fully open access documentation available to everyone.

Soil organic carbon density (SOC in kg $m^{-}3$) can be estimated at site level and is the central and most important variable of interest for global soil mapping. SOC density can be used to derive organic carbon stock in t ha^{-1} (Hengl and MacMillan, 2019):

$$\mathrm{SOCd}[\mathrm{kg\ m^{-3}}] = \frac{\mathrm{SOC}[\%]}{100} \cdot \mathrm{BD}[\mathrm{kg\ m^{-3}}] \cdot (1 - \frac{\mathrm{CF}[\%]}{100}) = \frac{\mathrm{SOCs}[\mathrm{kg\ m^{-2}}]}{\mathrm{HT}[m]} \quad (1)$$

where BD is the bulk density of fine earth, CF is the volumetric percent of coarse fragments, HT is the thickness of the horizon layer, and SOCs is the organic carbon stock of the soil for the specific depth interval. Correction for gravel content is necessary

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

because only material less than 2 mm is analyzed for SOC concentration. In principle, SOCd (kg m^{-3}) is strongly correlated with SOC content (g kg^{-1}). However, depending on soil mineralogy and coarse fragment content, SOCd can differ from the SOC content. SOCd can be estimated per depth interval (as indicated in Fig. 2), then aggregated to produce SOC stocks.

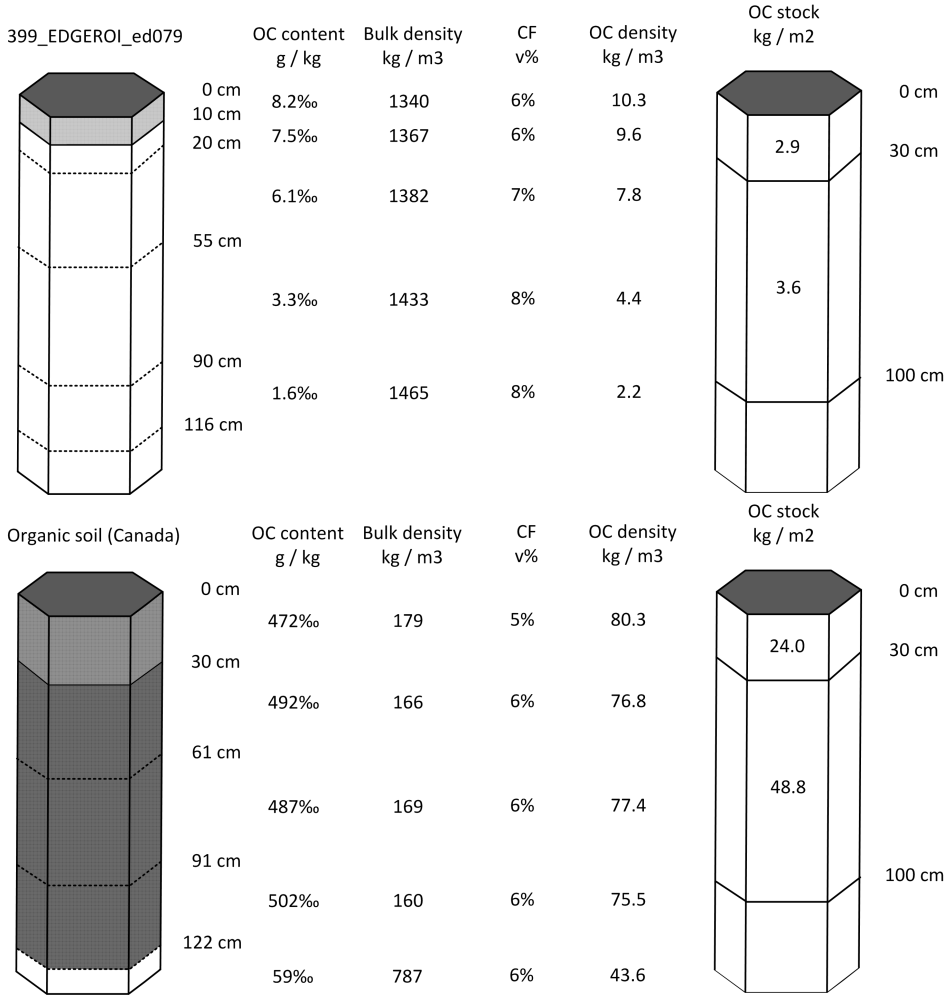

Figure 2. Determination of soil organic carbon density and stock for standard depth intervals: example of a mineral soil profile from Australia (above), and an organic soil profile from Canada (below). Image source: Hengl and MacMillan (2019).

Note also that values of SOCs in kg m^{-2} can also be expressed in t m^{-3}, in which case a simple conversion formula can be applied:

$$1 \cdot \text{kg m}^{-2} = 10 \cdot \text{tons ha}^{-1} \quad (2)$$

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Total SOC in tonnes for an area of interest can be derived by multiplying SOCs by total area e.g.:

$$120\text{tons ha}^{-1} \cdot 1\text{km}^{-2} = 120 \cdot 100 = 12,000\text{tons} \quad (3)$$

To determine stocks using global maps, one needs to first reproject SOCd predictions (kg m^{-3}) to some equal-area projection such as the Interrupted Goode Homolosine (EPSG:54052) (Steinwand, 1994). Next multiply the SOCd in kg m^{-3} with total area to obtain a total number of kg for the whole land mask. Another option is to determine the size of each pixel in WGS84 lon-lat projection system, although this can get computational. In this paper, we visualize all the maps and determine all areas using the IGH projection.

2.4 Preparation of training points

As training points for global soil mapping, we use a compilation of harmonized and quality-controlled soil O&M's listed at https://soildb.OpenLandMap.org/, which took several years to organize, import, standardize and harmonize. The data sources for the training data included:

- Original national or regional monitoring networks with probability sampling, quality controlled, and maintained by federal/national agencies (**L1**);
- Original national or regional 1–time surveys with probability sampling, quality controlled and fully documented (**L2**);
- Original regional or local soil sampling projects based on free-sampling (i.e. opportunistic sampling), but quality controlled, and fully documented (**L3**);
- Compiled national or regional soil legacy O&M's data sets, quality controlled and maintained; usually documented in a peer-review publication (**L4**);
- Compiled international, national or regional soil legacy O&M's data sets, quality controlled and fully documented, but with significant missing information about laboratory methods (**L5**);
- Compiled international, national or regional soil legacy O&M's data sets, usually not quality-controlled, based on unknown methods, including based on citizen-science data (**L6**);
- Other soil legacy O&M's data sets without a peer-review publication, with significant missing information about laboratory methods (**L7**);

We have put the greatest effort into importing and binding L1–L3 data sets such as the National Cooperative Soil Survey Characterization Database (http://ncsslabdatamart.sc.egov.usda.gov/) and the United States National Soil Information System, LUCAS soil (Orgiazzi et al., 2018), Brazilian PronaSolos (Polidoro et al., 2021), CSIRO's National Soil Site Database (CSIRO, 2024), Agriculture and Agri-Food Canada National Pedon Database (Geng et al., 2010), and the Mexican soil samples national inventory (Paz-Pellat and Velázquez-Rodríguez, 2018). These represent more than 80% of the training points used and were

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

essential to produce global predictions. The L1–L3 points are also the largest in volume, especially the NCSS Soil Characterization Database for USA, and a combination of LUCAS soil and national data sets for Europe.

From the 10 world's largest countries, the largest gaps in training data are because only very limited training data is available for 2000–2022 for India, the Russian Federation, China, and Kazakhstan. Although national data sets are available for Russia and China (Shangguan et al., 2013), these do not cover the 2000–2022 period and are relatively sparse. Likewise, the Canadian CUFS data set (Shaw et al., 2018) is a great open resource of soil laboratory data, however, it does not overlap in time with the 2000–2022+ period, and therefore was not used for modeling.

From the L4 data set, we should especially emphasize the following four (each covering larger region/continent): Africa Soil Profile Database (Leenaars et al., 2014), Latin America and Caribbean Soil Information System (SISLAC) database (Díaz-Guadarrama et al., 2024), Northern circumpolar permafrost soil profiles (Hugelius et al., 2013a), and the Mangroves soil data base (Maxwell et al., 2023). We also used several global or near-global databases produced as compilations from old reports and scientific papers (L5), for example: ISRIC's WoSIS (Batjes et al., 2024), Fine Root Ecology Database (FRED) (Iversen et al., 2017), Soil Health DB (Jian et al., 2020), and the International Soil Carbon Network Database (Harden et al., 2018). Many of these are actually compilations of the above-listed national or regional databases and, as such, do not necessarily need to be imported, as this could lead to many duplicates (these would hence be a compilation of compilations). Some, however, contain additional smaller data sets contributed by smaller organizations or individuals. Thus, it was important to import and check all available point data sets to avoid missing out.

From citizen science data (L6) the significant data set is the LandPKS app (Quandt et al., 2018) observations (165,000 observations with coordinates on December 2024), which is currently the biggest L6-type soil data set for global soil mapping. Beyond citizen science data, we also used a significant number of pseudo-observations (documented in the next sections) to help also represent areas with extreme climate/landscape conditions, e.g., shifting sands/deserts, mountain peaks, and bare rock areas. Pseudo-observations were added primarily to represent and integrate soil knowledge into ML.

We provide all import and harmonization steps in https://soildb.OpenLandMap.org/ and explain how to access the analysis-ready compiled and harmonized soil samples. Some training soil points are proprietary as we have signed a data sharing agreement that limits us to share them publicly, but we always provide preparation steps and a description of the data so that eventually users can detect any potential standardization / harmonization issues.

For mapping soil types (USDA subgroups), we used a compilation of points provided by the USDA (about 320,000 locations with soil classification) and extended it with harmonized soil profiles from various other projects, especially WoSIS points and national soil profile data sets. To reduce global gaps, we put particular effort into translating some compatible national soil classification systems, e.g., the Brazilian soil classification system and the Canadian soil classification systems. Usually, we translate the input Canadian or Brazilian classes to the 2 to 3 most probable soil types using the recommended translation tables (Krasilnikov et al., 2009); translating to multiple classes is more realistic, but results in many duplicate points. This inherent classification uncertainty is further propagated in the models. Translated classes were, however, only used for classification, but not for validation (as hold-out samples).

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

In principle, only USDA soil points with soil types are fully harmonized and can be considered analysis-ready, while other data sets required careful checks and preparation, so they could also be included in the analysis. To speed up the cleaning up of points for soil type mapping, we used the following three strategies:

- We use fuzzy search strategies to avoid missing out points with possible types or missing *"s"* at the end of the soil type. For example, a text containing *"typic haplaquoll, fine loamy mixed mesic"* will be matched with the targeted soil type *"typic haplaquolls"*. Fuzzy matching has been implemented using the `agrep` function in R with `max.distance=0.02`, `ignore.case=TRUE`; this has been shown to perform the best in removing only incompatible classes.
- We search for soil types in multiple columns in the soil profile databases. For example, in the case of the Australian CSIRO NatSoil database, some USDA soil classification is only available in comments.
- We record all translations and soil types cleaning in one large Google Sheet so that one can track all steps (see https://doi.org/10.5281/zenodo.4748499).

The import, translation and binding of soil-type training points are also fully documented in https://soildb.OpenLandMap.org/. In the end, these efforts provided a total of 332 thousand training points with soil type (USDA soil taxonomy subgroup), which yielded slightly more spatial locations than we prepared for soil property mapping. Unfortunately, most of the points (>80%) with USDA soil taxonomy are located in the USA and, as such, the North American continent is overrepresented in our models.

To quantify potential extrapolation problems due to spatial clustering and geographical gaps in point data, we run the Isolation Forest (Liu et al., 2008) on the training points and the selected most important covariates to produce an extrapolation risk probability map. This was only used to illustrate the effects of overrepresentation of training points, and of course to suggest to next generation projects where to place more samples in the future to help improve these predictions.

2.5 Standardization and harmonization

Before spatial analysis, it is important to standardize (convert to the same measurement units, the same physical standards) and harmonize (bring to the same laboratory reference methods) soil laboratory data to avoid potential bias in predictions and could also have serious consequences on decision making. From all the variables analyzed in soil science, the organic carbon and texture fractions of the soil must be carefully treated because different countries use contrasting laboratory methods and standards, and the difference in values can often be considerable (>5% in relative terms). For example, soil organic carbon has historically been analyzed using a diversity of laboratory methods, including (Chatterjee et al., 2009; Shamrikova et al., 2022):

- Walkley Black method (WB);
- Tyurin method;
- Dry Combustion method (DC);
- Loss on Ignition (LOI);

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

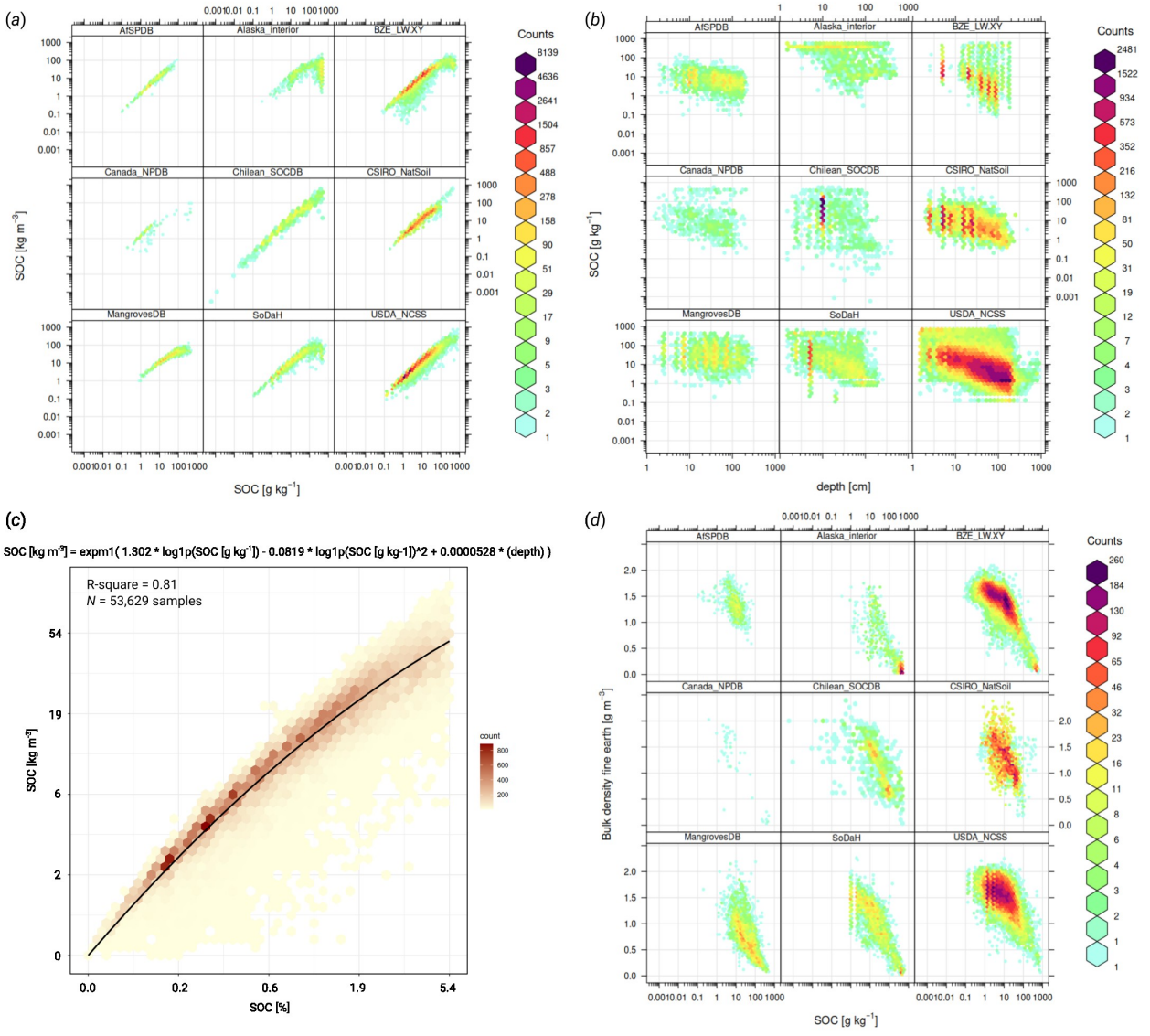

Figure 3. Comparison of imported soil laboratory data sets in terms of relationship between soil carbon density, carbon content, sampling depth, bulk density and others: (a) relationship between SOC content [g kg^{-1}] and SOC density (SOCD) [kg m^{-3}] is often close to linear, although this relationship is significantly more diffuse for organic soils; (b) soil carbon — depth plots usually indicate negative log-log relationship; (c) a global pedo-transfer function fitted using the highest quality laboratory data to gap-fill low SOC density [kg m^{-3}] values from SOC [g kg^{-1}] values; and (d) SOC [g kg^{-1}] and bulk density of fine-earth are also highly correlated and follow a bimodal distribution with one peak for mineral, and one for organic soils. AfSPDB = Africa Soil Profile Database (Leenaars et al., 2014), Alaska interior soil database (Manies et al., 2020), BZE-LW = Bodemzusandserhebung / German Agricultural Soil Inventory (Poeplau et al., 2020), Canada NPDB = Agri-Food Canada National Pedon Database (Geng et al., 2010), Chilean SOCDB = Chilean Soil Organic Carbon Database (Pfeiffer et al., 2020), CSIRO NatSoil (CSIRO, 2024), Mangroves soil database (Maxwell et al., 2023), SoDaH = the SOils DAta Harmonization database (Wieder et al., 2021), and USDA NCSS = National Cooperative Soil Survey Characterization Database.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

All four SOC determination methods can be considered compatible; however, values need to be corrected to a common standard, otherwise, this can lead to bias in total stocks. For example, the DC method, which is the current recommended standard for soil organic carbon (ISO 10694:1995), produced about 20–40% higher values than the WB method for the same samples. Locally, various groups have developed harmonization functions by analyzing the same soil samples using multiple laboratory methods (Chatterjee et al., 2009). Numerous harmonization studies producing functions and coefficients for translating SOC to the target laboratory method have been published in recent decades; however, these are often based on local data and therefore may not be globally applicable. Additionally, inter-laboratory comparisons analyzing samples from the same pedons have shown significant differences (Safanelli et al., 2023). This implies that the variation in the values of SOC, pH, and other soil properties comes in large part from short-range variability and the interlaboratory component, and not only from the harmonization strategy. Therefore, we have decided to use a simple harmonization principle described in Shamrikova et al. (2022):

$$\text{WB} \times 1.3 = \text{Tyurin} \times 1.15 = \text{DC} \quad (P = 0.95) \tag{4}$$

We have applied this harmonization to any SOC data set with the laboratory method explained in the metadata. Where metadata do not provide any information, we looked at the year of sampling and country of origin, and we estimated the laboratory method based on the indications from the literature. For most of the laboratory data (>90%) we had enough metadata to correctly determine the laboratory method used.

For carbon concentration and density values from the United States Soil Characterization Database (NCSS SCD) (United States Department of Agriculture and National Cooperative Soil Survey, 2023), several steps were taken to harmonize the different methods of estimating carbon, bulk density, and rock fragments. As carbon concentration measurement methods in NCSS SCD have shifted from WB to DC approaches (Soil Survey Staff, 2022), several regressions were used to harmonize all organic carbon measurements with WB to then integrate them into the larger global dataset by converting to DC using a previously fitted conversion model. Previous internal regressions that relate the SCD DC measurements to WB (Wills et al., 2013, 2014) have resulted in a contrasting relationship with the broader literature, so we decided to normalize all NCSS SCD carbon concentration values to WB to allow more widely accepted conversion equations to equivalent DC equations to be implemented. For all samples with DC total carbon (TC) estimates, we first regressed all samples with a 1:1 pH less than 7.4 (to exclude carbonates) against the WB measurements (WB = TC × 1.046, R^2 = 0.92, $N = 8,671$). This allowed all DC total carbon measurements with pH< 7.4 to be converted to WB units. Then, for additional samples with higher pH values that had DC organic carbon values, pre-adjusted for carbonates, we regressed those carbon values against WB again to convert them to a common unit (WB = DC × 1.037, R^2 = 0.90, $N = 175$).

In NCSS SCD, there are also more 1/3 bar bulk density (`BD.3`) measurements available than oven dry bulk density (`BDod`), therefore we also regressed these two methods to maximize our sample size (BDod = BD.3 × 1.102, R^2 = 0.99, $N = 90,230$). We also tested regression models with intercept values for both carbon- and bulk-density models. In all cases, zero intercept models had higher R^2 values, often by substantial amounts.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

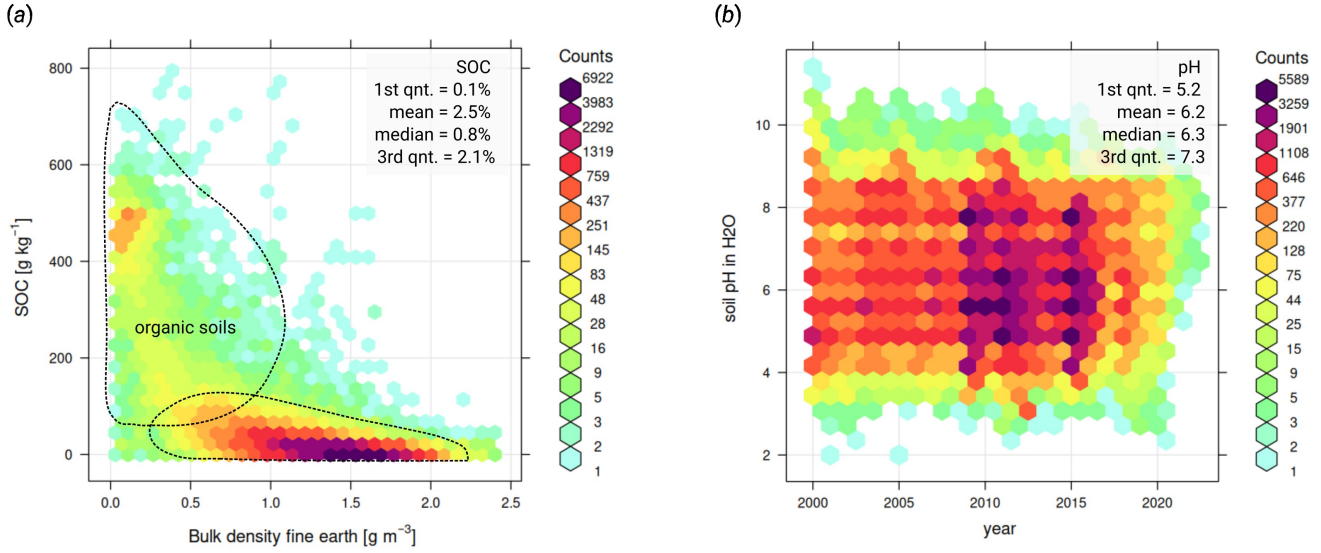

Figure 4. Visualized distributions of the final global harmonized soil organic carbon and soil pH data: (a) general relationship between soil organic carbon and bulk density with bimodal distribution of values (lines drawn by hand to illustrate overlap between organic and mineral soils), and (b) trend-plot showing, overall, no visible differences in soil pH distribution through time. Note that because SOC has a relatively right-skewed distribution, there is a significant difference in the mean value of SOC vs the median value. Soil pH, one the other hand, is already log-transformed and hence the mean and median values more or less match.

Finally, to adjust the carbon densities for rock fragments, we analyzed the US NCSS National Soil Information System (NASIS) for all SCD samples. The SCD rock estimates only include fragments less than about 4 inches in diameter, so we opted to use the NASIS field total rock volume estimates, which include all rock sizes. A rock density of 2.65 t m^{-3} was assumed for all samples. Similar corrections were applied to other L1–L3 data sets used in this work.

2.6 Insertion of pseudo-observations and gap-filling of missing values

Most legacy soil data sets in the world were not generated using probability sampling and/or strict experimental designs and, as such, are often not directly fit for spatial modeling (Hendriks et al., 2019). If we were to ignore that some areas are over-represented, resulting models fitted using such data could propagate potential bias in terms of, e.g., over-representation of agricultural land (Tian et al., 2024). That is why it is important to add covariate layers and additional points to assist machine learning models in producing predictions that also better match expert knowledge (Minasny et al., 2024).

Insertion of pseudo-observations is especially important for mapping chemical and physical soil properties, soil carbon stocks, as otherwise one could significantly over- or underestimate global stocks. Consider the following two examples: (1) the majority of soil surveys ignore taking samples from C horizons (parent material layer), semi-desert and shifting sand areas as it is obvious to surveyors that these contain no SOC; (2) mountainous areas, inaccessible areas such as swamps, jungles and

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

similar are also often under-represented due to inaccessibility. The world's deserts (polar deserts, Sahara, and similar) cover almost 33% the Earth's land surface: approximately 20% of the Earth's land surface are hot deserts; polar deserts (Antarctica and Greenland) cover another 10%. Very few soil surveys actually go to the middle of a desert or on top of a mountain to collect soil samples.

To avoid over-predicting SOC and under-predicting sand content for the world, we added pseudo-training points to help incorporate our soil knowledge in ML. To generate pseudo-points, we used primarily the GLANCE data set (Stanimirova et al., 2023), which is an extensive, quality-controlled point dataset covering years 1984 to 2020 and which is based on very high-resolution satellite images (usually about 30 cm resolution). We specifically used the classes *"Bare rock"* (5) and *"Shifting sand, deserts without any vegetation"* (6) as these are also easy to validate, and we believe that the risk of inserting erroneous pseudo-observations is low.

In addition to GLANCE points, we also used the global point data set with all major mountain tops (http://www.peaklist.org/ultras.html; about 1500 mountain tops), also to avoid generating extrapolation for the highest mountain chains, such as the Alps, the Himalayas, and similar. These areas are often under-sampled or not represented at all because they are extremely inaccessible. To avoid adding false 0 points for SOC, we double-checked the pseudo-observation points by overlaying them vs the 30 m resolution land cover map of the world GLC_FCS30D (Zhang et al., 2021). We only used the GLANCE point and the mountain tops that were also classified as *"190 Impervious surfaces"*, *"200 Bare areas"*, and/or *"220 Permanent ice and snow"*. In the end, this gave us 4680 high-quality pseudo-observations that are either permanent deserts, bare rock, or snow. At all these points, we have inserted 0 values for soil organic carbon (content and density), and in addition 100% sand content for all points classified as sand in the GLANCE data set.

Note that we insert pseudo-observations for modeling purposes — to better represent feature space, especially towards the edges of the feature space; however, after the modeling, we do not produce predictions for shifting sand areas and permanent ice as previously explained. This is for the following reasons: although we could have computed predictions for shifting sands and permanent ice, we believe that this would have increased production costs without adding significant value to the output maps. In addition, several covariates used for modeling are also often not accurate in such areas, potentially affecting the quality of the predictions. We, instead, advise users to gap-fill the maps using simple rules as indicated above or similar strategies (e.g. insert 0 SOC values and 100% sand content for shifting sands).

In addition to inserting 0 values for obvious shifting sand/deserts and bare rock areas, we also gap-filled around 5–6% soil carbon density points that only had SOC content but no bulk density. This was done by fitting a simple pedotransfer function (PTF) to estimate SOC [kg m^{-3}] directly from SOC [g kg^{-1}] measurements, avoiding estimating the bulk density that would be used to calculate SOC [kg m^{-3}]. We fit a bivariate quadratic function where the SOC density is a function of the SOC content and soil depth (shown in Fig. 3c), then use this function to fill in the missing values for the SOC density. We recommend using this PTF only for smaller values of SOC, e.g. <0.5% SOC, as the relationship for larger SOC values is of the order of magnitude more uncertain. In this work, we used this PTF to fill gaps for missing bulk density [kg m^{-3}] only where the SOC content is <0.4% or <4‰, as for this part of the range model is significant with R^2 >0.96. The relationship between the

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

density and content of SOC in soils with SOC >1% becomes proportionally more complex with eventually high uncertainty for SOC >10%, and therefore we recommend using this PTF only for small values.

2.7 Preparation of covariate layers

To integrate land use changes, soil management, and climate effects, we used more than 160TB of covariate data for modeling and prediction at 30 m resolution. The following four data sources are the largest in size and can be considered the most important:

- Landsat bimonthly and annual global composites described in Consoli et al. (2024) and derived products (Tian et al., 2025; Isik et al., 2025) (30 m spatial resolution);
- 6–scale Digital Terrain Model relief parameters described in Ho et al. (2025) (mix-scaled pyramid representation at 30, 60, 120, 240, 480, 960 m);
- CHELSA Climate time-series of climatic and bioclimatic variables v2.1 (Karger et al., 2017) (variable resolutions in kilometer scale);
- MODIS Land Surface Temperature MOD11A2 (https://doi.org/10.5281/zenodo.4527051) and Water Vapor data sets MCD19A2 (https://doi.org/10.5281/zenodo.8193738) (1 km spatial resolution);

From the Landsat archive, we use the Blue, Green, Red, NIR, SWIR1, SWIR2 bands, then also derivatives (biophysical indices) such as Normalized Difference vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Soil Adjusted Vegetation Index (SAVI), Bare Soil Index (BSI), Normalized Difference Tillage Index (NDTI), annual Bare Soil Frequency (BSF), Normalized Difference Snow Index (NDSI), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) and Gross Primary Productivity (Tian et al., 2025; Isik et al., 2025). Although we originally considered using the bimonthly values of all variables, winter months in the northern hemisphere and heavily clouded areas, like rain forests, have been shown to carry a significant amount of artifacts, which can propagate to predictions and lead to more serious artifacts. To avoid such issues, we decided to only use the lower (25%) quantile in the original bands instead of using bimonthly values or other quantiles. The decision to use the lower quantile comes from the fact that several artifacts originates from failing cloud mask, leading higher values in the raw bands, that are not impacting the lower quantiles. To keep a single consistent model, the usage of the lower quantile is applied at global scale, and not only in the areas with artifacts. However, it is possible that the prediction accuracy of soil soil properties could be further increased with further improvements in the Landsat composites.

From the DTM variables, we use 6–scale DTM global parameters derived at pixel resolutions of 30 m and of 60, 120, 240, 480, and 960 m, which were later resampled to 30 m using cubic splines. The DTM variables include terrain height, slope in degree, multidirectional hillshade, topographic wetness index, negative/positive openness, LS factor, minimum, maximum, profile, tangential, and ring curvature. This type of multi-scale nested terrain derivation is known as *"Mixed scaled Gaussian Pyramid"* (Behrens et al., 2018), designated to capture spatial dependencies and interactions of the landscape and soil at various scales. Relationships between different soil properties and terrain change at different scales are often in a non-linear

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

way. Hence, we prepare standard scales (microscale, mesoscale, and macroscale) of DTM derivatives that allow ML to model complex relationships and select an optimal set of the terrain representation.

Beyond the above listed-layers, we also use: peatland extent ensemble estimate (https://doi.org/10.5281/zenodo.13951438), bare rock extent based on the Local Climate zones map (Demuzere et al., 2022), forest and wetlands cover based on ESA CCI (https://doi.org/10.5281/zenodo.13951438), crop cover based on GLAD time-series 30 m (Potapov et al., 2022), World Karst Aquifer Map (WHYMAP WOKAM) (Chen et al., 2017), sediment types based on GUM v1.0 (Börker et al., 2018), bare soil fractions (mean and maximum) and photosynthetically active vegetation fractions based on GVFCP v3.1 (https://doi.org/10.5281/zenodo.11961219), Global WaterPack annual water extent probability (250 m) (Klein et al., 2017), snow probability P90 long-term MODIS-based (https://doi.org/10.5281/zenodo.5774953), soil salinity grade (250 m) (Ivushkin et al., 2019), Global Soil Bioclimatic variables (Winkler et al., 2021), geometric temperature, landform class based on the USGS EcoTapestry, and MERIT Hydro upstream area (Yamazaki et al., 2019). Because the Global Soil Bioclimatic variables are also based on SoilGrids (sand, silt, clay predictions) and soil salinity grade is also based on soil property predictions, we use these layers only for soil type mapping and not soil property mapping, to avoid possible circularity in the models.

For quantitative soil properties, we also use soil depth (center of the sample horizon) as a covariate. This means that all such models are 3D+T, i.e. we fit one model per property that can be used to predict values for any year and for any depth. As further detailed in the following, predictions are then averages over spatio-temporal blocks of five years (e.g. 2000–2005) and variable depths interval (0–30, 30–60 and 60–100 cm).

In summary, we used a total of 363 covariate layers for mapping soil properties and soil types, either as time series of bimonthly/annual images from 2000–2022+, long-term estimates of climate, or assumed static variables (DTM parameters, lithology types, and similar). For soil type mapping, we used a much smaller number (229) of covariate layers because we excluded all time-varying layers, and hence only long-term estimates of climate, vegetation and similar are used. Not all layers were used in the final prediction as the feature selection process would typically reduce the number of initial number of layers to 60–120, removing layers that marginally contributed to the final model.

2.8 Variables transformation for soil properties

Soil organic carbon models, both content and density, properties were transformed into a natural log (with offset = 1, `log1p()` R function) to improve the prediction performance of soil properties with a highly skewed distribution (Dangal et al., 2019). Predictions were then back-transformed (`expm1()` R function) in the original space. We report error metrics for log-normal variables in both original and transformed spaces.

Soil texture fractions are transformed using a modified version of the additive log-ratio (ALR) transform, that for the forward transformation reads

$$\begin{aligned} \text{Texture}_1 &= \log_2\left(\frac{\frac{\text{Sand}}{a}+1}{\frac{\text{Clay}}{a}+1}\right), \\ \text{Texture}_2 &= \log_2\left(\frac{\frac{\text{Silt}}{a}+1}{\frac{\text{Clay}}{a}+1}\right). \end{aligned} \tag{5}$$

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

where a is a normalization factor corresponding to the summation value of the three fractions (e.g. 100 if they are represented in %). The new transformation removes the singularities that are present in the ALR transformation if one or more of the textures fractions is equal to zero. Furthermore, the usage of $\log_2$ and the normalization by a of each texture in the forward transformation guarantees that both variables in the transformed space are in the range -1 and 1. For the data collected in this work, the distributions of Texture_1 and Texture_2 are close to a uniform distribution and a normal distribution, respectively. These properties facilitate the modeling phase compared to having skewed, sparse distributions or numerically noisy values that were observed using standard ALR transform. Texture_1 and Texture_2 are then modeled and predicted separately. This leads to no guarantees that after applying a straightforward back-transformation to the predictions, the texture fractions would sum up to a, nor that each of them is greater than or equal to 0. For this reasons, the back-transformation applied to the prediction is slightly modified to guarantee that these constraints are respected, and it reads:

$$x_1 = 2^{\text{Texture}_1}, \quad x_2 = 2^{\text{Texture}_2}, \tag{6}$$

$$C = \max\left(0, \frac{3 - x_1 - x_2}{1 + x_1 + x_2}\right), \tag{7}$$

$$S = \max\left(0, x_1 C + x_1 - 1\right), \tag{8}$$

$$L = \max\left(0, x_2 C + x_2 - 1\right), \tag{9}$$

$$T = S + L + C, \tag{10}$$

$$\text{Sand} = \frac{a}{T} S, \quad \text{Silt} = \frac{a}{T} L, \quad \text{Clay} = \frac{a}{T} C. \tag{11}$$

2.9 Model calibration and prediction of soil properties

For each property, the data set is first partitioned into three subsets: (1) calibration, (2) training, and (3) stratified test sets, with an approximate ratio of 1:8:1. The calibration set is used for feature selection and hyperparameter tuning, the training set for model development, and the hold-out test set for final evaluation. The hold-out test set is not used for any other purpose but for accuracy assessment. When the data set is large, to prevent the excessive data volume from skewing the process, we cap the calibration and test set sizes at 8,000 and 6,000 samples, respectively. For the calibration and test sets, we use spatial subsetting with a standard density of points per 100 km by 100 km tile (for example maximum 2 points per tile). This ensures that the overall density of points is standard and that there are no geographical groups (Roberts et al., 2017), similarly to the approach used in Poggio et al. (2021). The data set partitioning scheme is represented in Fig. 5.

For feature selection, we use Repeated Subsampling-Based Cumulative Feature Importance (RSCFI), a variant of Recursive Feature Elimination with Cross-Validation (RFECV) (Wadoux et al., 2020). RSCFI optimizes model performance while efficiently eliminating less relevant covariates, achieving results comparable to those of RFECV. Hyperparameter tuning is performed using `HalvingGridSearchCV` (Pedregosa et al., 2011), a resource-efficient approach that iteratively narrows down the best parameter combinations, optimizing the Lin's concordance correlation coefficient (CCC).

After calibration and accuracy assessment, the whole dataset was used to train the **Tree-Based Quantile Regression Forest** (TB-QRF) and the RF models. We used the compiled versions of these models to produce predictions at 30 m resolution. In

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

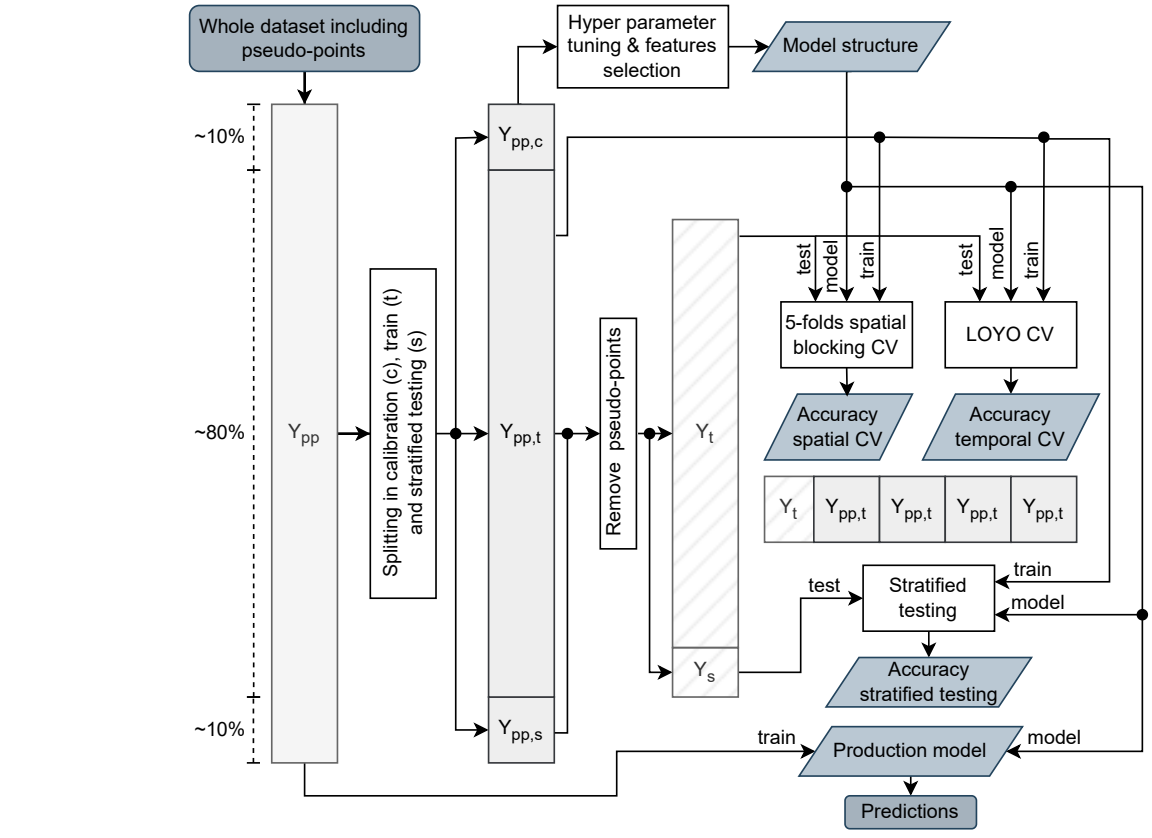

Figure 5. Schematic partitioning of the soil properties dataset. For each property. the whole dataset including pseudo-points (Y_{pp}) is divided in calibration ($Y_{pp,c}$), training ($Y_{pp,t}$), and stratified test ($Y_{pp,s}$) sets, with an approximate ratio of 1:8:1. The calibration dataset is used to perform hyper parameters tuning and features selection. The optimized model structures is trained with the $Y_{pp,t}$ set under three different validation setups: stratified testing, 5–folds spatial blocking CV and leave-one-year-out (LOYO) CV. In all the testing phases of the validation, the pseudo-points were removed from the test sets, so using Y_s or splits of Y_t. The obtained results are used to derive all the reported accuracy metrics. The whole dataset is instead used to train the final model used for predictions.

addition, we used the non-compiled version of the models to retrieve the single-tree outputs to produce 120 m resolution maps that also include quantiles 0.16 and 0.84 for uncertainty estimation. The entire pipeline has been developed using open-source code and integrated into the scikit-map library (https://github.com/openlandmap/scikit-map).

The predictions are run per 1° by 1° (∼120 km by ∼120 km) tiles using parallel computing over 10 CPU servers and by reading from 17 storage servers to the central storage data lake (based on SeaweedFS file service). The world land mask can be represented with about 18,000 120 km tiles. After predicting target variables per tile, global mosaics are built using GDAL

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

to produce complete, consistent Cloud-Optimized GeoTIFFs, one global scale (whole land mask) file for each combination of variable, time-frame, depth-range, and mean or quintile. Prediction is the most costly part of the data production, with each soil property taking at least 3 days of HPC with about 1500 threads and 14TB of RAM to produce space-time predictions of a single soil property. The final output mosaics contain variable type, reference method and measurement units, depth interval, and reference begin / end year in the file name (see https://github.com/openlandmap/soildb for more details).

2.10 Derivation of prediction uncertainty

To produce per-pixel uncertainty, we use the TB-QRF, where the output of each single tree in the RF has been used to derive the prediction intervals (Meinshausen, 2006), following the scheme described in Fig. 6. Note that compared to other QRF the distribution is obtained from the tree outputs and not from the single leafs. We decided to predict the quantiles 0.16 and 0.84 to lead to a 68% interquartile range (IQR). assuming a Gaussian distribution, 68% interval corresponds to the ± 1 standard deviation; to derive 1 standard deviation from the lower and upper intervals, users should calculate the range and then divide by 2. In addition, compared to 90% or 95% IQRs, this allows us to have a smaller number of trees (e.g. 64) in the RF without leading to artifacts in case of noisy trees or covariates that are generally in the extremes of the distribution, and therefore speed-up computing. For variables with more complex distributions, for example, log-normal, gamma, or multi-modal distributions, dividing the upper minus lower range by 2 should be used with caution, as also the prediction distributions per pixel are often skewed, and hence the true errors might not match the approximated 1 std.

We produce in memory point predictions for the years 2000, 2005, 2010, 2015, 2020, and 2022, and soil depths 0 cm, 30 cm, 60 cm, and 100 cm; these are then averaged in 2 by 2 spatio-temporal blocks that represent timeframes of 5 years with one year overlap (excluding the last timeframe) and variable depth ranges:

$$\text{spb}_{i,j} = \frac{P_{i,j} + P_{i+1,j} + P_{i,j+1} + P_{i+1,j+1}}{4}, \tag{12}$$

following the notation in Fig. 6. We decided to use block predictions as most of the users require predictions of soil properties per standard depth intervals. We also block-predictions in time primarily to reduce interannual variability, especially interannual oscilations coming from Landsat-derived indices. Note that while depth is a feature of the model, the prediction year is not. However, prediction year was used to determine with specific layers to use in time-dependent features, so the models are fully temporally consistent.

To derive both predictions and prediction uncertainty on a global scale, we used a hybrid Python/C++ implementation of TB-QRF and RF Python / C++ where the models are fitted using the scikit-learn library in Python, then translated to C++ source code and compiled using tl2cgen. Spatio-temporal overlay and predictions were also performed using C++ interfaced with Python within the scikit-map library. Finally, although TB-QRF is a fairly robust method and is applied to all regression problems, sometimes it can over- or under-estimate actual prediction errors, and hence we also test the accuracy of the prediction intervals using the procedure described in Tian et al. (2024).

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

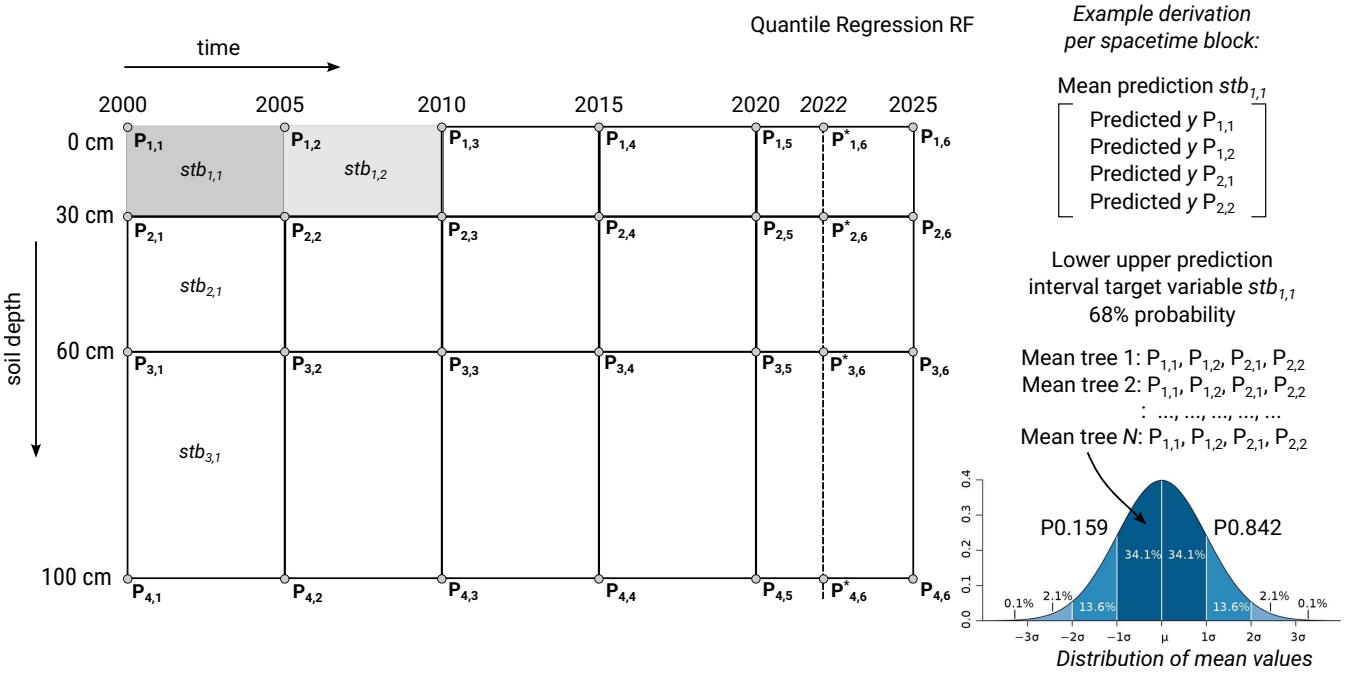

Figure 6. Spatio-temporal prediction blocks scheme: predictions are generated for four space-time points, then averaged to derive mean prediction and lower and upper prediction intervals with 68% probability interval. Note that currently no ARD Landsat data is yet available for 2023–2025, hence for the last period the block support is <5 years.

2.11 Model calibration and prediction for soil types

For modeling and mapping soil types, we also use the RSCFI framework, but with the difference that we develop two models: RF and LightGBM (Ke et al., 2017) models; final predictions are then generated as an ensemble model by averaging the two. The approach is based on fitting models to features that are potentially valuable and selecting them based on the mean decrease in impurity. To enhance robustness, the model was trained 50 times, each time using a different bootstrap-sampled subset (80%) of the calibration dataset, selected by spatial blocking (100 by 100 km blocks). Features below the mean importance threshold were discarded in each iteration. To optimize computational burden, we selected 100 features that were consistently repeated in at least 25 model runs in both models. The final selected features were applied to the calibration and validation data sets before hyperparameter tuning.

2.12 Cross-validation and quality control

We decided to run the evaluation of soil properties models in three different modalities: (i) on a test set derived from stratified sampling based on Köppen–Geiger climate classification from CHELSA V.2.1 (Karger et al., 2017), (ii) with a 5–fold spatial blocking CV with 100 km by 100 km tiles, and temporal CV using the leave-one-year-out (LOYO) approach (Fig. @ 5). We

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

consider the results of the accuracy assessment using the test set to give a *optimistic* estimate of the mapping accuracy and the results of temporal and spatial CV to give a *pessimistic* estimate: we expect that the actual accuracy is between the two numbers.

For each model, we report RMSE, mean error (bias), R-squared (R^2), Lin's concordance correlation coefficient (CCC), defined as:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \\ bias &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \\ CCC &= \frac{2 r s_y s_{\hat{y}}}{(\bar{y} - \bar{\hat{y}})^2 + s_y^2 + s_{\hat{y}}^2}, \end{aligned} \tag{13}$$

and the fraction of Tweedie deviance explained (D^2) (Hastie et al., 2015; Pedregosa et al., 2011), calculated as:

$$\begin{aligned} \mathrm{d}(y_i, \hat{y}_i) &= \begin{cases} (y_i - \hat{y}_i)^2, & p = 0 \quad \text{(Normal)} \\ 2\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - y_i + \hat{y}_i\right], & p = 1 \quad \text{(Poisson)} \\ 2\left[\log\left(\frac{\hat{y}_i}{y_i}\right) + \frac{y_i}{\hat{y}_i} - 1\right], & p = 2 \quad \text{(Gamma)} \\ 2\left[\frac{y_i^{2-p}}{(1-p)(2-p)} - \frac{y_i \hat{y}_i^{1-p}}{1-p} + \frac{\hat{y}_i^{2-p}}{2-p}\right], & \text{otherwise} \end{cases} \\ D^2 &= 1 - \frac{\sum_{i=1}^{n} \mathrm{d}(y_i, \hat{y}_i)}{\sum_{i=1}^{n} \mathrm{d}(y_i, \bar{y})}, \end{aligned} \tag{14}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean value, n is the total number of samples, r is the Pearson correlation between y and $\hat{y}$, s_y^2 is the variance of the observed values, $s_{\hat{y}}^2$ is the variance of predicted values, and $\bar{\hat{y}}$ is the mean of predicted values. Finally, to asses performance in quantifying uncertainty, we also report Prediction Interval Coverage Probability (PICP), computed as the ratio of actual values that reside inside a model's estimated confidence intervals for the corresponding predictions.

2.13 Spatial dependence analysis for residuals

To evaluate the spatial structure of the prediction residuals, we computed empirical semivariograms of the absolute prediction errors. In an operational setting, this means that variograms are fitted per each of the six continents (Antartica is excluded). Prediction errors were obtained through 10–fold cross-validation explained in the previous sections. The coordinates were reprojected to continent-specific azimuthal equidistant projections (Equi7) to assist in the distance calculation (Bauer-Marschallinger et al., 2014). For each continent, pairwise distances and squared differences in prediction errors were computed, and the empirical variogram was derived by binning these differences in 5 km distance intervals, up to a maximum lag of 125 km. A Locally

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Weighted Scatterplot Smoothing (LOWESS) smoothing line was fitted to the binned semi-variance estimates to visualize spatial trends. To support interpretation, we also fit spherical variogram models to data within a truncated spatial range of 125 km. GLanCE pseudo-observations were excluded from the analysis to avoid distortion of spatial dependencies.

2.14 Soil property change analysis against land cover change

To compare changes in soil properties for 2000 to 2022+ versus land cover change, we used a total of 12,500 unique spatial locations sampled following the strategy described in Hackländer et al. (2024). This is a point data set generated using the stratified random sampling approach and excluding areas covered by permanent water or ice (Brus, 2022). Each sampled point was overlaid with predicted maps of SOCD and pH for the periods 2000–2005 and 2020–2022, as well as the ESA CCI land cover maps for the years 2000 and 2020 (ESA, 2017). Based on this overlaid dataset, spatially matched changes in SOCD and pH were derived and linked to the corresponding land cover transitions for analysis and visualization. For each change class (e.g. broad-leave forest to pasture), we derive the mean SOCD and soil pH change value and the distribution of values. These values are then reported and sorted to see which land use change categories result in larger changes in soil properties, i.e. to detect which are the key drivers of change.

2.15 Extrapolation risk assessment

Extrapolation often leads to decreased performance in machine learning models, but it is an unavoidable aspect of large-scale spatial mapping. Several methods exist to identify predictions made in dissimilar feature spaces, including the Area of Applicability (AOA) (Meyer and Pebesma, 2021), Isolation Forest (Liu et al., 2008), and Homosoils (Nenkam et al., 2022). Given the extensive spatial scope and computational demands of this study, we selected the Isolation Forest algorithm due to its efficiency and suitability for non-normally distributed multivariate datasets (Liu et al., 2008). Isolation Forest detects regions of the feature space that differ from the training data by recursively partitioning the dataset and isolating individual samples. It works by constructing an ensemble of randomly generated trees and calculating an anomaly score based on the average path length required to isolate a sample. Samples located in low-density or unfamiliar regions of the feature space generally require fewer partitions, resulting in shorter path lengths and thus higher anomaly scores. We employed the `ensemble.IsolationForest` implementation from scikit-learn (Pedregosa et al., 2011) to generate these scores. The average path length within the training data set serves as a baseline threshold to distinguish between *in-sample* and *out-of-sample* predictions (Liu et al., 2008). To effectively communicate the extrapolation risk to users, we normalize the anomaly scores to a scale 0–1, where higher values represent greater extrapolation risk for a given sample or pixel. The threshold separating the *in-sample* and *out-of-sample* regions was similarly rescaled to this normalized scale, ensuring consistency with the extrapolation risk probability maps delivered to end users.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025


Open Access Earth System Science Data Discussions

3 Results

3.1 Harmonization of training data

After multiple rounds of import, binding, and internal tests, we finally prepared about 216,000 soil samples with soil carbon density (kg m^{-3}), 408,000 soil samples with soil carbon content (g kg^{-1}), 272,000 samples with soil pH in H_2O, 363,000 samples with clay, silt, and sand (%), and 134,000 samples with bulk density oven dry (t m^{-3}), which we consider to be analysis-ready. The additional samples from pseudo-observations from the PTF helped us increase the number of training points for mapping the SOC density from 227,000 to 305,820. The final density of the training points prepared for the soil carbon, pH, soil texture fraction mapping, and soil type mapping is shown in Fig. 7. Compared to some previous global modeling attempts (Poggio et al., 2021; Padarian et al., 2022b), our training data is harmonized to a single standard, e.g., DC method for SOC and try to equally represent the diversity of biomes and land use systems: from agricultural soils, forests, and specific biomes such as tropical peatlands to mangrove forests. The final harmonized points are available via https://soildb.openlandmap.org (the publicly available data; exclude LUCAS soil samples and similar) and will be continuously updated.

3.2 Accuracy of soil properties predictions

Results of validation using the stratified test data (hold-out samples) show RMSE of 17.7 [kg m^{-3}] (0.486 in log-scale) and CCC of 0.88 for SOC density, RMSE of 51.3 [g kg^{-1}] (0.574 in log-scale) and CCC of 0.87 for SOC content, RMSE of 0.15 [t m^{-3}] and CCC of 0.92 for bulk density of fine-earth, RMSE of 0.51 and CCC of 0.91 for soil pH, RMSE of 8.4% and CCC of 0.87 for soil clay content, and RMSE of 12.6% and CCC of 0.84 for soil sand content respectively. These accuracy levels match or exceed the accuracy levels reported in Poggio et al. (2021). Our predictions appear to be potentially more accurate for soil pH (our results RMSE 0.51 vs. 0.77), bulk density (our results RMSE 0.15 vs. 0.19), and texture fractions (our results RMSE 8.4% vs. 13% for clay content). Note that RMSE as an accuracy metric for log-normal / skewed variables is of limited use and probably should be avoided as RMSE is highly sensitive on few high values (e.g., organic soils); hence we are not able to compare our results to the results from SoilGrids V2 for SOC content. Based on the D^2 metric (distribution independent), the best performing variables appear to be soil pH, bulk density, and texture fractions, but all numbers are in principle comparable and in the range 0.70–0.85 for the holdout samples. We recommend to other groups to also report their D^2 metric as this seems to be distribution-independent; in the case of log-normal variables, we recommend estimating RMSE also in the log-space (natural logarithm).

The PICPs for the target prediction interval (68%) for the SOC density, SOC content, bulk density and pH models are respectively 63%, 67%, 38% and 57%. While for SOC density and content the values are quite close to the ideal scenario, pH and in particular bulk density, the PICPs are quite smaller than the target PI. This motivated us to also check the quantile coverage probability (QCP), from which we can see that for the pH, the difference is reasonable and symmetrically deriving from upper and lower quantiles. For bulk density instead, the lower density is drastically off, and only converging to good performance around PI 90%. In future versions we will focus on improving the PICP for bulk density. Finally, the PICPs for textures fractions are 57%, 64% and 57% for sand, sild and clay, respectively.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

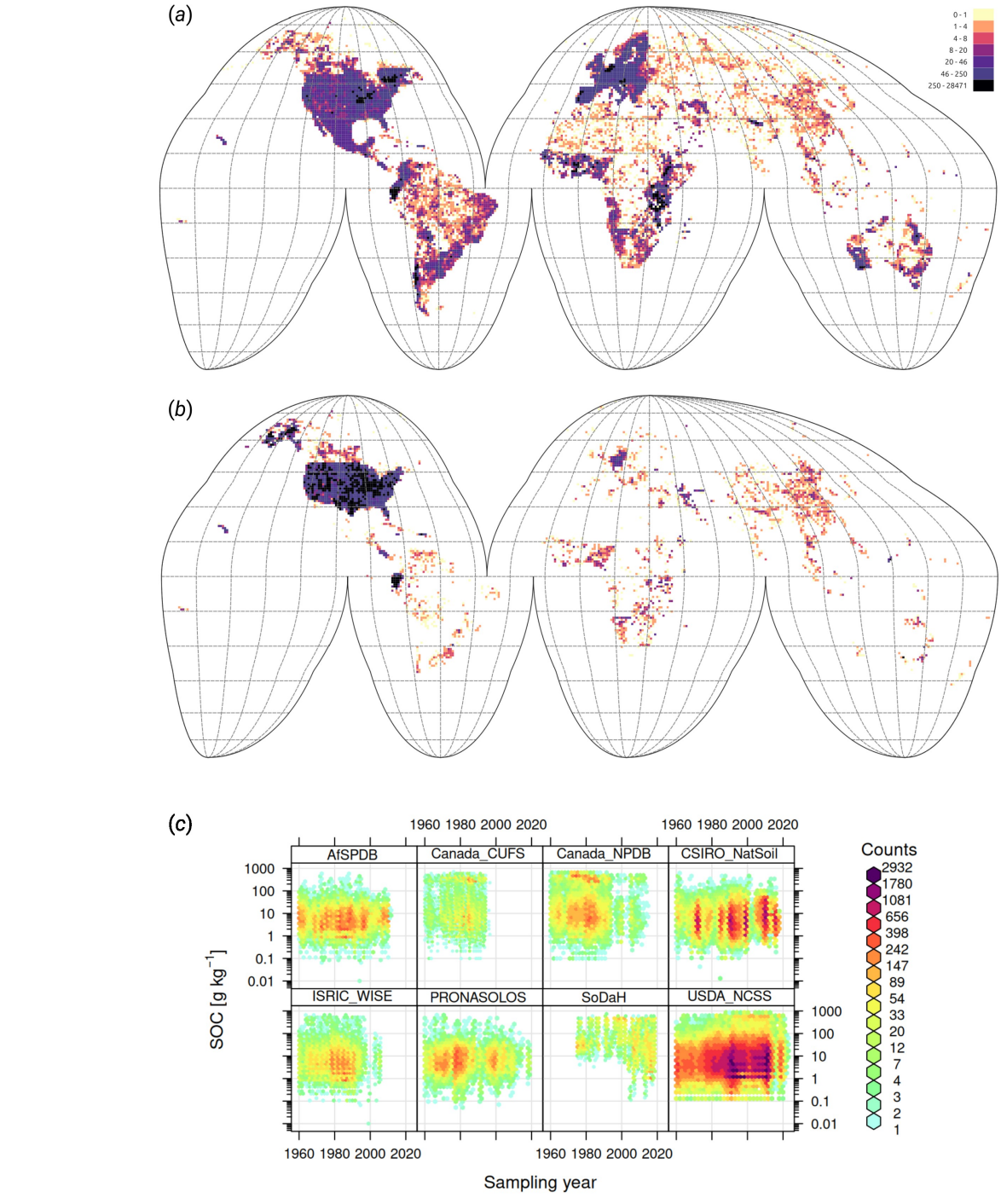

Figure 7. Density of training points used to build global predictive mapping models, and quality control plots for a number of key variables: (a) soil samples with soil organic carbon and/or soil pH, (b) soil profiles with soil taxonomy class, and (c) temporal coverage of samples from several larger datasets. Only points collected after year 1999 were used for modeling soil properties. For soil type mapping and to match high resolution covariates, we prioritize using points that are collected with GPS accuracy.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

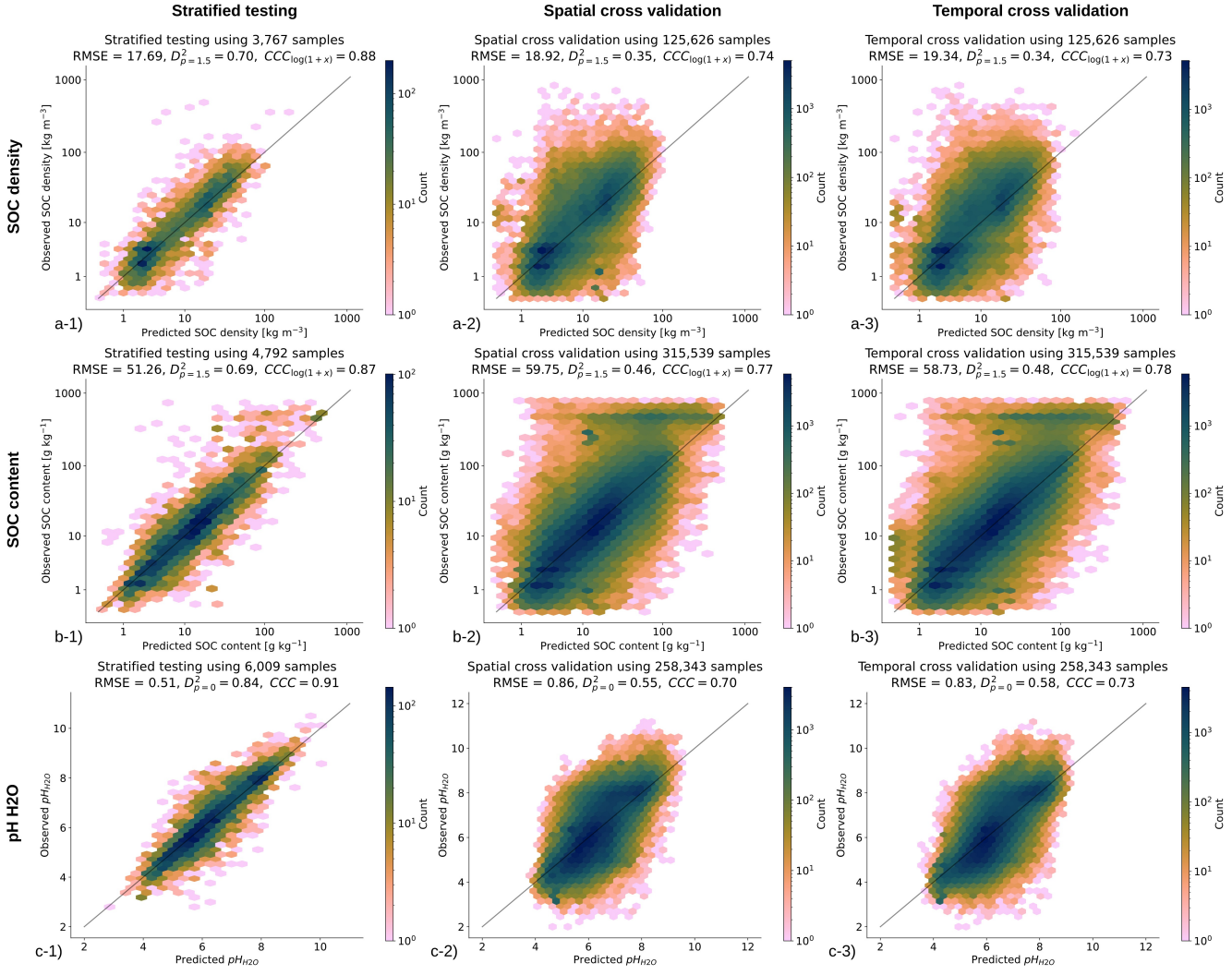

Figure 8. Accuracy plots for (a) soil organic carbon density [kg m^{-3}], (b) soil organic carbon content [g kg^{-1}], and (c) soil pH H_2O based on the (left) stratified testing set, (center) spatial cross-validation, and (right) temporal cross-validation (LOYO). CCC for SOC density and content are derived in log-scale; RMSE based on stratified testing for SOC density and SOC content in log-scale is 0.486 and 0.574 respectively.

The accuracy results for different validation strategies are shown in Fig. 8. These show a clear difference between stratified testing and spatial CV (with blocking), which was also expected. In general, we consider that temporal and spatial CV give the most pessimistic accuracy results and stratified testing gives independent results, but because we do not really have a probability sample, we consider these results potentially over-optimistic. For example, for predicting SOC density, CCC is between 0.73–0.88; for soil pH RMSE is between 0.51 and 0.83. The difference in D^2 for all variables between stratified hold-

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

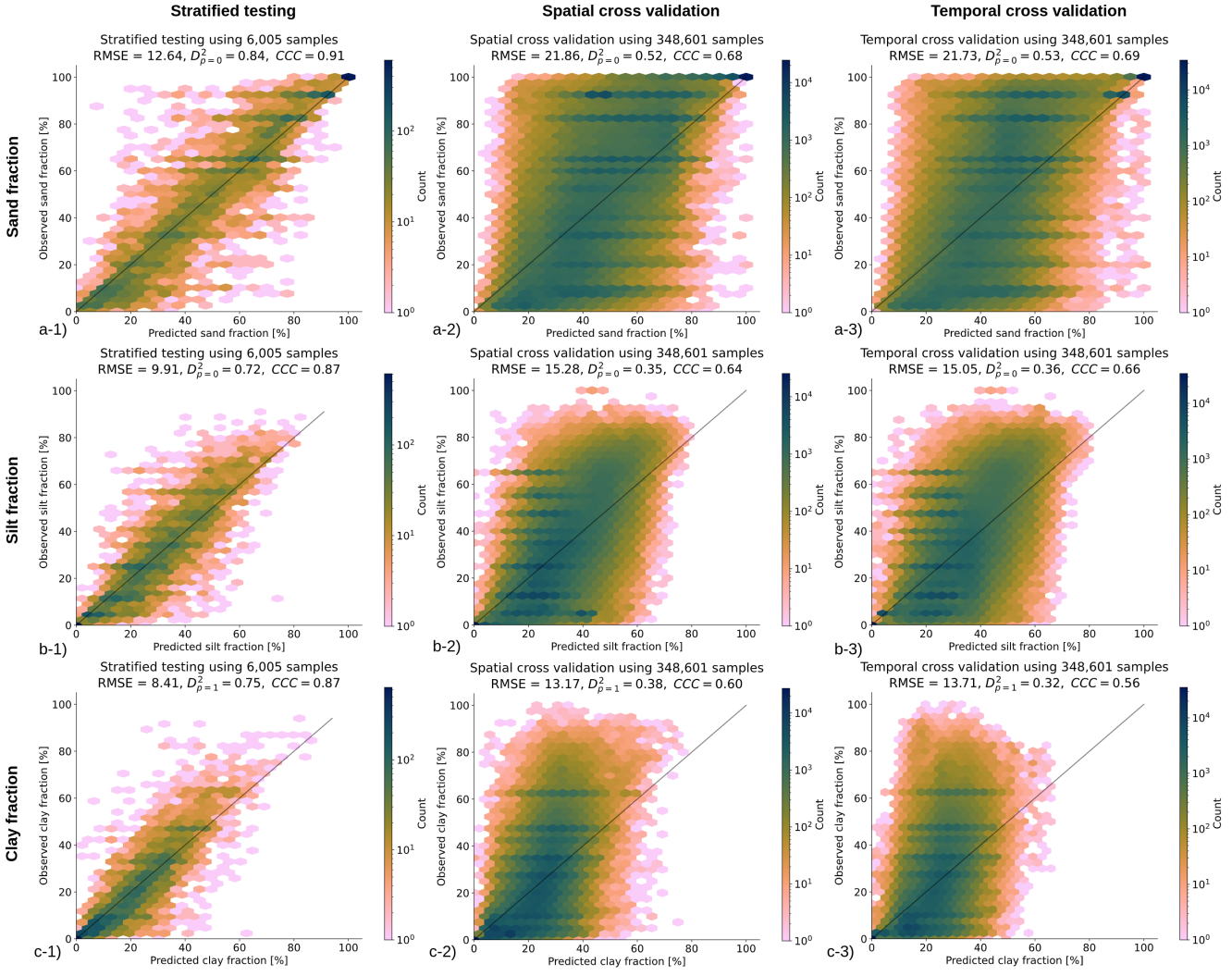

Figure 9. Accuracy plots for (a) sand fraction [%], (b) silt fraction [%], and (c) clay fraction [%] based on the (left) stratified testing set, (center) spatial cross-validation, and (right) temporal cross-validation (LOYO).

outs and spatial blocking appears to be the largest, with values for SOC density, for example, ranging between 0.68–0.84. It is interesting to observe that the temporal CV achieves accuracy similar to that of the spatial CV indicating that indeed models fitted over a longer period of years (25+ years) can be used to predict also new years e.g. 2025, 2026 for which we maybe have no new training points. The predicted soil property maps also show very gentle changes with most of pixels ($> 90\%$) not changing much from period to period.

Table 1 shows the results of the accuracy assessment for the target soil properties for different standard depths: 0–30, 30–60 and 60–100 cm. These results indicate that, as expected, the highest accuracies for SOC density and content are achieved for

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

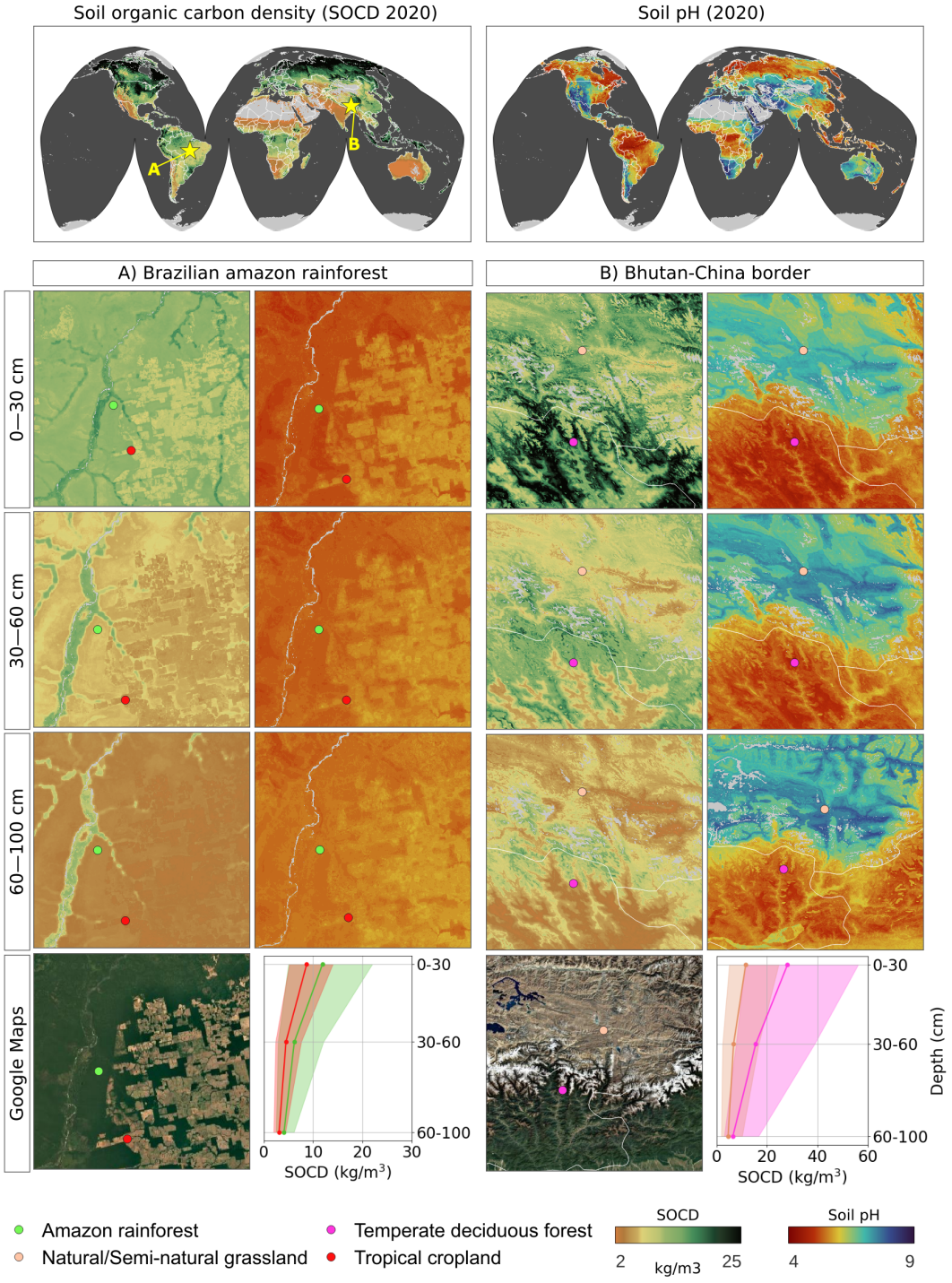

Figure 10. Predicted soil organic carbon and soil pH at 30 m resolution with zoom-in on two sample areas, with corresponding satellite images from Google Maps 2025. Soil-depth plots indicate 68% probability prediction intervals based on the Quantile Regression Random Forest.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Table 1. Model performance for SOCD, SOC content, BD, and pH_{H2O} across different depth intervals, calculated on the testing set. The values signed with * are computed in the $\log(1+x)$ space, as illustrated in Fig. 8. Note that "All points" can include also points that are deeper then 100 cm.

Property	Depth (cm)	RMSE	CCC	D^2	R^2	bias
SOCD	0–30	18.7	0.840*	0.617	0.729*	-3.83
	30–60	20.1	0.780*	0.615	0.648*	-2.26
	60–100	13.3	0.758*	0.569	0.621*	-2.03
	All points	17.7	0.882*	0.700	0.792*	-2.85
SOC	0–30	57.8	0.816*	0.665	0.699*	-10.4
	30–60	44.4	0.784*	0.637	0.658*	-7.11
	60–100	25.5	0.726*	0.477	0.590*	-4.07
	All points	51.3	0.866*	0.685	0.768*	-8.56
BD	0–30	0.141	0.916	0.846	0.846	-0.00160
	30–60	0.170	0.893	0.809	0.809	0.00296
	60–100	0.157	0.913	0.845	0.845	-0.00367
	All points	0.148	0.916	0.847	0.847	-0.00209
pH_{H2O}	0–30	0.528	0.895	0.814	0.814	0.00883
	30–60	0.444	0.926	0.867	0.867	-0.000479
	60–100	0.490	0.922	0.863	0.863	0.0203
	All points	0.508	0.908	0.836	0.836	0.00484

the top soil: CCC drops from 0.84 to 0.76 going from 0–30 cm to 60–100 cm. However, the difference in accuracy between depths in our results appears to be in general minor, with most values oscillating ±5–10% between different depths (Table 1). This is a somewhat unexpected result, although for SOC density and similar the values at higher depth are also significantly lower, so possibly this is why the errors are also in average lower even though models are typically based on much less points than what is available for top-soil.

Our results of cross-validation also show some bias in predicting SOC content and SOC density and clay content, with our models potentially over-predicting smaller SOC values and under-predicting higher clay content. This indicates that it is important to use prediction intervals (we provide lower and upper prediction intervals as maps as shown in Fig. 10) together with predictions to incorporate the uncertainty of these models.

Semivariograms representing spatial autocorrelation of model residuals for SOCd are shown in Fig. 11. Except for North America, residuals show either no spatial autocorrelation structure, or spatial dependence at shorter distances i.e. up to maximum 10–20 km. Considering that only a fraction of the points are available at distances of <10 km. We hence do not consider

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

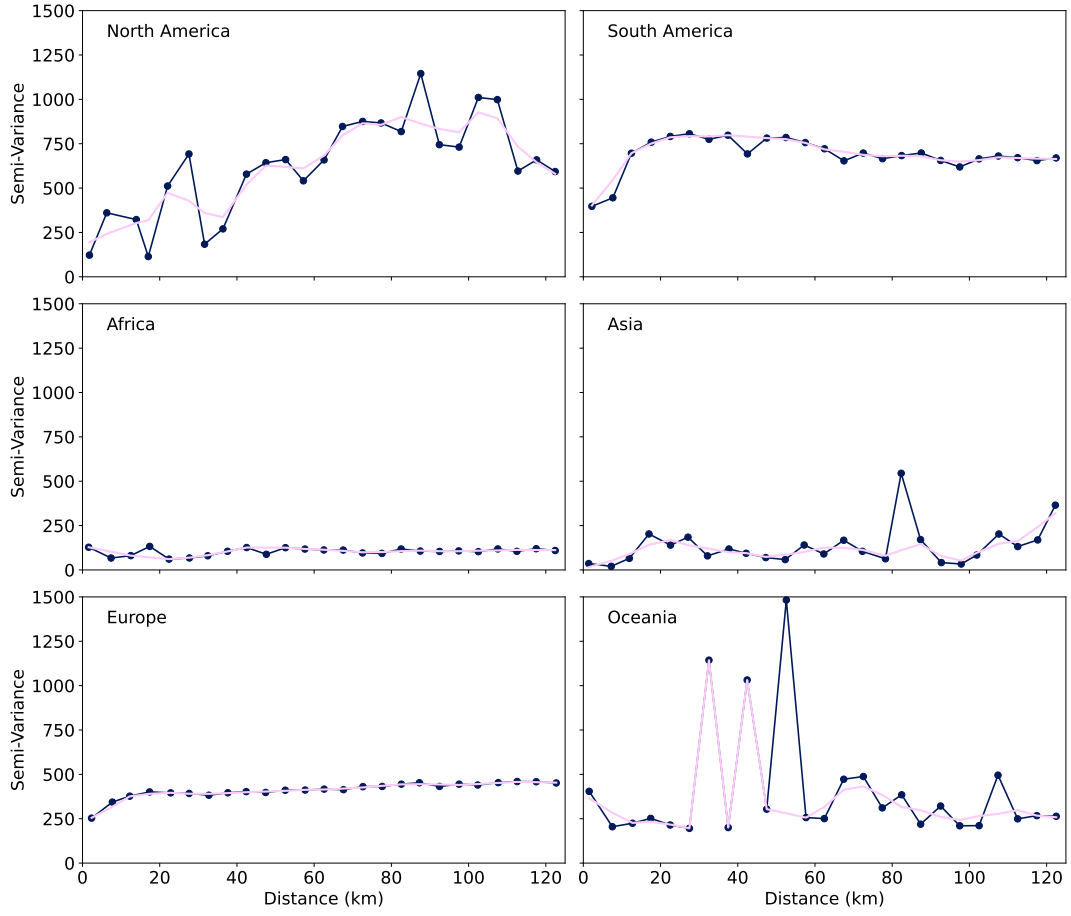

Figure 11. Sample semi-variograms of SOCd prediction residuals across six continents (10–fold CV), with distances computed using continent-specific equal-area projections. Binned every 5 km up to 125 km (dark blue dots an line), smoothed by LOWESS (pink). GLanCE points (pseudo-observations) were excluded.

kriging of residuals for these data, although for further merging with local data combining variogram modeling with RF could help increase accuracy.

3.3 Key covariates explaining global distribution of targeted soil variables

Results of variable importance for the soil variables of interest are shown in Fig. 12. For SOC density, it is especially interesting to see that Landsat-derived GPP (30 m resolution, bimonthly aggregated to annual) comes in the top three most important covariates (see R1). Conceptually speaking, we expect that primary productivity is the key source of SOC, at least in natural vegetation systems. As expected, depth explains almost 30% of variability in the SOC distribution and is distinctly at the top, which justifies the use of soil depth as a covariate.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

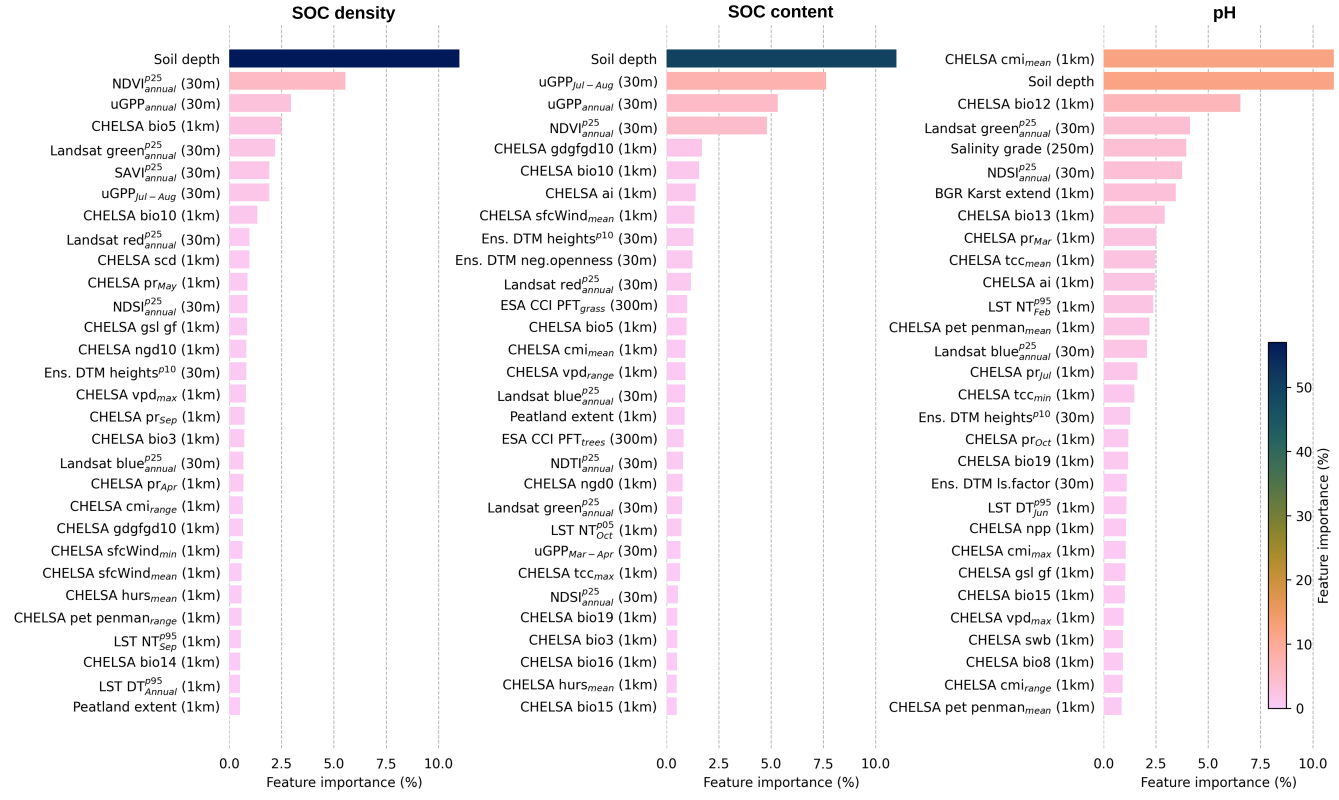

Figure 12. Variable importance plots for soil organic carbon density, soil organic carbon content, and soil pH.

The global distribution of soil pH can be primarily explained by the CHELSA Aridity Index (long-term), annual precipitation, and the grade of salinity (Ivushkin et al., 2019). The correlation plots in Fig. @ 13 show how the top 4 most important variables listed for SOC density relate in 1:1 density plots: higher GPP / higher vegetation index and cooler climates convert to higher SOC. SOC density and soil depth are close to linearly correlated on a log-log scale, as are SOC density and GPP. This also illustrates that the uncertainty of individual driving factors is still relatively high.

3.4 Accuracy of soil type predictions

The results of the accuracy assessment using spatial blocking for the soil subgroups (818) show that, as expected for this high number of classes, the F1 score does not exceed 0.30. We observed log loss of 2.49, 2.74 and 2.46 for RF, LightGBM and the ensemble model; and a F1 score of 0.23, 0.30 and 0.30 respectively. Overall, the ensemble model seems to be justified, although the difference in accuracy is marginal.

Figure 14 shows the 30 most important variables for the RF and LGB ML models displayed together. The features are sorted in descending order according to the importance values of the LightGBM model. This indicates that both models agree with

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

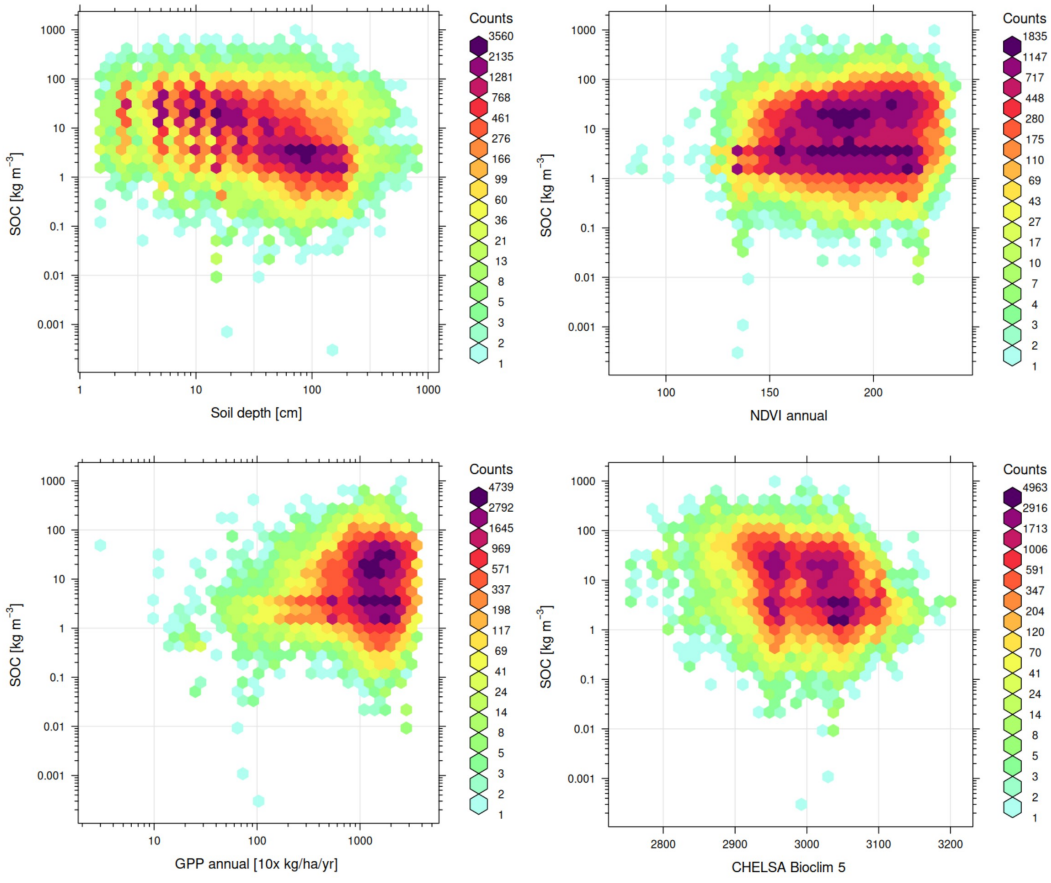

Figure 13. Scatterplots for the top 4 most important variables for modeling SOC density. GPP = annual Gross Primary Productivity based on Landsat; NDVI = Normalized Difference Vegetation Index (0–255 values); CHELSA Bioclim 5 = the highest temperature of any monthly daily mean maximum temperature.

each other in terms of important features, but there are changes in the order. Elevation (GEDTM30) seems to be the most important feature in both models. In general, climate variables from the CHELSA product at 1 km spatial resolution dominate the list of important features.

Figure 15 shows an example of soil type prediction maps for Lithic Haploxerolls. The global map reveals places where Lithic Haploxerolls are more probable, especially in North America, the Mediterranean, and central Asia. For a more detailed view, we focused on a small area to illustrate the small-scale variations compared to the land features depicted in the satellite imagery. We compared the probability maps from the OpenLandMap 2018 product, which has 250 m resolution. Since OpenLandMap 2018 does not include soil subgroups, in this case Lithic Haploxerolls, we aggregated all Haploxeroll subgroups in our map

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

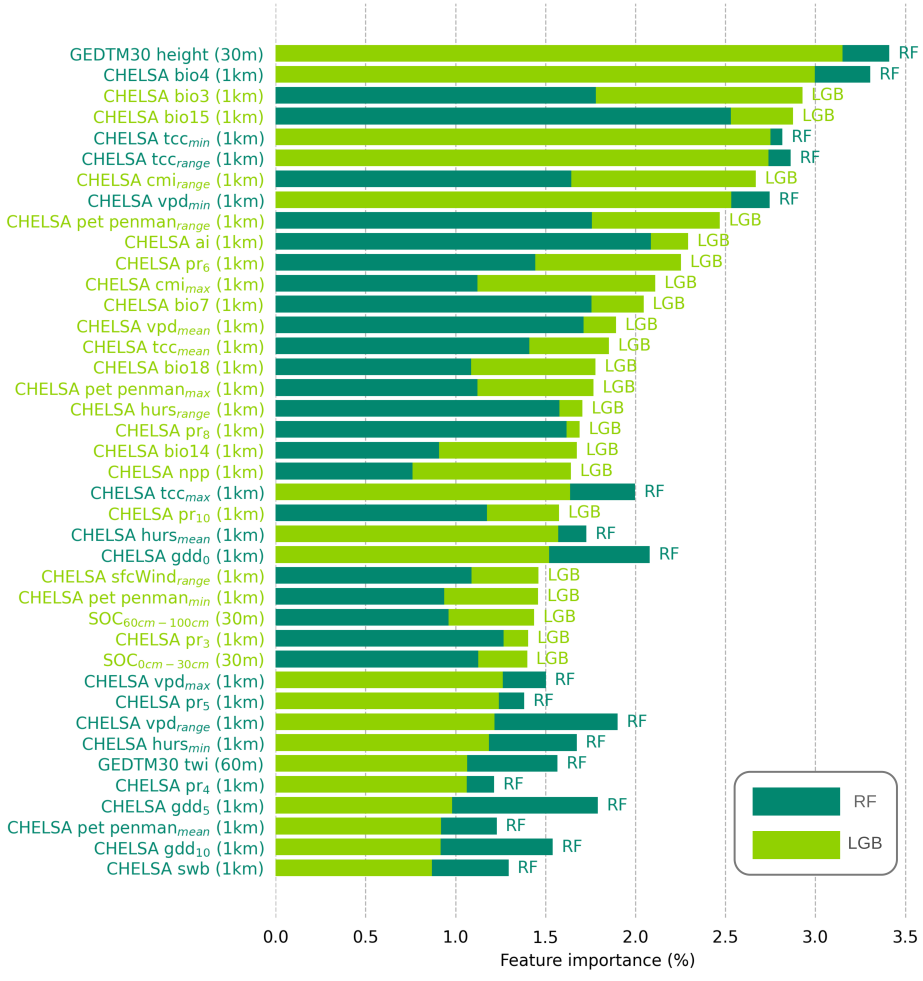

Figure 14. Variable importance plots for soil type mapping using RF and LGB models. Labels are colored based on the model (RF or LGB) that assigned the highest importance to each feature.

to generate a comparable probability layer (see Fig. 15d). Overall, the increase in spatial detail is a positive result: predictions help detect many local features, and could be potentially used for farm-scale decisions.

3.5 Comparison with other similar global data sets

Fig. 16 shows the difference in spatial detail and general patterns for an area in Germany. This illustrates the difference between 30 and 250 m spatial resolution, which in a case of managed land can be drastic with 250 m completely missing field boundaries and within-field patterns. The SOC content predictions from SoilGrids V2 seem to overpredict the SOC values by a factor of 2–3 times, which is a known problem with SOC predictions where models have limited accuracy and most of low values are

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

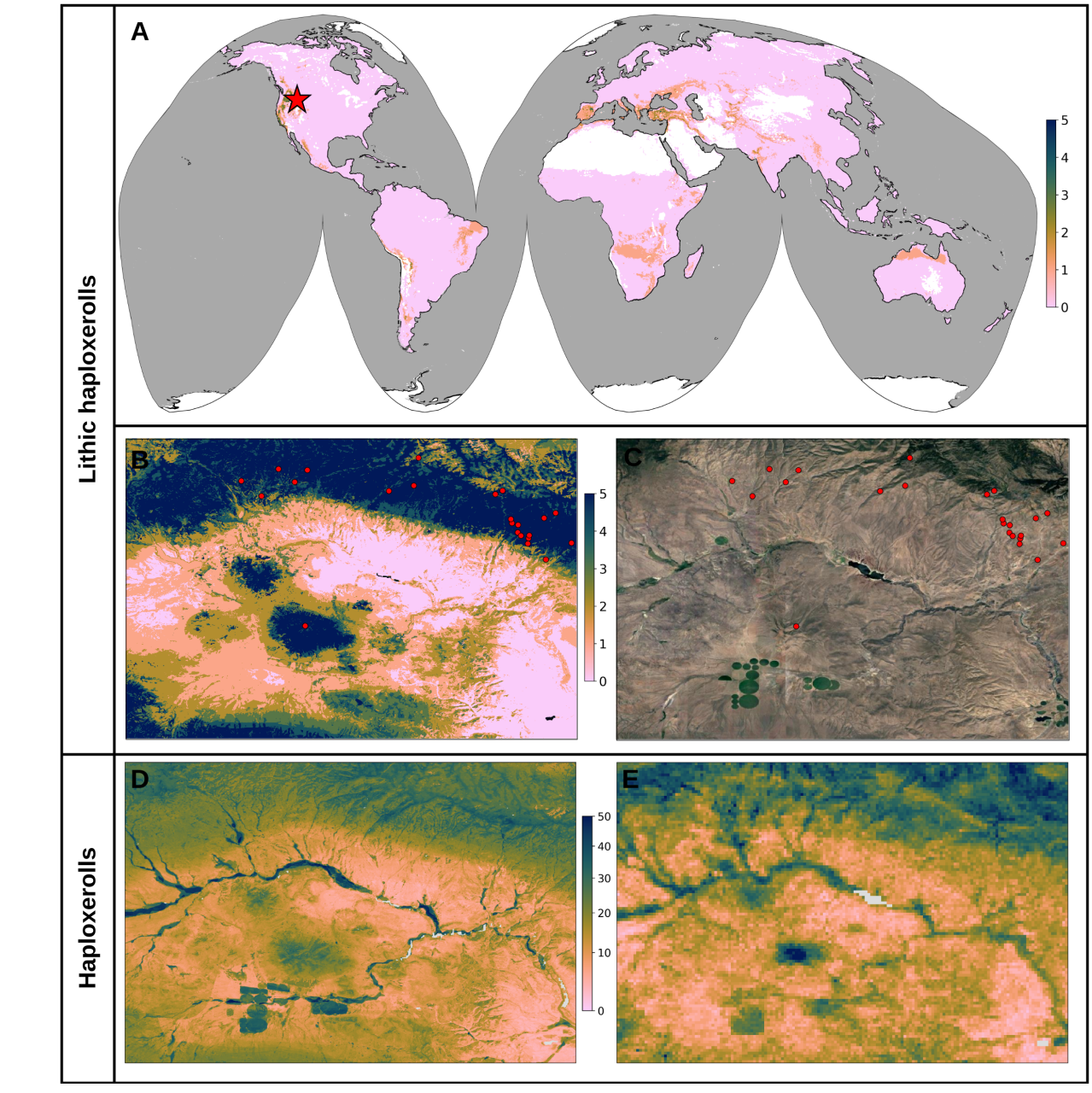

Figure 15. Example probability maps of soil types: (a) Global probability of Lithic Haploxerolls, (b) Regional probability of Lithic Haploxerolls in the area marked with a red star in the USA in (a), (c) Google Satellite view of the same region, (d) Aggregated probability of all Haploxerolls, and (e) Probability of Haploxerolls from OpenLandMap 2018 map at 250 m resolution. Lithic haploxerolls training points are shown as red circles in (a) and (b). Note: The probability maps for Lithic Haploxerolls and aggregated Haploxerolls use different legends.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

over-predicted. Note that it is not easy to compare all possible SOC and pH maps as our predictions relate to specific time intervals (e.g. 2000–2005), while many soil mapping products ignore time-dimension. In the case of soil texture fractions, bulk density, coarse fragments, clay minerology, there is probably no need to map these at shorter time intervals e.g. <20 years. In the case of chemical soil properties, our results show that differences (changes) will be visible at 5 year intervals, although in general changes in all soil properties are relatively minor (gradual) and the users need to carefully look and zoom in to notice changes.

In general, the investment in processing global 30 m resolution data seem to be paying off and the spatial detail of our predictions is comparable to that of Helfenstein et al. (2024). The differences between our predictions and national predictions open an opportunity for further local-global data fusion. Some ideas on how to implement this are mentioned further in the discussion section.

3.6 Detected trends in soil organic carbon density and soil pH

Based on the predictions, we also derived changes in soil properties corresponding to land cover transitions using the sampled point data set. The distributions of SOCD and pH changes between 2000 and 2022 across the most prominent land cover change classes are visualized in Fig. 17. In general, both SOCD and pH exhibit decreasing trends for these change classes. Transitions involving tree loss such as *'Tree cover broadleaved evergreen-Mosaic cropland or natural vegetation'* (TREBE-MCRNV), *'Tree cover broadleaved deciduous-Mosaic cropland or natural vegetation'* (TREBD-MCRNV), *'Tree cover broadleaved deciduous-Cropland rainfed'* (TREBD-CRPRF), *'Tree cover needleleaved deciduous-Mosaic tree and shrub or herbaceous cover'* (TREND-MTSHH) and *'Tree cover needleleaved evergreen-Grassland'* (TRENE-GRASS), are associated with stronger negative trends in SOCD.

These results align with the findings that SOC loss in the tropics is largely driven by deforestation (Fig. 18), although increasing droughts and forest fires may also contribute to this trend (Naval et al., 2025). For other land cover transitions, the decrease in SOCD is less pronounced. Changes in pH exhibit a relatively uniform distribution in all examined land cover change classes, with a slight trend toward acidification. We finally estimate that the world has lost 11 Pg of SOC from 2000 to 2022 based on these results. Note that, due to the limited availability of training data, especially for the Russian Federation, the actual loss of SOC could be even higher.

4 Discussion

4.1 Summary findings

We implemented a High Performance Computing system (EO-SoilMapper) to map dynamic soil properties at multiple depths (0–30, 30–60 and 60–100 cm) over time (5–year intervals from 2000–2022+) and with uncertainty quantified per pixel. This allowed us to produce complete, consistent and current predictions of some key soil properties such as SOC content, SOC

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

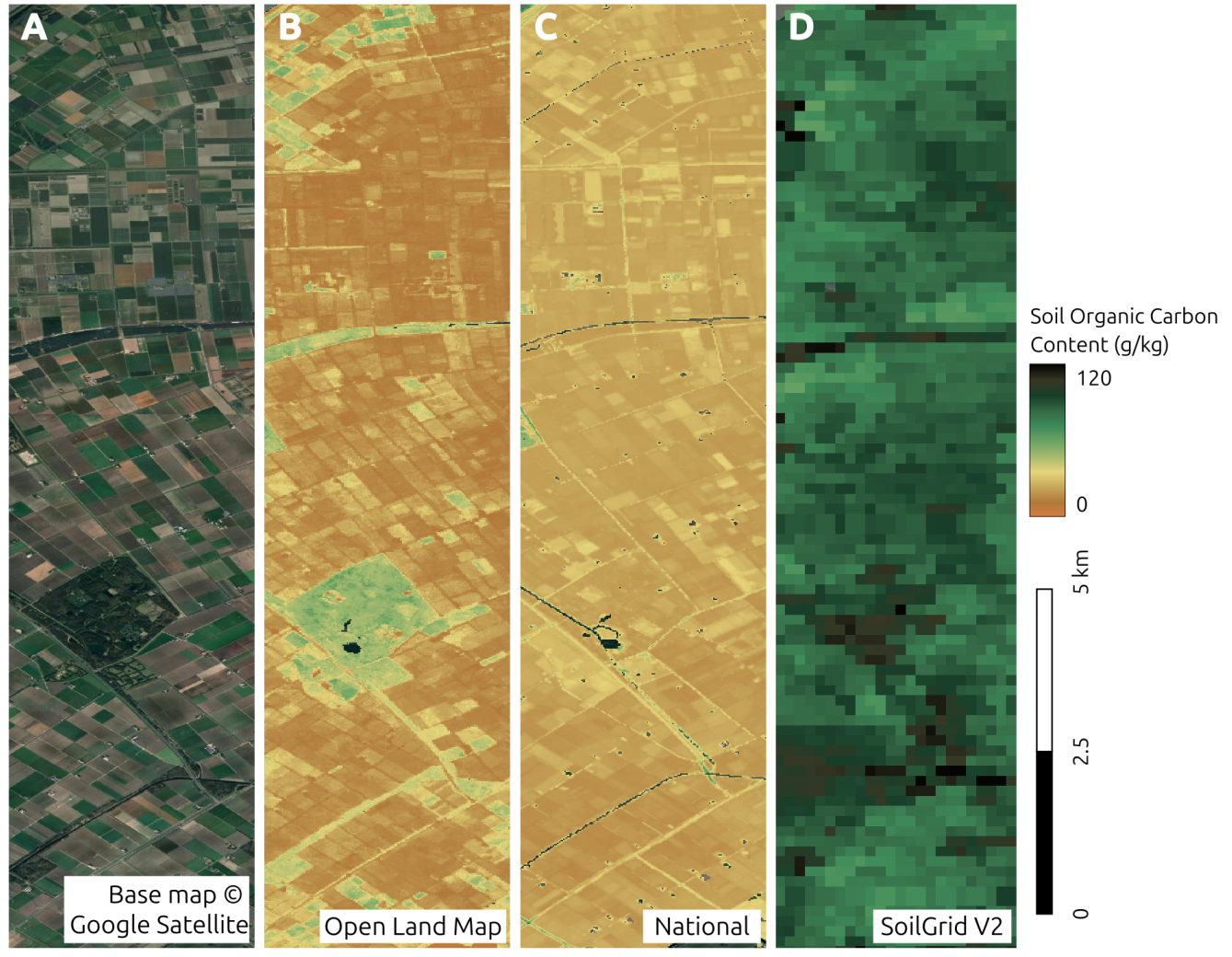

Figure 16. Comparison between SOC content maps of the region in the Netherlands, based on: (A) satellite imagery from Google Satellite (last access: 21th, May); (B) our model predictions for the 2020–2021 period (0–30 cm depth, 30 m resolution). (C) locally generated national SOC maps (0–5 cm depth, 25 m resolution, converted from SOM) for the year 2020, produced by Helfenstein et al. (2024); (D) SoilGrids V2 data (0–5 cm depth, 250 m resolution) released in 2020 (Poggio et al., 2021).

density, soil pH, and soil types at unprecedented spatial resolution. We refer to this data set as the *"OpenLandMap-soildb"*. Our ambition is to continuously update, expand and improve these data to serve the global good.

We evaluated the accuracy of prediction models using stratified testing based on climate zones, 5–fold spatial blocking cross-validation, and LOYO using best quality data. The results show improvements in terms of spatial detail (Fig. 16) and accuracy (Fig. 8) for SOC content, SOC density, and soil pH, compared to previous global soil mapping initiatives (Hengl

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

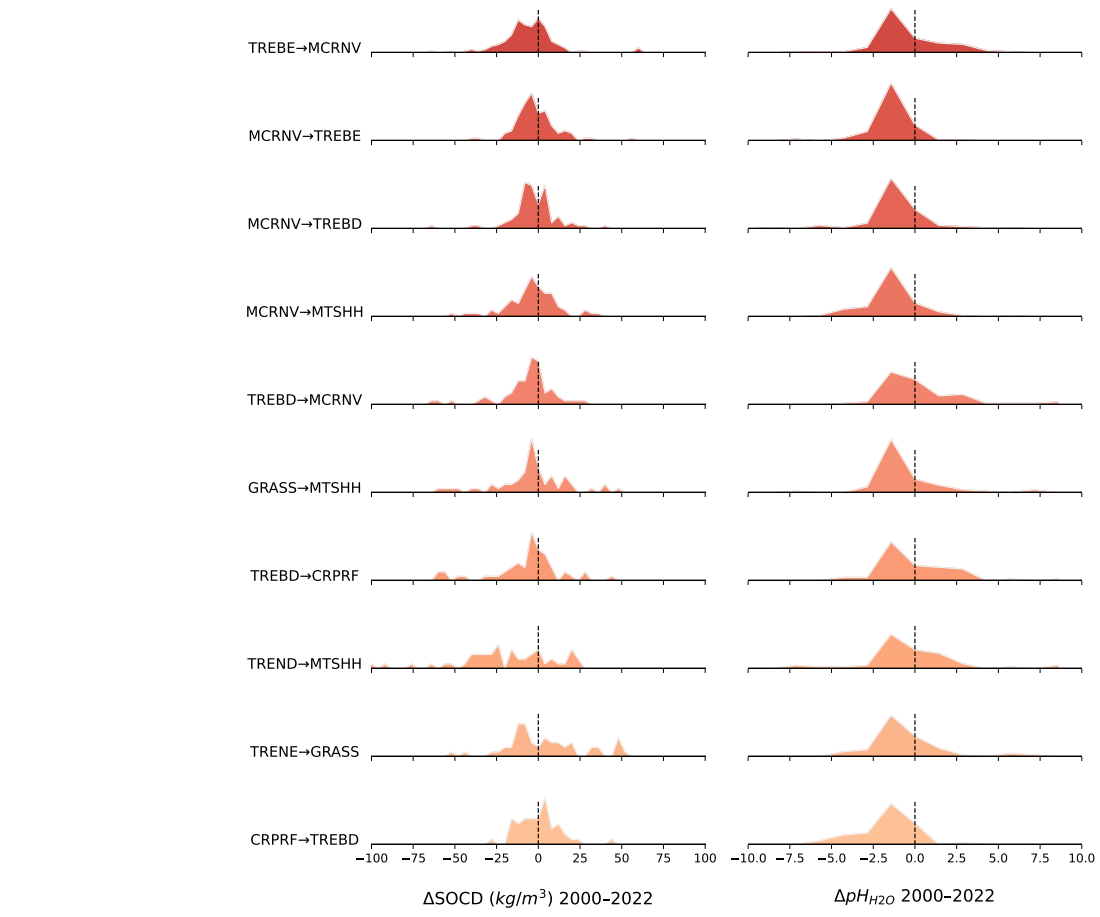

Figure 17. Distribution of soil property changes from 2000–2005 to 2020–2022+ (ridgeline plots; left: SOCD, right: pH) across the top 10 most frequent land cover change classes. **TREBE** — Tree cover broadleaved evergreen; **MCRNV** — Mosaic cropland or natural vegetation; **TREBD** — Tree cover broadleaved deciduous; **MTSHH** — Mosaic tree and shrub or herbaceous cover; **GRASS** — Grassland; **CRPRF** — Cropland rainfed; **TREND** — Tree cover needleleaved deciduous; **TRENE** — Tree cover needleleaved evergreen. The intensity of the color indicates the relative density (frequency) of occurrences for each land cover change class within the sampled dataset.

et al., 2017; Poggio et al., 2021). Having time-series of predictions based on a single model allows us to compare changes over time, which is especially interesting when it comes to SOC and soil pH (Fig. 17). The loss of carbon density is known to be related to land degradation, which often begins with deforestation, draining of wetlands, and similar. Naval et al. (2025) found that annual forest burning depletes soil C stocks (0–30 cm) by 16%, triennial burning by 19%, and long-term agriculture by 38% (compared to undisturbed forest in the tropics). Our results predict that the SOC losses for the last 25+ years are primarily driven by deforestation and the removal of peatlands.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

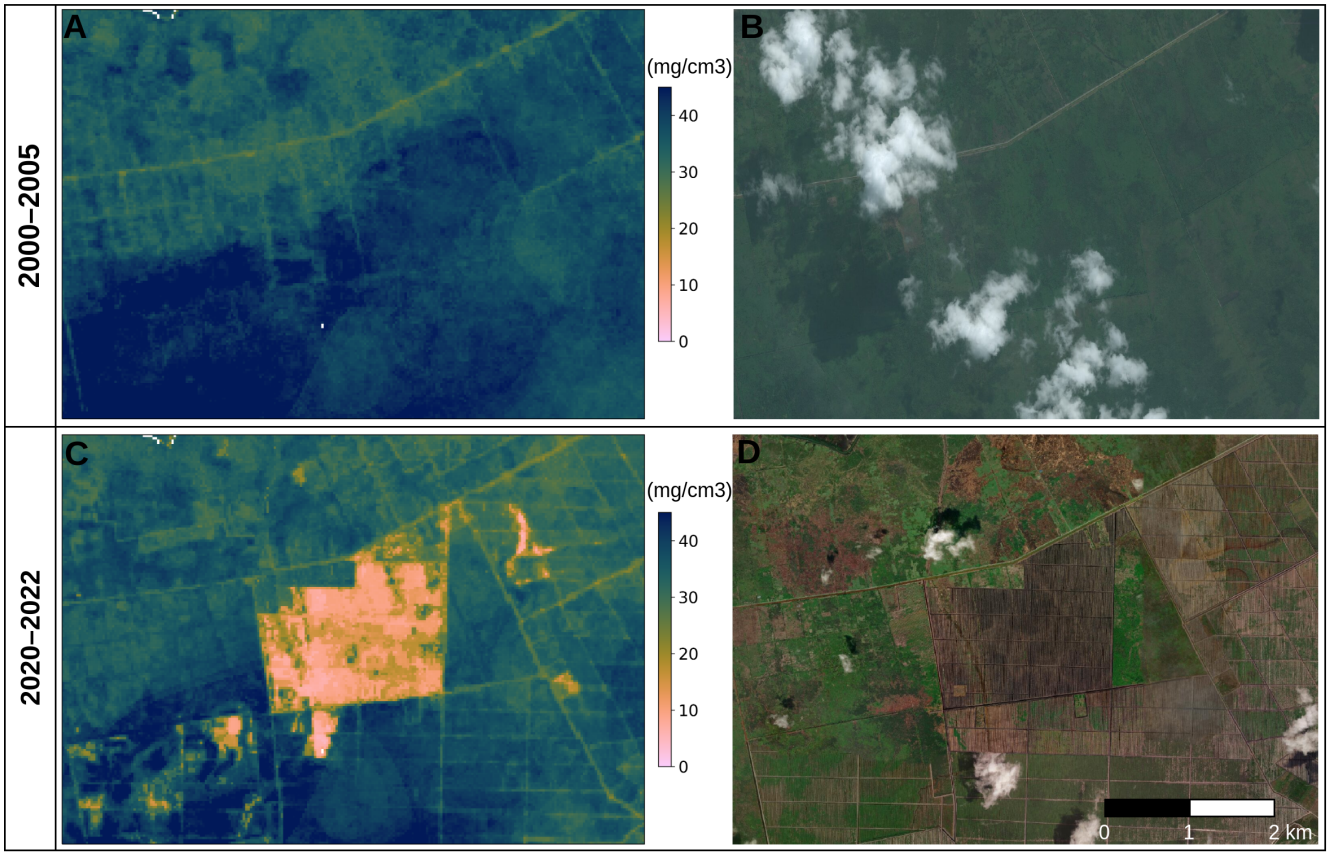

Figure 18. Comparison predictions of SOC density for 2000–2005 and 2020–2022 periods for area in Indonesia. (A) SOCD map for the period 2000–2005, (B) ESRI satellite imagery from 2014, indicating largely intact forest cover, (C) SOCD map for the period 2020–2022, and (D) ESRI satellite imagery from 2022, revealing extensive clearing and agricultural conversion.

These results of the accuracy assessment confirm that the time invested in preparing these data at high spatial resolution (30 m) was worthwhile. This required significant efforts to prepare and fine-tune the training data and input covariate layers, in addition to the technical challenges of processing these data in a cost-effective way. It was especially tedious to import and bind all national and international soil laboratory measurements and observations into a single analysis-ready training data set. Laboratory soil data are often only available in parts and without any standard schema: to fully document all harmonization steps can be extremely lengthy, and eventually we had to often make expert decisions, resulting in the extensive code-base provided at https://soildb.openlandmap.org.

We have released all the data produced as open data (CC-BY license) and have exposed our workflows currently implemented via the scikit-map package calling for the establishment of open development communities, comparable to the Open Soil

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Table 2. Total carbon stocks (Pg) by period and soil depth.

Period	Depths	Lower (p16)	Mean	Upper (p84)
2000–2005	0-30	249	472	899
	30-60	151	302	608
	60-100	133	289	635
2005–2010	0-30	245	468	898
	30-60	147	296	602
	60-100	129	284	629
2010–2015	0-30	243	465	891
	30-60	148	297	600
	60-100	132	285	626
2015–2020	0-30	242	462	885
	30-60	147	296	596
	60-100	131	285	623
2020–2022	0-30	239	461	890
	30-60	144	293	599
	60-100	128	283	627

Spectral Library (Safanelli et al., 2025), to help maintain and improve these data. In the next sections, we discuss limitations of the data, suggest some recommended uses of them, and envision future development directions.

4.2 Towards a more accurate estimate of soil carbon dynamics

One of the significant results of this work is that we have been able to estimate the SOC stocks based on detailed SOC density maps. Our results show (Table 2) the global stocks for each spatio-temporal block. For example, the carbon stock for the 0–30 cm depth interval in the most recent time-frame (2020–2022) is estimated to be 461 Pg (Peta grams) for 114 million km^2 (excluding Antartica, Greenland, deserts and permanent ice/snow) with a 68% probability range of 239–890 Pg. We further estimate that the total stocks for 0–1 m of the soil depth is 1037 Pg. This number is somewhat higher than what is reported by Padarian et al. (2022b), but also significantly less than what several other sources suggest (Jackson et al., 2017; Lin et al., 2022).

The significant amount of SOC in our predictions in the subsoil is mainly contributed to the northern hemisphere and tropical peatlands. In the rest of the world, deeper soils typically contain only a fraction of the total SOC e.g. 10–15% for 30–200 cm. Our predictions of high SOC stocks for northern latitudes (> 55 degrees) should be taken with caution, as we had limited training data for Russia. Our predictions for northern latitudes are likely based on the two main data sources: Northern Circumpolar Soil Carbon Database (NCSCD) (Hugelius et al., 2013b) and Interior Alaska Carbon and Nitrogen stocks (Manies et al., 2020). These data sets appear to represent northern peatlands and wetlands with obvious right skewness toward high SOC

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

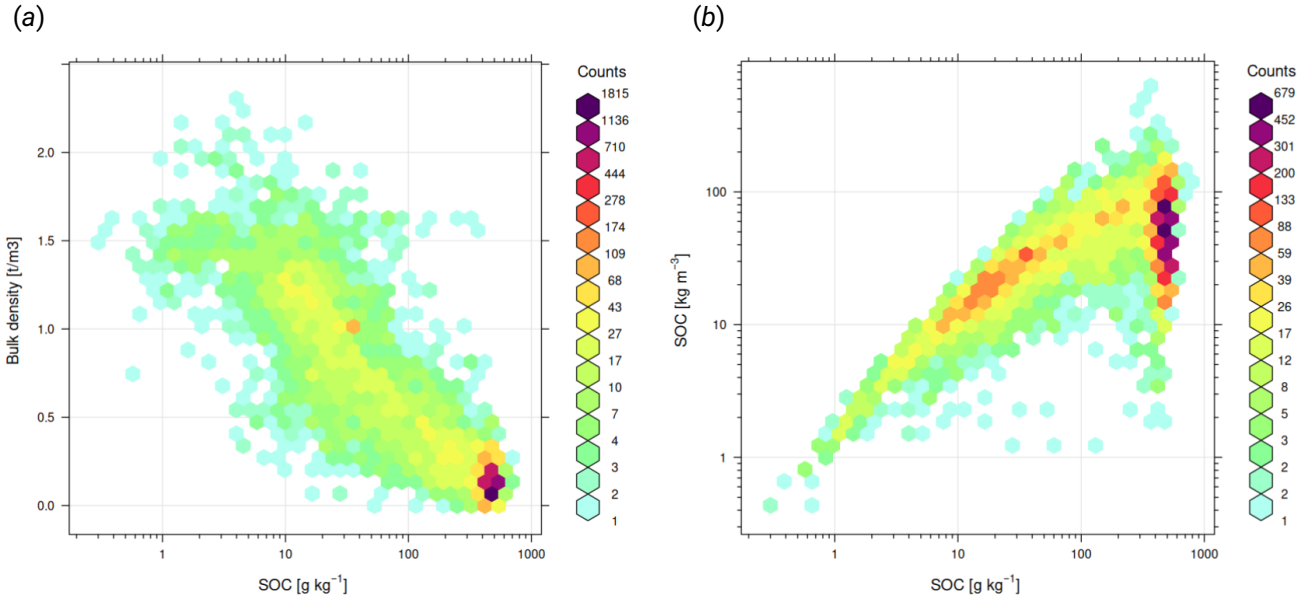

Figure 19. Density plot showing: (a) relationship between SOC density and Bulk Density (fine-earth), and (b) SOC content and SOC density based on the Northern Circumpolar Soil Carbon Database (NCSCD) (Hugelius et al., 2013b). The right figure shows how the main source of uncertainty in SOC stock estimates for the world are likely the high variation in the the stocks for soils with >10% of SOC.

density (Fig. 19). Most of the literature agrees that most SOC stocks in the world belong practically to Canada and the Russian Federation (Scharlemann et al., 2014; Crowther et al., 2016). We now provide predictions of SOC density at high spatial detail. An important note here is that many of the tundra and taiga areas of the world, although probably have a high SOC content, are also shallow soils with a significant amount of coarse fragments. Although we corrected for coarse fragments during the derivation of the SOC content, many soil profiles do not report coarse fragmentation, so the actual stocks we estimated could be somewhat lower in fact. We plan to add coarse fragments, depth to bedrock, and similar to the list of variables for global soil mapping in the next update.

Our results further show that the planet has lost at least 11 Pg of SOC for 0–30 cm in the period 2000–2022+. We think that this is probably a conservative estimate as our models possibly smooth out and also miss especially some peatlands in the tropics. It is important to note that this number is derived directly from the data, that is, @ is not based on any assumptions about the processes and / or drivers of the SOC change (as in, for example, Padarian et al. (2022b)). We hope to improve this estimate with each new update of the predictions, which will hopefully be driven mainly by the addition of more and higher quality training points.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

4.3 The OpenLandMap approach to global soil mapping

Having a 30 m resolution data product means that the state-of-the-art global soil mapping is on a path to becoming compatible (in terms of spatial detail, consistency, and completeness) with high-resolution global layers such as land cover (Potapov et al., 2020), cereal extent (Van Tricht et al., 2023), forest canopy (Turubanova et al., 2023), and similar. Our predictions span almost 25 years and therefore can be used to detect changes, i.e. this is now potentially a farm-scale, decision-ready geospatial soil database. Our trend analysis based on space-time predictions visually shows that the decrease in SOC is correlated with deforestation, especially in countries rich in organic soils such as Indonesia (Fig. 18).

We model dynamic soil properties using a data fusion approach, that is, using an extensive combination of time series of EO biophysical indices, climatic variables, terrain variables, variables representing human impact, and using pseudo-observations to help incorporate soil knowledge into ML (Fig. 1). Compare with the approach of van Wesemael et al. (2024), for example, who decided to fit two separate models — one for areas with enough bare-soil spectra, one for areas permanently covered with vegetation such as grasslands and forests (bare-soil spectra are often only available for 1/3 to 1/2 of the land mask or less). In our opinion, direct use of bare-soil spectra, for example from Landsat or Sentinel optical images, although shown to be promising for mapping SOC in agricultural areas, seems to be applicable only for a narrow niche of mapping top soil in intensely managed agricultural soils. In our framework, we use instead a much denser number of long time-series of Landsat indices (bi-monthly to annual) to represent both bare surface and vegetated spectra. This makes our OpenLandMap-soildb global soil mapping approach an order of magnitude more computational, more hyper-dimensional than the approach of van Wesemael et al. (2024), and this is probably a downside. On the other hand, the advantage of our approach is that we did not have to fit separate models, then fix boundary issues, etc. Another disadvantage of our approach is that we underused the potential of extracting bare-surface soil spectra. In summary, using bare-earth soil spectra and using dense time-series of monthly/bimonthly spectral signatures (our approach) are both valid approaches and would need to be compared versus the same test data to objectively compare differences.

From a personal perspective, the scale of the product, due to the 30 m/120 m spatial resolution, different timeframes and depths, quantiles and mean, and numerous properties, required a high computational effort. In general, the production phase required hundreds of thousands of CPU hours and resulted in over 30 TB of output storage size. Consequently, we had to make compromises in terms of selection of properties and temporal resolution — currently we only map 5–year periods (averaged over 5–year blocks: 2000–2005, 2005–2010, 2010–2015, 2015–2020, 2020–2022+). For top soil SOC density we also produced a more granular product with overlapping biannual time-frames to be used as annual product in the range 2015–2022+. In addition, we originally wanted to also map coarse fragments, macro and micronutrients, and similar (Fig. 20), but these would have pushed us beyond the project budget. To illustrate data volumes, the world's land mask at 30 m is about 220 billion pixels (Ho et al., 2025), therefore, if we include multiple depths and uncertainty, only one time period for one soil property contains more than a trillion pixels.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

	Variable - measurement unit	*ISO code**	**SoilGrids V2** 250 m	**OpenLandMap-soildb** 30, 120, 240 m	*Revisit planned**
primary soil properties	Soil organic carbon density [kg/m^3]	10694.1995	☒	☑	Annual
	Soil organic carbon [g/kg]	10694.1995	☑	☑	Annual
	Soil pH in H_2O [-]	10390.1994	☑	☑	Annual
	Soil pH in $CaCl_2$ [-]	10390.1994	☑	☒	Annual
	Extractable P, K [ppm]	1952:2008	☑	☒	Seasonal
	Cation exchange capacity [cmol(c)/kg]	11260.1994	☑	☒	Annual
	Sum of total nitrogen [g/kg]	13878.1998	☑	☒	Annual
	Electrical conductivity [mS/m]	11265.1994	☒	☒	Seasonal
	Bulk density fine earth [kg/m^3]	11272.2017	☑	☑	Long-term
	Soil texture fraction clay-silt-sand [%]	11277.2020	☑	☑	Long-term
	Coarse fragments volumetric [%]	11277.2020	☑	☒	Long-term
	Depth to bedrock [cm]		☒	☒	Long-term
	Toxic metals in soil [ppm]	21365:2019	☒	☒	Annual
	USDA subgroup taxa [-]		☒	☑	Long-term
landscape variables	Summary annual bare surface [%]		☒	☑	Annual
	Mean annual GPP [kg/ha/yr]		☒	☑	Annual
	Median annual canopy height [cm]		☒	☑	Annual
	Land cover annual [-]		☒	☑	Annual

Figure 20. Primary soil properties of interest for global soil mapping (Chen et al., 2022), ISO code, temporal granularity and current availability in OpenLandMap-soildb. The ISO standard code is approximated. Toxic metals include: arsenic, cadmium, cobalt, chromium, copper, nickel, and lead.

The preparation, harmonization, and binding of points (training data) and the preparation of the covariate layers took almost 60–80% of the OpenLandMap-soildb project time and was difficult to predict. Consider, for example, the peatland extent map of the world: currently, there are at least four overlapping data sets that claim to represent the extent of the world's peatlands:

1. WRI's Global Peatlands extent map at 30-m (250-m effective) (Gumbricht et al., 2017; Xu et al., 2018);
2. Peat-ML at cca 8 km (Melton et al., 2022);
3. PEATGRIDS at 1-km (Widyastuti et al., 2024);
4. Global Peatlands Map 2.0 produced by the Global Peatlands Initiative (https://globalpeatlands.org/);

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

A practical logical solution for us was to average between the multiple sources to produce an ensemble extent map with values 0–100% (a probability map of the world's peatlands we produced is available at https://doi.org/10.5281/zenodo.13951438). However, this takes time and requires maintenance; therefore, many global soil mapping projects in essence need to budget for excessive preparation, gap filling, and harmonization of both target training points and covariates. We currently do not see how this type of work could be replaced with AI as a team of experts is needed to open, analyze, compare maps and design an original procedure to combine data.

4.4 OpenLandMap-soildb data limitations

Even though our cross-validation results show that our predictions are significant with CCC often exceeding 0.8, it is also important to list some observed limitations of these data. The following three limitations should be especially emphasized:

- *Soil laboratory data harmonization issues*: Although we have fully documented import, harmonization and binding of soil laboratory data, we admit that harmonization based on a simple translation formula (Eq.4) and the potentially difference in SOC and soil pH values between different data sets is unknown. We consider this a noise component and assume that it is random, but this has not been tested. Many soil observations and measurement data sets are discontinued / no longer maintained; hence it would be difficult to find all original data producers and check all reference laboratory methods used. Liu et al. (2024) shows how even trivial things, such as differences in soil sample grinding and drying processes, can lead to significant differences in the SOC estimate at laboratory level. It is very well possible that we have missed some important metadata and that our code could be further optimized; we call soil scientists and soil laboratory data curators to look at our code (https://soildb.openlandmap.org) and help improve the consistency of data import.

- *Large geographical gaps and spatial clustering of training points*: Unfortunately, availability of training data follows the well-known paradox of all physical geography, where places with highest biodiversity / geodiversity usually have proportionally fewer ground observations and measurements. There is no simple solution for this. Nevertheless, we have at least made sure that the hold-out samples (5–10% best quality samples) are equally distributed to avoid over-representing USA and Europe.

- *Depth is used as covariate resulting in redundancy of other covariates*: soils are 3D, but our covariates usually only represent the soil surface (with few exception e.g. the soil bioclimatic variables (Lembrechts et al., 2022)). This means that values of all covariates are basically copied across all soil depths, leading to redundancy in training data. Although technically this is not a problem for decision-tree based algorithms, redundancy is obvious and this does not appear optimal. One solution to this problem is to fit separate models for different depths as in Nauman et al. (2024); another option is to fit multi-response models where models for multiple depths are fitted at once, but this would require that we gap-fill all missing values as multi-response models require that all values are available across regression matrix.

- *Smoothing of some lower/higher values and omission of potential hotspots*: We have compared our predictions of clay content with the SOLUS predictions for USA (Nauman et al., 2024) to discover that our predictions miss several hotspots

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

of higher clay content (>50% clay fraction) for example in the Missisipi river delta and similar. This is a known issue of regression smoothing out values and soil laboratory values over-representing agricultural soils. Although there is no simple solution to this problem (without collecting more training points), we at least recommend that users incorporate our prediction intervals into their decision frameworks.

- *Limited data with repetitions over time / lack of soil monitoring stations*: Unfortunately, unlike meteorological data, very few soil monitoring projects produce repeated measurements over time (in meteorology, these are referred to as *"stations"*, or in statistics, as *"longitudinal data"*). Exceptions are LUCAS soil points and a few other national/sub-national permanent soil monitoring networks (e.g. Broeg et al. (2024) and Keel et al. (2019)), where soil surveyors return to the same locations every few years. The ideal data set for dynamic modeling is where the majority of points contain repetitions over time; in this case we would need at least 4–5 repetitions so that we can also observe changes in soil properties per site. In our case, only a few areas (e.g. Europe with LUCAS soil) have repetitions of measurements through time, which majority of data (>80%) basically does not overlap spatially. This is a serious limitation and can only be improved by more countries setting up permanent monitoring stations where exactly the same soil properties are measured at least every 2–3 years.

Our modeling also suffers from limited harmonization of the training data. The values of soil properties can differ only because different laboratory standards and different sampling designs are used. For example, a country that only samples agricultural soils and makes all their soil laboratory data on sampling and monitoring agricultural land available could significantly underestimate national SOC stocks, as it would completely miss various SOC pools in forests, wetlands, and peatlands. In order to produce unbiased global estimates of SOC changes at the highest possible spatial resolution, the world needs global unbiased predictive mapping models that can account for large spatial clustering of training points and where all values are at least standardized, at best fully harmonized using interlaboratory collaborations (Safanelli et al., 2023).

Unfortunately, most international organizations cannot still fully agree on the standard for the sampling, analysis and registration of SOC stocks (Even et al., 2024). For example, the UN's Convention to Combat Desertification (UNCCD) currently measures Land Degradation Neutrality at 300 m spatial resolution focusing on 0–30 cm top-soil only (Cowie et al., 2018); other organizations require 0–100 cm estimates, while in Europe soils have been sampled for 0–20 cm depth intervals (Orgiazzi et al., 2018). In that sense, we also have high hopes for all the harmonization and networking initiatives of the FAO's Global Soil Partnership (GSP). In particular, the Global Soil Laboratory Network (GLOSOLAN) is a promising platform to find international standards that work for everyone. If these are application-centered and are released as open data / with code on Github or similar, this could solve many problems of data harmonization.

In this work, we also promote adding pseudo-observations to help incorporate soil knowledge into machine learning. Note, however, that from the total set of measurements used for model building, only about 5–10% of the total training points were pseudoobservations (pseudo-points are available at https://soildb.openlandmap.org); however, their spread around the global land surface is consistent, and thus they may appear as being overrepresented. It is important to emphasize again that we used pseudo-observations only for the final model fitting and not for validation or hold-out testing of the predictions. Also note that,

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

even though we use pseudo-observations representing deserts, permanent ice, and rock outcrops, we do not predict values for them. We believe that the addition of high-quality pseudo-observations helps produce more realistic predictions, especially at the edges of feature space where arid landscapes and climate tend to dominate (as also illustrated in Tian et al. (2024)).

The soil type maps we produced in this study are based on a relatively simple ML approach of basically putting all points and covariates together and then fitting the best model possible. This approach ignores multiple aspects of the data:

- Soil is three-dimensional, that is, for soil classification, it would be important to have more information about vertical stratification in the sense of diagnostic horizons and parent material, which we could not present with the covariates we use except soil depth. In some parts of the world, ground-penetrating sensors are used, for example, to produce Gamma radiometric images (Ng et al., 2023) or similar. To our knowledge, no such data are available globally (and might not be available in the decades to come).
- We ignore hierarchical relationships (proximity; parent-child relationships) between soil classes. To our knowledge, there is currently no ML method in which a hierarchical relationship can be integrated into the model fitting, but it makes sense to continue exploring this option further.
- For training models, unfortunately, we did not have global maps of diagnostic properties. For a global list of soil types, the number of diagnostic properties would have been excessive, e.g., a few hundreds of properties. This was currently beyond the scope of this project.

We could not gain access to the national soil profile data set of China (Liu et al., 2022), LUCAS for the year 2022, and the Soils4Africa project pan-continental collection of soil samples for the purpose of global soil mapping. We still have huge geographic gaps in training points for the 4–5 largest countries: Russian Federation, China, India, Kazakhstan, and similar (Fig. 7). Hopefully, if the data curators of the previously listed point data sets recognize the benefits of using and contributing to OpenLandMap-soilDB, we would be happy to integrate these data and update predictions. We are open to signing data sharing agreements that protect these laboratory data from misuse.

4.5 Detection limits and standard change rates

As shown in Fig. 6, we decided to aggregate the predictions into space-time blocks (4–points) and then serve only block predictions further e.g. predictions for 0–30 cm depth interval for the period 2000–2005. This increases usability of these data as most users are interested in depth intervals (e.g. 0–30 cm) and block-predictions in time (e.g. 2000–2005) can be matched with land cover change maps as in e.g. Potapov et al. (2022), which are also centered on 5–year periods 2000–2005, 2005–2010, …, 2020–2025.

Averaging predictions and prediction intervals has the following effects on the output data:

- The prediction intervals of the blocks are about 30–40% narrower than for the original point predictions (Fig. 21).
- Blocking predictions between two years reduces inter-annual variability, which is usually not of interest for SOC mapping. For example, climatic oscillations between years can result in significant differences in Landsat-based indices from

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

year to year (as in climatic modeling, it is important to smooth out random variation), which could then reflect on soil predictions. This makes trend analysis cumbersome, as the values oscillate from year to year.

– As the prediction errors of the means are narrower than the individual prediction errors, this allows users to detect changes in SOC at shorter periods even using sampling methods of limited precision (Fig. 21).

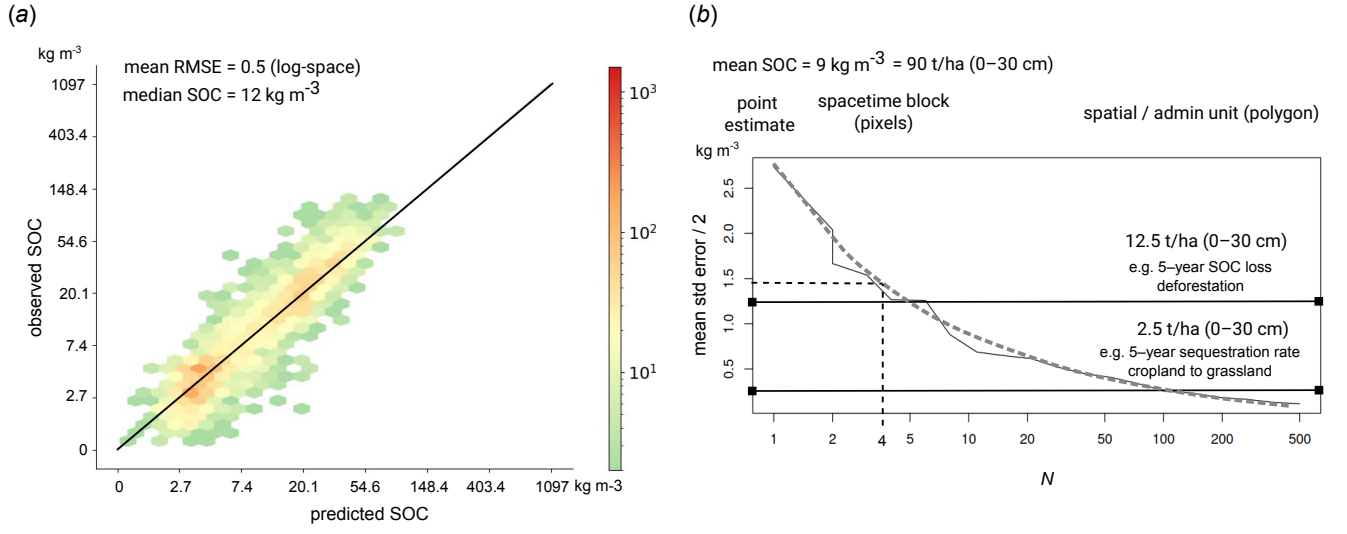

Figure 21. Simulated example of how SOC density prediction errors estimated through validation (a) relate to detection limits for different SOC sequestration / SOC loss rates (b).

From a practical point of view, most users of soil maps expect that soil predictions refer to some standard depth interval e.g. 0–30 cm. Likewise, soil properties change gradually and often slowly; hence it is sensible to expect that, if one were to produce predictions of SOC content for every year, most of pixels in the map would not change much and this change could be significantly lower than the average prediction error. Broeg et al. (2024) showed, using revisited sites for Bavaria, that although the prediction accuracy of the SOC was high, direct validation of the derived SOC trends revealed a significantly higher uncertainty. In this work, because we also opted to generate predictions at 30 m spatial resolution, which makes this a relatively large data set, we also decided to average the values to somewhat reduce the data volumes from hundreds of terabytes to a few tens of terabytes. One could argue that we could have produced only point predictions, then let users aggregate values how ever they prefer, again we have estimated that in that case data volumes would have expanded beyond what we can handle (in terms of computing time / costs), but in the future saving the whole distributions of predictions would be an option (provided that there is enough storage for these data).

Fig. 21 shows an example of the simulated effect on the standard prediction error (error of the mean) assuming averaging values of 1 to 500 points under the assumption that the SOC density follows a log-normal distribution. Assuming that the standard prediction error (RMSE) of our model is 0.5 on the logarithmic scale, and assuming that the mean value is 9 kg m^{-3}

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

for 0–30 cm (agricultural soil), it is easy to show that the standard error of the mean for an average of 4 points (e.g. 0 and 30 cm depth and years 2000 and 2005) will be about 2.8 kg m^{-3}. Following the Nyquist theorem (Hengl et al., 2013), the detection limit is RMSE/2, hence we would be able to detect per-pixel changes which are >1.4 kg m^{-3} per 5–year period. Assume the standard SOC carbon sequestration rates for the conversion of cropland to grassland of about 0.5 t ha^{-1} yr^{-1} for 0–30 cm (which corresponds to ΔSOC of about 1.25 kg m^{-3} for 5–year period). This means that with the model error of 0.5 on the log-scale, one would probably not be able to detect changes in SOC on a scale of 5–years, but a 20–year scale would be required (Fig. 21b). On the other hand, with 1.4 kg m^{-3} detection limit, one should be able to estimate SOC loss at a time period of 5 years, assuming 2.5 t ha^{-1} yr^{-1} SOC loss rates for 0–30 cm depth interval, i.e. loss of about 1.25 kg m^{-3} for a 5–year period. If you compare with the accuracy plot for SOC density (Fig. 8) this shows that our RMSE for the SOC density on the log scale is approximately 0.5 (based on stratified sampling), which corresponds to the numbers we used above. This gives us some confidence that our predictions can at least be used to detect the serious effects of land degradation on the SOC changes, as also illustrated in Fig. 18. Although averaging the error seems to help increase the detection limit (following the $1/\sqrt{N}$ rule), we should emphasize that this does not change our average prediction error, so there is still some work to do to try to improve the prediction errors for local farms. Having shown this calculus, RMSE of our predictions of SOCd is 0.5 on the logarithmic scale refers to point predictions. For block predictions, the error of the mean value is possibly about 30–40% less than 0.5, so the detectable difference between two periods is possibly even more optimistic. Also note that, because most of the training data come from the USA and EU (Fig. 7), it is very well possible that our prediction errors are narrower for the two continents than for the whole land mask. On the other hand, in countries where we have major gaps (Russia, Kazahstan, China, India, tropical forests parts of Africa etc) we possibly perform worse than global average.

4.6 Combining global and local efforts across scales

Global predictive soil mapping efforts, such as the one described in this manuscript, overlap with local (national or regional) efforts. Is this redundancy inhibiting soil data production and confusing users of soil data? Feeney et al. (2022) compared global *vs* local SOC predictions for Great Britain (GB) and discovered surprising inconsistencies, leading to the conclusion that we probably often underestimate the uncertainty of SOC predictions from predictive soil mapping. Users, land owners, and land managers can become overwhelmed by the amount of data offered and question which to use. We instead believe that as long as some shared standards / shared open data licenses are used, local and global mapping efforts can be combined to the benefit of end users. We specifically support groups in using our global predictions as covariates in local modeling or as input for data fusion (see Fig. 22) and *"federated learning"* frameworks (Gallios et al., 2025). To demonstrate the synergy of local and global modeling, we are currently discussing global-local data fusion of soil carbon density predictions based on national, pan-EU (Tian et al., 2024) and our global predictions.

Although many soil mapping projects seem to overlap (as in the case of land cover mapping, for example), we believe that there is still a lot of room for multiple initiatives, as many projects are, in fact, delivering different standards. Consider the following technical specification of soil carbon in OpenLandMap-soildb (one of the Essential Climate Variables of the Global Climate Observing System):

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

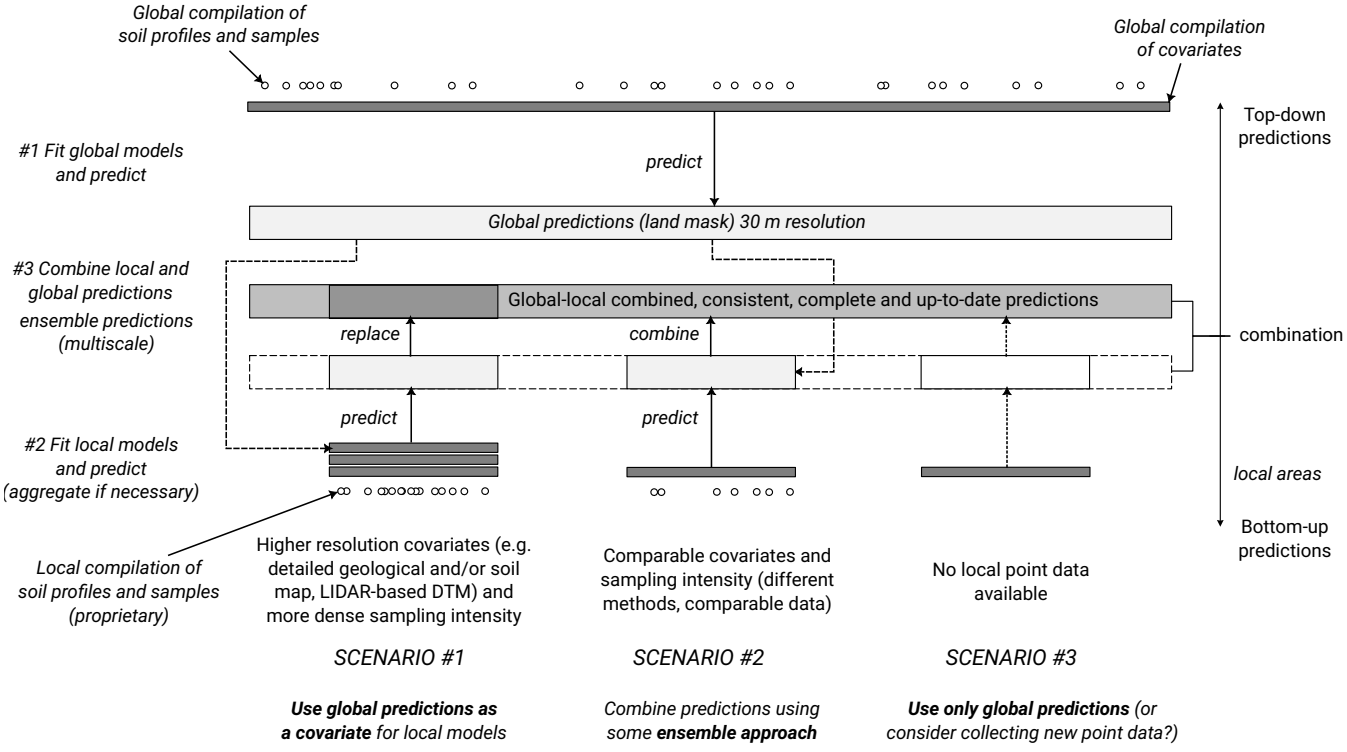

Figure 22. Proposed scheme for merging global and local soil predictive mapping outputs based on three scenarios: (1) local predictions are significantly more accurate, hence can be used to replace global predictions; (2) local and global predictions are comparable accuracy and can be best statistically combined; (3) only global predictions are available.

- Referent variable: soil organic carbon density in kg m^{-3};
- Referent laboratory method: DC ISO 10694:1995(E);
- Measurement support size: 1×1 m horizontal, 5 cm vertical;
- Prediction depth interval: 0–30 cm;
- Prediction time-interval: 2000–2005;
- Prediction error distribution: 68% probability (1 standard deviation);

In order to combine the predictions of two projects as shown in Fig. 22, both local and global should match in all of the specifications; otherwise the differences in values could be unrelated to the accuracy of each individual map. A *"soil carbon map"* tag is no longer specific enough. We probably need to start using more specific standards and specifications where soil variables match at least in reference laboratory methods, measurement units, and temporal coverage — for example, *"5–year*

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

soil carbon stocks for 0–30 cm depth interval for 2000–2005". In some cases, products from different projects could possibly be harmonized, but in general, without complying with the same standard, it is probably not necessary to compare or criticize what are essentially different soil data products; for example, the predictions of the organic carbon content of the soil (%) are not 1:1 with predictions of soil organic carbon density / stocks (kg m^{-3}), which is also illustrated in Fig. 3a and Fig. 19b.

4.7 Future development directions

In 2008, a group of world leaders in digital soil mapping launched the idea of mapping soils for the entire world at a high spatial resolution of 100 m: the GlobalSoilMap.net project (Hempel et al., 2014). The main idea of GlobalSoilMap was to produce global maps in a resolution compatible with the publicly available SRTM DEM (90 m), at that time one of the most popular global environmental products. In the original plan, the proposal was to achieve this per country, then to combine all data together. Although stitching high-resolution national soil maps to produce global data sets is technically possible and politically correct (see e.g. FAO (2022)), it has been shown to lead to significant differences at political borders. In addition, often a large number of countries are left blank, leading to limited usability of such data.

In 2009 it seemed that GlobalSoilMap could be achieved in a few years, but this was a gross underestimate. It took almost 18 years to produce complete and consistent soil property maps with comparable spatial resolution. Thanks to the exponential development of computing and Machine Learning, we are now able to predict not only soil properties at 100 m, but at a 10× finer volumes and in space-time. However, this long delay in producing soil data that matches most land cover and vegetation products indicates that global soil mapping is complex, especially with soils being hidden, buried, and impacted by multiple soil-forming processes working at the same time in a non-linear way (compared to vegetation and land cover mapping where EO images often suffice), having high short-range variability, and often based on unrecorded historic processes including extreme events such as flooding, landslides, vertical movement of materials, and similar. The mapping of soils remains one of the most challenging tasks in physical geography.

We foresee the following future development directions in dynamic soil mapping (unsorted):

- **Hybrid Machine-Learning / Process-Based modeling**: there is increasing interest in the so-called *"Knowledge-based ML"*. Liu et al. (2024) shows how a relatively detailed model ecosys can be combined with ground measurements to recalibrate modeling and optimize accuracy. The ecosys framework requires a large number of inputs, produces daily values, and is currently optimized for agricultural systems. The computational load required to run ecosys at 30 m for all land mask would be enormous. In the meantime, we recommend to all soil mappers to at least put effort to develop ML models following *"Soil Science-Informed ML"* (e.g. by specifying observational priors, pseudo-observations, adding model structure design and loss functions) (Minasny et al., 2024).

- **Global digital soil twin**: assuming that we manage to integrate state-of-the-art process-based models with high resolution EO data and in-situ laboratory measurements, one could expect that one day we will be able to model soil-vegetation-land-use-climate interactions within a paradigm of digital soil twin (digital copy of the world soils is connected with the physical twin and data automatically flows in two directions).

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

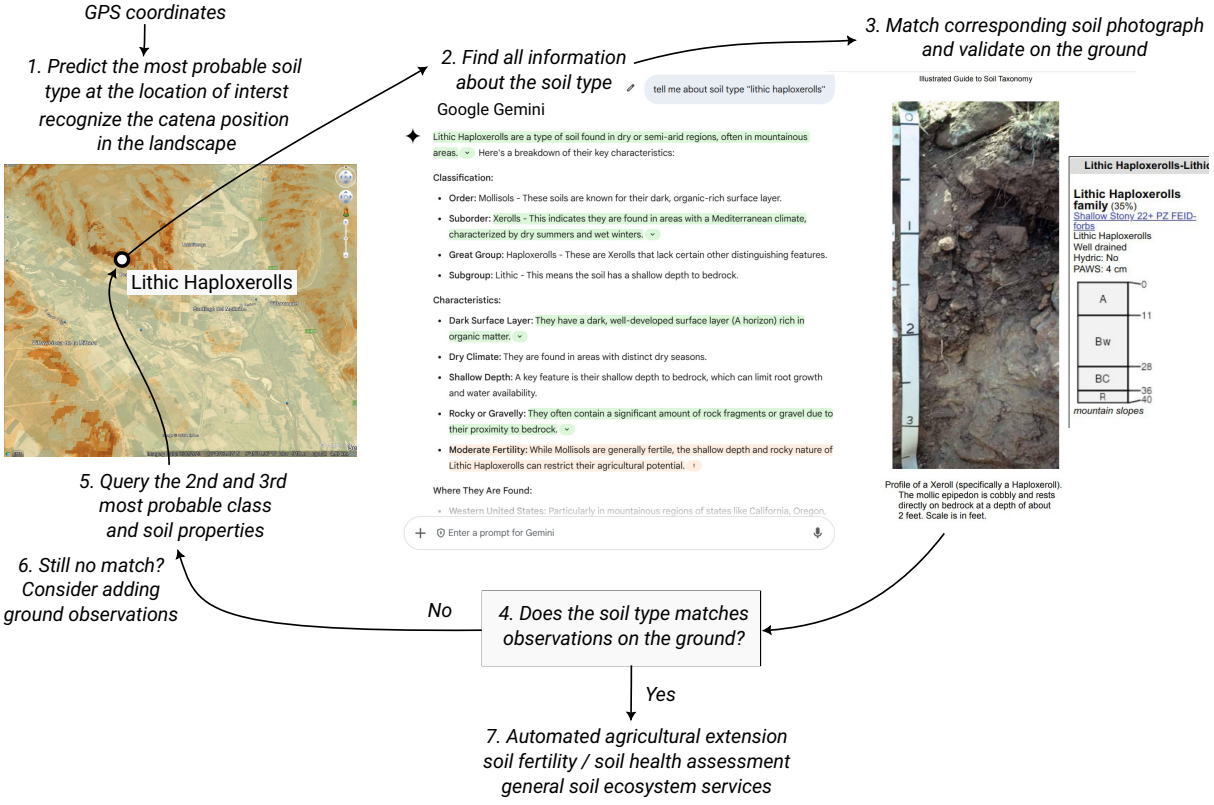

Figure 23. Example of how LLM's (in this case Google's Gemini™) can be directly combined with the global data sets we have produced to help with understanding the soils better and with taking actions. Image and text source: USDA's Illustrated Guide to Soil Taxonomy.

– **Development of rapid and cost-effective *in-situ* soil sampling instruments**: soil laboratory analysis remains costly and global training data are often limited to most developed nations. Modern in situ technologies such as soil spectroscopy and similar show that the costs of measuring soil chemical and physical soil properties can be reduced to fractions of traditional soil laboratory costs (e.g. about $150 per sample). In addition, land owners often do not have patience to wait weeks until the results of laboratory analysis are released. Here, one of the most striking examples is the LUCAS soil survey, where soil laboratory data takes almost 3–5 years until they are released. Some recent results with relatively cost-effective NIR instruments show that there is an opportunity to use a handheld near-infrared device (NIR, 1350–2550 nm) for near-real-time SOC measurements (Kalopesa et al., 2025). Results of using the YardStick™ instrument (Gyawali et al., 2025) also shows satisfactory results with an opportunity to scan changes in soil properties every 1 cm of depth (continuous functions).

– **Finer-temporal resolution modeling**: in this paper we mapped soil properties using annual values of bio-physical indices and climatic variables. Many soil variables could also be mapped at bimonthly or monthly temporal resolutions.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

For example, many geochemical soil properties, such as available N, P, K, have been shown to vary within a few weeks between sampling; soil moisture is even more variable with values changing within hours. Again, to map the world at monthly intervals at high spatial resolution would be extra computationally intensive and is certainly beyond our state of technology. However, we have recently produced bimonthly 30 m resolution GPP (Isik et al., 2025), so this is definitely possible, but we would likely have to limit any such models to top-soil only and 1–2 soil variables.

- **Using multi-response models to model and predict all variables using a single model**: In this work we fit multiple separate models for each property. This means that we assume that all target variables are independent, but they are not. To deal with multi-colinearity and overlap in target variables, one could fit a very holistic single model able to predict multiple depths and soil proprieties all at once. This multi-output approach works better for values that are somehow correlated and also is more elegant as the prediction would be for all properties at once. Random Forest is again applicable statistical method here as it supports multi-response models and predictions.

- **Using global Generative Foundation Models**: currently there is increasing interest in building global foundation models or World foundation models (WFMs) that would basically include all literature on soils (tabular, textual data, schemes and scientific visualizations) and would be able to represent world soil distribution and properties. Bodnar et al. (2025) recently released *"Aurora"*, a large-scale foundation model trained on more than one million hours of diverse geophysical data, and which can outperform several existing weather forecasting operational systems while also being orders of magnitude faster. Once such models become robust and convincing, they could be used to connect users directly with the data, which could make many traditional soil scientists and agronomists redundant. An example of how LLMs can already be used to boost knowledge about soils is shown in Fig. 23.

On the one hand, we can anticipate many exciting developments in soil science in the near future; on the other hand, we recognize that soil science seems to be still lagging behind many other environmental fields. For example, when it comes to soil laboratory data, there are now 5+ independent initiatives where global soil point data have been prepared and made ready for modeling:

1. WoSIS soil profiles and samples (Batjes et al., 2024);
2. International Soil Carbon Network (ISCN) (Harden et al., 2018);
3. Open Soil Spectral Library (OSSL) (Safanelli et al., 2025);
4. SoDaH: the SOils DAta Harmonization database (Wieder et al., 2021);
5. SoilHive (https://www.soilhive.ag/) hosted by the Varda corporation, shows all soil data available for any location in the world (both point, polygon and raster layers);

Compare with databases produced in the fields of biodiversity, biology, meteorology, or similar:

1. GBIF (the Global Biodiversity Information Facility; https://www.gbif.org) with 3 billion occurrence records;

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

2. NOAA's hosted GHCN (Global Historical Climatology Network) with over 100,000 meterological stations and 1B of daily measurements;

3. FLUXNet (https://fluxnet.org/data/) with hourly measurements of some 100+ biophysical variables at over 250 automated measurement stations worldwide;

4. GLANCE (https://doi.org/10.34911/rdnt.x4xfh3) with over 2M observations of land cover;

5. GEOROC Database (Geochemistry of Rocks of the Oceans and Continents; https://georoc.eu/georoc/) containing almost 30M of values of major and trace element concentrations, radiogenic and nonradiogenic isotope ratios etc;

It is easy to notice that soil science is seriously lagging behind GBIF and similar initiatives. In our opinion, global soil science critically misses: (a) a global (permanent) soil monitoring network of at least 300–500 permanent stations where soil properties can be tracked on a monthly / annual basis, (b) professional infrastructure where various groups can enter and access point data (as in https://www.gbif.org/dataset/, and (c) agreements on data sharing, soil sampling (Even et al., 2024), and open soil laboratory standards including a universal soil classification system blessed by the IUSS (International Union of Soil Science).

An inspiring model for monitoring soil properties over time is the International Soil Moisture Network (https://ismn.earth/) that provides open access to approximately 3200 automated measurement stations with hourly measurements of soil moisture, temperature, precipitation, and similar (Dorigo et al., 2021). Another infrastructure critically missing in soil science, in our opinion, are globally applicable mobile phone apps that allow anyone to take photographs of soil, soil spectral scans, and similar and share (as in e.g. iNaturalist app). Many soil enthusiasts and agronomists have asked us in the past *"How do I share my data?"*. At the moment, we can only recommend to those colleagues who register their (*in situ*) data on Zenodo.org, SoilHive.ag or similar, obtain a DOI and then let us know that we can import and integrate your data into these models to help build better soil maps for everyone. The LandPKS app (https://landpotential.org/mobile-app/) could here potentially play an important role if it extends its functionality to soil laboratory data, soil profile photographs, and soil spectral scans.

Aroca-Fernandez et al. (2025) recently developed a framework called WALGREEN, which runs on top of Google Earth Engine and the Copernicus Data Space Ecosystem, and shows how to generate SOC predictions *on-demand* (that is: select an area of interest, drop training points, and download SOC predictions). WALGREEN is primarily based on EO images, and the performance in terms of accuracy and robustness is currently unknown. Automation of modeling, predictions, even automated deployment of maps in an app is a reality today. However, knowledge of soil science is needed more than ever. The application of all these (AI) tools has become much easier, so we have to spend much less time in actually implementing machine learning and statistics, but at the same time we still need to know what we do and have a clear picture of what to put the outcome into perspective, to actually be able to have a proper interpretation of the outcome and the results; and without a sufficient background knowledge we will not be able to put it into perspective.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Conclusion

We have produced the first batch of 30 m resolution soil property annual dynamic maps for 2000–2024. This is an open data output aiming at serving international modeling and monitoring projects, especially the United Nations Convention to Combat Desertification (UNCCD) Land Degradation Neutrality programme, FAO's Global Soil Partnership and similar, to help countries and land owners get baselines of their soil stocks and better understand the soil dynamics. For modeling, we use state-of-the-art harmonized soil laboratory data sets (training points) that we have been collecting and improving over the years. We make import procedures transparent and fully documented via https://soildb.OpenLandMap.org/. The results indicate that (R1) Landsat-derived biophysical indices rank high in the variable importance results, especially with GPP coming in the top three most important variables and showing a clear positive correlation with SOC density. These results indicate clear interconnection between GPP and SOC accumulation, and a warning to all land use systems that decrease annual GPP — this will likely also result in significant losses in SOC. The prediction accuracy assessment indicates that (R2) the best achievable mapping accuracy is RMSE of 17.7 [kg m^{-3}] (0.486 in log-scale) for SOC density, RMSE of 51.3 [g kg^{-1}] (0.574 in log-scale) for SOC content, RMSE of 0.15 [t m^{-3}] for bulk density of fine-earth, RMSE of 0.51 for soil pH, RMSE of 8.4% for soil clay content, and RMSE of 12.6% for soil sand content, respectively. Further analysis of trends in SOC density and soil pH indicates that (R3) the key drivers of negative changes in SOC is land degradation, primarily conversion of tropical forests to cash crops; for soil pH, the most important explanatory variable appears to be CHELSA Aridity Index (long-term), annual precipitation, and salinity grade. Finally, (R4) the remaining hot spots of global SOC are boreal peatlands (of Canada and Russia) and tropical peatlands storing majority of the total soil carbon. Especially Canada and the Russian Federation seem to contain most of the world's soil carbon. We estimate that the world has lost at least 11 Pg of SOC in the top soil in the period 2000–2022 and that the current SOC stock of the land is 461 Pg for 0–30 cm.

The 30 m resolution predictions of soil properties and USDA subgroups show an unprecedented level of detail; however, we warn users that prediction errors are still relatively wide and that this uncertainty (per pixel) should be incorporated into decision making to prevent taking high-risk decisions. Further recommended uses of these data include: continental soil carbon dynamics monitoring, derivation of secondary soil variables such as soil hydraulic properties using pedotransfer functions, and land degradation and soil health assessment.

We plan to update these predictions for each subsequent year and also as the new point data sets become available and as we receive feedback from local experts. We currently engage in Data-Sharing Agreements for various soil datasets that are not in the public domain, and we would like to engage with more countries outside of Europe in a similar manner. In addition, for those countries which consider accurate position of sample locations (GPS coordinates) to be private data, we would like to extend this to the federated model space, which would further enable global products to be aligned with (finer resolution) national products becoming available in a few jurisdictions. Additional improvements in the accuracy of the dynamic predictions can be implemented by combining global and local models, especially at national scales. The only requirement for further data fusion is that all parties use the same standard (e.g. 0–30 cm; 5–year time blocks; DC reference laboratory methods and similar).

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

We call on research groups to use these data to derive secondary soil properties and test their applicability for real-world applications.

5 Data availability

The produced data products can be accessed at https://doi.org/10.5281/zenodo.15470431 (Consoli et al., 2025), while the training dataset is available at https://doi.org/10.5281/zenodo.4748499 (Hengl and Gupta, 2025). Due to Zenodo's storage limitations and the large size of the dataset, only portions of the data are stored in Zenodo, distributed across multiple buckets. We provide predictions of soil properties in 120 m resolution with uncertainty (16th percentile, mean, and 84th percentile) for only the first and the last period (2000–2005 and 2020–2022+). In addition, we provide soil type probability maps in 120 m resolution based on USDA Soil Taxonomy, organized at the subgroup level. Complete global 30 m resolution mosaics are available through the Google Earth Engine (https://code.earthengine.google.com/?asset=projects/global-pasture-watch/assets/gsm-30m).

Each data set layer follows a standardized naming format, structured into 10 key fields: generic variable name, variable procedure combination, position in the probability distribution or variable type, spatial support, depth reference, time reference (including start and end times), bounding box, EPSG code, and version code. Each metadata field serves a specific purpose in assessing the datasets' fitness for use. For the data sets presented in this study, the metadata specify a uniform spatial support of 30 m resolution, a depth reference denoted as *s* (depth from the surface), a bounding box identified as *go*, and an EPSG code of *EPSG:4326*. For the other fields: the generic variable name helps users identify the required predictor layer; the variable procedure combination provides indications of how the data were derived and its source; the time reference, comprising the start and end dates, ensures users can select layers matching their relevant temporal scope; and the version code, which reflects the creation date of the corresponding layer, facilitates tracking and version control.

6 Code availability

The code used to harmonize the training points and generate predictions is available under the MIT license at https://doi.org/10.5281/zenodo.15608971 (Hengl et al., 2025).

7 Competing interests

Tomislav Hengl, Davide Consoli, Xuemeng Tian, Mustafa Serkan Isik, Leandro Parente, Yu-Feng Ho & Rolf Simoes are employed by OpenGeoHub.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Acknowledgments

This research was supported by multiple grants, primarily by the Land & Carbon Lab grant from the Bezos Earth Fund. The Open-Earth-Monitor Cyberinfrastructure project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101059548. The AI4SoilHealth project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101086179.

We are grateful to the USDA NRCS National Soil Survey Center, the Africa Soil Information Service (AfSIS), the European Soil Data Center, ISRIC — World Soil Information, and the soil data community for publishing and providing high-quality soil samples and observations, without which it would not have been possible to produce these maps. The Landsat bands were made available, with many thanks, from the University of Maryland's GLAD lab.

The authors acknowledge the use of a generative AI language model, ChatGPT developed by OpenAI, to assist in improving the clarity and readability of the manuscript.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

References

Aroca-Fernandez, J. M., Diez-Pastor, J. F., Latorre-Carmona, P., Elvira, V., Camps-Valls, G., Pascual, R., and Garcia-Osorio, C.: A Collaborative Platform for Soil Organic Carbon Inference Based on Spatiotemporal Remote Sensing Data, https://doi.org/10.48550/ARXIV.2504.13962, 2025.

Atwood, T. B., Connolly, R. M., Almahasheer, H., Carnell, P. E., Duarte, C. M., Ewers Lewis, C. J., Irigoien, X., Kelleway, J. J., Lavery, P. S., Macreadie, P. I., et al.: Global patterns in mangrove soil carbon stocks and losses, Nature Climate Change, 7, 523–528, https://doi.org/10.1038/nclimate3326, 2017.

Batjes, N. H., Calisto, L., and de Sousa, L. M.: Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023), Earth System Science Data, 16, 4735–4765, https://doi.org/10.5194/essd-16-4735-2024, 2024.

Bauer-Marschallinger, B., Sabel, D., and Wagner, W.: Optimisation of global grids for high-resolution remote sensing data, Computers & Geosciences, 72, 84–93, https://doi.org/10.1016/j.cageo.2014.07.005, 2014.

Behrens, T., Schmidt, K., MacMillan, R. A., and Rossel, R. V.: Multiscale contextual spatial modelling with the Gaussian scale space, Geoderma, 310, 128–137, https://doi.org/10.1016/j.geoderma.2017.09.015, 2018.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A foundation model for the Earth system, Nature, 641, 1180–1187, https://doi.org/10.1038/s41586-025-09005-y, 2025.

Börker, J., Hartmann, J., Amann, T., and Romero-Mujalli, G.: Terrestrial sediments of the earth: development of a global unconsolidated sediments map database (GUM), Geochemistry, Geophysics, Geosystems, 19, 997–1024, https://doi.org/10.1002/2017GC007273, 2018.

Borrelli, P., Robinson, D. A., Fleischer, L. R., Lugato, E., Ballabio, C., Alewell, C., Meusburger, K., Modugno, S., Schütt, B., Ferro, V., Bagarello, V., Oost, K. V., Montanarella, L., and Panagos, P.: An assessment of the global impact of 21st century land use change on soil erosion, Nature Communications, 8, 2013, https://doi.org/10.1038/s41467-017-02142-7, 2017.

Broeg, T., Don, A., Wiesmeier, M., Scholten, T., and Erasmi, S.: Spatiotemporal Monitoring of Cropland Soil Organic Carbon Changes From Space, Global Change Biology, 30, e17 608, https://doi.org/https://doi.org/10.1111/gcb.17608, e17608 GCB-24-2377.R1, 2024.

Brus, D.: Spatial Sampling with R, Chapman & Hall/CRC The R Series, CRC Press, Boca Raton, ISBN 9781000600056, 2022.

Chatterjee, A., Lal, R., Wielopolski, L., Martin, M. Z., and Ebinger, M.: Evaluation of different soil carbon determination methods, Critical Reviews in Plant Science, 28, 164–178, 2009.

Chen, S., Arrouays, D., Mulder, V. L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., et al.: Digital mapping of GlobalSoilMap soil properties at a broad scale: A review, Geoderma, 409, 115 567, https://doi.org/10.1016/j.geoderma.2021.115567, 2022.

Chen, Z., Auler, A. S., Bakalowicz, M., Drew, D., Griger, F., Hartmann, J., Jiang, G., Moosdorf, N., Richts, A., Stevanovic, Z., et al.: The World Karst Aquifer Mapping project: concept, mapping procedure and map of Europe, Hydrogeology Journal, 25, 771, https://doi.org/10.1007/s10040-016-1519-3, 2017.

Consoli, D., Parente, L., Simoes, R., Şahin, M., Tian, X., Witjes, M., Sloat, L., and Hengl, T.: A computational framework for processing time-series of Earth Observation data based on discrete convolution: global-scale historical Landsat cloud-free aggregates at 30 m spatial resolution, PeerJ, accepted for publication, https://doi.org/10.21203/rs.3.rs-4465582/v1, preprint posted at Research Square, 2024.

Consoli, D., Tian, X., Isik, S., Simoes, R., and Hengl, T.: OpenLandMap-soildb: soil organic carbon density (mg/cm3) 2000-2005 0-30cm below ground, https://doi.org/10.5281/zenodo.15470432, 2025.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Cowie, A. L., Orr, B. J., Sanchez, V. M. C., Chasek, P., Crossman, N. D., Erlewein, A., Louwagie, G., Maron, M., Metternicht, G. I., Minelli, S., et al.: Land in balance: The scientific conceptual framework for Land Degradation Neutrality, Environmental Science & Policy, 79, 25–35, https://doi.org/10.1016/j.envsci.2017.10.011, 2018.

Crowther, T. W., Todd-Brown, K. E. O., Rowe, C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B., Snoek, B. L., Fang, S., Zhou, G., Allison, S. D., Blair, J. M., Bridgham, S. D., Burton, A. J., Carrillo, Y., Reich, P. B., Clark, J. S., Classen, A. T., Dijkstra, F. A., Elberling, B., Emmett, B. A., Estiarte, M., Frey, S. D., Guo, J., Harte, J., Jiang, L., Johnson, B. R., Kröel-Dulay, G., Larsen, K. S., Laudon, H., Lavallee, J. M., Luo, Y., Lupascu, M., Ma, L. N., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L. L., Schmidt, I. K., Sistla, S., Sokol, N. W., Templer, P. H., Treseder, K. K., Welker, J. M., and Bradford, M. A.: Quantifying global soil carbon losses in response to warming, Nature, 540, 104–108, https://doi.org/10.1038/nature20150, 2016.

CSIRO: CSIRO National Soil Site Database. v10, CSIRO, Canberra, Australia, https://doi.org/https://doi.org/10.25919/c4br-0r30, 2024.

Dangal, S., Sanderman, J., Wills, S., and Ramirez-Lopez, L.: Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library, Soil Systems, 3, 11, https://doi.org/10.3390/soilsystems3010011, 2019.

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I. D., Van Vliet, J., and Bechtel, B.: A global map of Local Climate Zones to support earth system modelling and urban scale environmental science, Earth System Science Data Discussions, 2022, 1–57, https://doi.org/10.5194/essd-14-3835-2022, 2022.

Díaz-Guadarrama, S., Varón-Ramírez, V. M., Lizarazo, I., Guevara, M., Angelini, M., Araujo-Carrillo, G. A., Argeñal, J., Armas, D., Balta, R. A., Bolivar, A., Bustamante, N., Dart, R. O., Dell Acqua, M., Encina, A., Figueredo, H., Fontes, F., Gutiérrez-Díaz, J. S., Jiménez, W., Lavado, R. S., Mansilla-Baca, J. F., de Lourdes Mendonça-Santos, M., Moretti, L. M., Muñoz, I. D., Olivera, C., Olmedo, G., Omuto, C., Ortiz, S., Pascale, C., Pfeiffer, M., Ramos, I. A., Ríos, D., Rivera, R., Rodriguez, L. M., Rodríguez, D. M., Rosales, A., Rosales, K., Schulz, G., Sevilla, V., Tenti, L. M., Vargas, R., Vasques, G. M., Yigini, Y., and Rubiano, Y.: Improving the Latin America and Caribbean Soil Information System (SISLAC) database enhances its usability and scalability, Earth System Science Data, 16, 1229–1246, https://doi.org/10.5194/essd-16-1229-2024, 2024.

Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., Preimesberger, W., Xaver, A., Annor, F., Ardö, J., Baldocchi, D., Bitelli, M., Blöschl, G., Bogena, H., Brocca, L., Calvet, J.-C., Camarero, J. J., Capello, G., Choi, M., Cosh, M. C., van de Giesen, N., Hajdu, I., Ikonen, J., Jensen, K. H., Kanniah, K. D., de Kat, I., Kirchengast, G., Kumar Rai, P., Kyrouac, J., Larson, K., Liu, S., Loew, A., Moghaddam, M., Martínez Fernández, J., Mattar Bader, C., Morbidelli, R., Musial, J. P., Osenga, E., Palecki, M. A., Pellarin, T., Petropoulos, G. P., Pfeil, I., Powers, J., Robock, A., Rüdiger, C., Rummel, U., Strobel, M., Su, Z., Sullivan, R., Tagesson, T., Varlagin, A., Vreugdenhil, M., Walker, J., Wen, J., Wenger, F., Wigneron, J. P., Woods, M., Yang, K., Zeng, Y., Zhang, X., Zreda, M., Dietrich, S., Gruber, A., van Oevelen, P., Wagner, W., Scipal, K., Drusch, M., and Sabia, R.: The International Soil Moisture Network: serving Earth system science for over a decade, Hydrology and Earth System Sciences, 25, 5749–5804, https://doi.org/10.5194/hess-25-5749-2021, 2021.

Drake, T. W., Van Oost, K., Barthel, M., Bauters, M., Hoyt, A. M., Podgorski, D. C., Six, J., Boeckx, P., Trumbore, S. E., Cizungu Ntaboba, L., et al.: Mobilization of aged and biolabile soil carbon by tropical deforestation, Nature geoscience, 12, 541–546, https://doi.org/10.1038/s41561-019-0384-9, 2019.

ESA: Land Cover CCI Product User Guide Version 2, European Space Agency (ESA) Climate Change Initiative (CCI), https://www.esa-landcover-cci.org/?q=webfm_send/84, 2017.

Even, R. J., Machmuller, M. B., Lavallee, J. M., Zelikova, T. J., and Cotrufo, M. F.: Large errors in common soil carbon measurements due to sample processing, https://doi.org/10.5194/egusphere-2024-1470, 2024.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

FAO: Global Soil Organic Carbon Map – GSOCmap v.1.6: Technical report, Food & Agriculture Org., Rome, Italy, ISBN 9789251358993, https://doi.org/10.4060/cb9015en, 2022.

FAO & IIASA: Harmonized World Soil Database version 2.0, FAO, Rome and Laxenburg, https://doi.org/10.4060/cc3823en, 2023.

Fatoyinbo, L.: Vast peatlands found in the Congo Basin, Nature, 542, 38–39, https://doi.org/10.1038/542038b, 2017.

Feeney, C., Cosby, B., Robinson, D., Thomas, A., Emmett, B., and Henrys, P.: Multiple soil map comparison highlights challenges for predicting topsoil organic carbon concentration at national scale, Scientific Reports, 12, 1379, https://doi.org/10.1038/s41598-022-05476-5, 2022.

Gallios, G., Tsakiridis, N., and Tziolas, N.: Federated learning applications in soil spectroscopy, Geoderma, 456, 117 259, https://doi.org/https://doi.org/10.1016/j.geoderma.2025.117259, 2025.

Gautam, S., Mishra, U., Scown, C. D., Wills, S. A., Adhikari, K., and Drewniak, B. A.: Continental United States may lose 1.8 petagrams of soil organic carbon under climate change by 2100, Global Ecology and Biogeography, 31, 1147–1160, https://doi.org/10.1111/geb.13489, 2022.

Geng, X., Fraser, W., VandenBygaart, B., Smith, S., Waddell, A., Jiao, Y., and Patterson, G.: Toward digital soil mapping in Canada: Existing soil survey data and related expert knowledge, Digital soil mapping: bridging research, environmental application, and operation, pp. 325–335, https://doi.org/10.1007/978-90-481-8863-5_26, 2010.

Gottschalk, P., Smith, J., Wattenbach, M., Bellarby, J., Stehfest, E., Arnell, N., Osborn, T. J., Jones, C., and Smith, P.: How will organic carbon stocks in mineral soils evolve under future climate? Global projections using RothC for a range of climate change scenarios, Biogeosciences, 9, 3151–3171, https://doi.org/10.5194/bg-9-3151-2012, 2012.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., et al.: No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America, SOIL Discussions, 2018, 1–20, 2018.

Gumbricht, T., Roman-Cuesta, R. M., Verchot, L., Herold, M., Wittmann, F., Householder, E., Herold, N., and Murdiyarso, D.: An expert system model for mapping tropical wetlands and peatlands reveals South America as the largest contributor, Global Change Biology, 23, 3581–3599, https://doi.org/10.1111/gcb.13689, 2017.

Guo, L. B. and Gifford, R. M.: Soil carbon stocks and land use change: a meta analysis, Global Change Biology, 8, 345–360, https://doi.org/10.1046/j.1354-1013.2002.00486.x, 2002.

Gyawali, A. J., Wiseman, M., Ackerson, J. P., Coffman, S., Meissner, K., and Morgan, C. L.: Measuring in situ soil carbon stocks: A study using a novel handheld VisNIR probe, Geoderma, 453, 117 152, https://doi.org/https://doi.org/10.1016/j.geoderma.2024.117152, 2025.

Hackländer, J., Parente, L., Ho, Y.-F., Hengl, T., Simoes, R., Consoli, D., Şahin, M., Tian, X., Jung, M., Herold, M., et al.: Land potential assessment and trend-analysis using 2000–2021 FAPAR monthly time-series at 250 m spatial resolution, PeerJ, 12, e16 972, https://doi.org/10.7717/peerj.16972, 2024.

Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., Loisel, J., Malhotra, A., Jackson, R. B., Ogle, S., Phillips, C., Ryals, R., Todd-Brown, K., Vargas, R., Vergara, S. E., Cotrufo, M. F., Keiluweit, M., Heckman, K. A., Crow, S. E., Silver, W. L., DeLonge, M., and Nave, L. E.: Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter, Global Change Biology, 24, e705–e718, https://doi.org/https://doi.org/10.1111/gcb.13896, 2018.

Hastie, T., Tibshirani, R., and Wainwright, M.: Statistical learning with sparsity, Monographs on statistics and applied probability, 143, 8, 2015.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Helfenstein, A., Mulder, V. L., Hack-ten Broeke, M. J., van Doorn, M., Teuling, K., Walvoort, D. J., and Heuvelink, G.: BIS-4D: mapping soil properties and their uncertainties at 25 m resolution in the Netherlands, Earth System Science Data, 16, 2941–2970, https://doi.org/10.5194/essd-16-2941-2024, 2024.

Hempel, J., McBratney, A. B., Arrouays, D., McKenzie, N. J., and Hartemink, A.: GlobalSoilMap project history, in: GlobalSoilMap: Basis of the global spatial soil information system; Proceedings of the 1st GlobalSoilMap conference, Taylor & Francis Group, Orléans, France, 2014.

Hendriks, C., Stoorvogel, J., Lutz, F., and Claessens, L.: When can legacy soil data be used, and when should new data be collected instead?, Geoderma, 348, 181–188, https://doi.org/10.1016/j.geoderma.2019.04.026, 2019.

Hengl, T. and Gupta, S.: An Open Compendium of Soil Datasets: Soil Observations and Measurements, https://doi.org/10.5281/zenodo.15593990, 2025.

Hengl, T. and MacMillan, R.: Predictive Soil Mapping with R, OpenGeoHub foundation, Wageningen, the Netherlands, ISBN 978-0-359-30635-0, https://soilmapper.org, 2019.

Hengl, T., Nikolić, M., and MacMillan, R.: Mapping efficiency and information content, International Journal of Applied Earth Observation and Geoinformation, 22, 127–138, https://doi.org/10.1016/j.jag.2012.02.005, 2013.

Hengl, T., Jesus, J. M. d., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLOS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., Yemefack, M., Wendt, J., MacMillan, R. A., Wheeler, I., and Crouch, J.: African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning, Scientific Reports, 11, 6130, https://doi.org/10.1038/s41598-021-85639-y, 2021.

Hengl, T., Consoli, Davide, T. X., and Isik, M. S.: openlandmap/soildb: v0.0.2, https://doi.org/10.5281/zenodo.15608972, 2025.

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., Van Den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., and Sanderman, J.: Machine learning in space and time for modelling soil organic carbon change, European Journal of Soil Science, 72, 1607–1623, https://doi.org/10.1111/ejss.12998, 2021.

Ho, Y.-f., Parente, L., Lindsay, J., Grohmann, C. H., Reuter, H. I., and Hengl, T.: Global Ensemble Digital Terrain modeling and parametrization at 30 m resolution (GEDTM30): a data fusion approach based on ICESat-2, GEDI and multisource data, PeerJ, in review, 1–39, https://doi.org/10.21203/rs.3.rs-6280607/v1, 2025.

Hou, D., Jia, X., Wang, L., McGrath, S. P., Zhu, Y.-G., Hu, Q., Zhao, F.-J., Bank, M. S., O'Connor, D., and Nriagu, J.: Global soil pollution by toxic metals threatens agriculture and human health, Science, 388, 316–321, https://doi.org/10.1126/science.adr5214, 2025.

Hugelius, G., Bockheim, J. G., Camill, P., Elberling, B., Grosse, G., Harden, J. W., Johnson, K., Jorgenson, T., Koven, C. D., Kuhry, P., Michaelson, G., Mishra, U., Palmtag, J., Ping, C.-L., O'Donnell, J., Schirrmeister, L., Schuur, E. A. G., Sheng, Y., Smith, L. C., Strauss, J., and Yu, Z.: A new data set for estimating organic carbon storage to 3 m depth in soils of the northern circumpolar permafrost region, Earth System Science Data, 5, 393–402, https://doi.org/10.5194/essd-5-393-2013, 2013a.

Hugelius, G., Tarnocai, C., Broll, G., Canadell, J. G., Kuhry, P., and Swanson, D.: The Northern Circumpolar Soil Carbon Database: spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions, Earth System Science Data, 5, 3–13, 2013b.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Isik, M. S., Parente, L., and Consoli, D. e. a.: Light Use Efficiency (LUE) based bimonthly Gross Primary Productivity (GPP) for global grasslands at 30 m spatial resolution (2000-2022), PeerJ, in review, https://doi.org/10.21203/rs.3.rs-5587863/v1, 2025.

Iversen, C. M., McCormack, M. L., Powell, A. S., Blackwood, C. B., Freschet, G. T., Kattge, J., Roumet, C., Stover, D. B., Soudzilovskaia, N. A., Valverde-Barrantes, O. J., et al.: A global Fine-Root Ecology Database to address below-ground challenges in plant ecology, New Phytologist, 215, 15–26, https://doi.org/10.1111/nph.14486, 2017.

Ivushkin, K., Bartholomeus, H., Bregt, A. K., Pulatov, A., Kempen, B., and de Sousa, L.: Global mapping of soil salinity change, Remote Sensing of Environment, 231, 111 260, https://doi.org/https://doi.org/10.1016/j.rse.2019.111260, 2019.

Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., and Piñeiro, G.: The ecology of soil carbon: pools, vulnerabilities, and biotic and abiotic controls, Annual review of ecology, evolution, and systematics, 48, 419–445, 2017.

Jenny, H.: Factors of soil formation: a system of quantitative pedology, Courier Corporation, 1994.

Jian, J., Du, X., and Stewart, R. D.: A database for global soil health assessment, Scientific Data, 7, 16, https://doi.org/10.1038/s41597-020-0356-3, 2020.

Jones, C., McConnell, C., Coleman, K., Cox, P., Falloon, P., Jenkinson, D., and Powlson, D.: Global climate change and soil carbon stocks; predictions from two contrasting models for the turnover of organic carbon in soil, Global Change Biology, 11, 154–166, https://doi.org/10.1111/j.1365-2486.2004.00885.x, 2005.

Kalopesa, E., Tziolas, N., Tsakiridis, N. L., Safanelli, J. L., Hengl, T., and Sanderman, J.: Large-Scale Soil Organic Carbon Estimation via a Multisource Data Fusion Approach, Remote Sensing, 17, https://doi.org/10.3390/rs17050771, 2025.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M.: Climatologies at high resolution for the earth's land surface areas, Scientific data, 4, 1–20, https://doi.org/10.1038/sdata.2017.122, 2017.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems, edited by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., vol. 30, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf, 2017.

Keel, S. G., Anken, T., Büchi, L., Chervet, A., Fliessbach, A., Flisch, R., Huguenin-Elie, O., Mäder, P., Mayer, J., Sinaj, S., et al.: Loss of soil organic carbon in Swiss long-term agricultural experiments over a wide range of management practices, Agriculture, Ecosystems & Environment, 286, 106 654, https://doi.org/10.1016/j.agee.2019.106654, 2019.

Klein, I., Gessner, U., Dietz, A. J., and Kuenzer, C.: Global WaterPack – A 250m resolution dataset revealing the daily dynamics of global inland water bodies, Remote Sensing of Environment, 198, 345–362, https://doi.org/https://doi.org/10.1016/j.rse.2017.06.045, 2017.

Kraamwinkel, C. T., Beaulieu, A., Dias, T., and Howison, R. A.: Planetary limits to soil degradation, Communications Earth & Environment, 2, 249, https://doi.org/10.1038/s43247-021-00323-3, 2021.

Krasilnikov, P., Martí, J.-J. I., Arnold, R., and Shoba, S.: A handbook of soil terminology, correlation and classification, Routledge, London, https://doi.org/10.4324/9781849774352, 2009.

Leenaars, J., van Oostrum, A., and Gonzalez, M. R.: Africa Soil Profiles Database, Version 1.2, A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Report, 1, https://www.isric.org/projects/africa-soil-profiles-database-afsp, 2014.

Lembrechts, J. J., van den Hoogen, J., Aalto, J., Ashcroft, M. B., De Frenne, P., Kemppinen, J., Kopeckỳ, M., Luoto, M., Maclean, I. M., Crowther, T. W., et al.: Global maps of soil temperature, Global change biology, 28, 3110–3144, 2022.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Li, T., Cui, L., Kuhnert, M., McLaren, T. I., Pandey, R., Liu, H., Wang, W., Xu, Z., Xia, A., Dalal, R. C., and Dang, Y. P.: A comprehensive review of soil organic carbon estimates: Integrating remote sensing and machine learning technologies, Journal of Soils and Sediments, 24, 3556–3571, https://doi.org/10.1007/s11368-024-03913-8, 2024.

Liang, Z., Chen, S., Yang, Y., Zhou, Y., and Shi, Z.: High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling, Science of The Total Environment, 685, 480–489, https://doi.org/10.1016/j.scitotenv.2019.05.332, 2019.

Lin, Z., Dai, Y., Mishra, U., Wang, G., Shangguan, W., Zhang, W., and Qin, Z.: On the magnitude and uncertainties of global and regional soil organic carbon: A comparative analysis using multiple estimates, https://doi.org/10.5194/essd-2022-232, 2022.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A.-X., and Zhang, G.-L.: Mapping high resolution national soil information grids of China, Science Bulletin, 67, 328–340, https://doi.org/10.1016/j.scib.2021.10.013, 2022.

Liu, F. T., Ting, K. M., and Zhou, Z.-H.: Isolation forest, in: 2008 eighth ieee international conference on data mining, pp. 413–422, IEEE, https://doi.org/https://doi.org/10.1109/ICDM.2008.17, 2008.

Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., Zhu, Q., Till, J., Jia, X., Jiang, C., Wang, S., Qin, Z., Kong, H., Grant, R., Mezbahuddin, S., Kumar, V., and Jin, Z.: Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems, Nature Communications, 15, 357, https://doi.org/10.1038/s41467-023-43860-5, 2024.

López-Ballesteros, A., Nielsen, A., Castellanos-Osorio, G., Trolle, D., and Senent-Aparicio, J.: DSOLMap, a novel high-resolution global digital soil property map for the SWAT + model: Development and hydrological evaluation, CATENA, 231, 107 339, https://doi.org/10.1016/j.catena.2023.107339, 2023.

Manies, K., Waldrop, M., and Harden, J.: Generalized models to estimate carbon and nitrogen stocks of organic soil horizons in Interior Alaska, Earth System Science Data, 12, 1745–1757, https://doi.org/10.5194/essd-12-1745-2020, 2020.

Maxwell, T. L., Hengl, T., Parente, L. L., Minarik, R., Worthington, T. A., Bunting, P., Smart, L. S., Spalding, M. D., and Landis, E.: Global mangrove soil organic carbon stocks dataset at 30 m resolution for the year 2020 based on spatiotemporal predictive machine learning, Data in Brief, 50, 109 621, https://doi.org/10.1016/j.dib.2023.109621, 2023.

Meinshausen, N.: Quantile Regression Forests, Journal of Machine Learning Research, 7, 983–999, http://jmlr.org/papers/v7/meinshausen06a.html, 2006.

Melton, J. R., Chan, E., Millard, K., Fortier, M., Winton, R. S., Martín-López, J. M., Cadillo-Quiroz, H., Kidd, D., and Verchot, L. V.: A map of global peatland extent created using machine learning (Peat-ML), Geoscientific Model Development, 15, 4709–4738, https://doi.org/10.5194/gmd-15-4709-2022, 2022.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods in Ecology and Evolution, 12, 1620–1633, https://doi.org/https://doi.org/10.1111/2041-210X.13650, 2021.

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., et al.: Soil carbon 4 per mille, Geoderma, 292, 59–86, 2017.

Minasny, B., Bandai, T., Ghezzehei, T. A., Huang, Y.-C., Ma, Y., McBratney, A. B., Ng, W., Norouzi, S., Padarian, J., Rudiyanto, Sharififar, A., Styc, Q., and Widyastuti, M.: Soil Science-Informed Machine Learning, Geoderma, 452, 117 094, https://doi.org/https://doi.org/10.1016/j.geoderma.2024.117094, 2024.

Montgomery, D. R.: Soil erosion and agricultural sustainability, Proceedings of the National Academy of Sciences, 104, 13 268–13 272, https://doi.org/10.1073/pnas.0611508104, 2007.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

National Academies of Sciences, Engineering, and Medicine and others: Exploring a dynamic soil information system: proceedings of a workshop, The National Academies Press, Washington, DC, https://doi.org/10.17226/26170, 2021.

Nauman, T. W., Kienast-Brown, S., Roecker, S. M., Brungard, C., White, D., Philippe, J., and Thompson, J. A.: Soil landscapes of the United States (SOLUS): Developing predictive soil property maps of the conterminous United States using hybrid training sets, Soil Science Society of America Journal, 88, 2046–2065, https://doi.org/10.1002/saj2.20769, 2024.

Naval, M. L. M., Bieluczyk, W., Alvarez, F., da Silva Carvalho, L. C., Maracahipes-Santos, L., de Oliveira, E. A., da Silva, K. G., Pereira, M. B., Brando, P. M., Junior, B. H. M., et al.: Impacts of repeated forest fires and agriculture on soil organic matter and health in southern Amazonia, CATENA, 254, 108 924, https://doi.org/10.1016/j.catena.2025.108924, 2025.

Nenkam, A. M., Wadoux, A. M.-C., Minasny, B., McBratney, A. B., Traore, P. C., Falconnier, G. N., and Whitbread, A. M.: Using homosoils for quantitative extrapolation of soil mapping models, European Journal of Soil Science, 73, e13 285, https://doi.org/https://doi.org/10.1111/ejss.13285, 2022.

Ng, W., Minasny, B., McBratney, A., de Caritat, P., and Wilford, J.: Digital soil mapping of lithium in Australia, Earth System Science Data, 15, 2465–2482, https://doi.org/10.5194/essd-15-2465-2023, 2023.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., and Fernández-Ugalde, O.: LUCAS Soil, the largest expandable soil dataset for Europe: a review, European Journal of Soil Science, 69, 140–153, https://doi.org/10.1111/ejss.12499, 2018.

Padarian, J., Minasny, B., McBratney, A., and Smith, P.: Soil carbon sequestration potential in global croplands, PeerJ, 10, e13 740, https://doi.org/10.7717/peerj.13740, 2022a.

Padarian, J., Stockmann, U., Minasny, B., and McBratney, A.: Monitoring changes in global soil organic carbon stocks from space, Remote Sensing of Environment, 281, 113 260, https://doi.org/10.1016/j.rse.2022.113260, 2022b.

Panagos, P., Montanarella, L., Barbero, M., Schneegans, A., Aguglia, L., and Jones, A.: Soil priorities in the European Union, Geoderma Regional, 29, e00 510, 2022.

Paz-Pellat, F. and Velázquez-Rodríguez, A. S.: Base de datos de perfiles de suelos en México, Elementos para Políticas Públicas, 2, 210–235, https://www.elementospolipub.org/ojs/index.php/epp/article/view/16, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, https://www.jmlr.org/papers/v12/pedregosa11a.html, 2011.

Pfeiffer, M., Padarian, J., Osorio, R., Bustamante, N., Olmedo, G. F., Guevara, M., Aburto, F., Albornoz, F., Antilén, M., Araya, E., Arellano, E., Barret, M., Barrera, J., Boeckx, P., Briceño, M., Bunning, S., Cabrol, L., Casanova, M., Cornejo, P., Corradini, F., Curaqueo, G., Doetterl, S., Duran, P., Escudey, M., Espinoza, A., Francke, S., Fuentes, J. P., Fuentes, M., Gajardo, G., García, R., Gallaud, A., Galleguillos, M., Gomez, A., Hidalgo, M., Ivelic-Sáez, J., Mashalaba, L., Matus, F., Meza, F., Mora, M. D. L. L., Mora, J., Muñoz, C., Norambuena, P., Olivera, C., Ovalle, C., Panichini, M., Pauchard, A., Pérez-Quezada, J. F., Radic, S., Ramirez, J., Riveras, N., Ruiz, G., Salazar, O., Salgado, I., Seguel, O., Sepúlveda, M., Sierra, C., Tapia, Y., Tapia, F., Toledo, B., Torrico, J. M., Valle, S., Vargas, R., Wolff, M., and Zagal, E.: CHLSOC: the Chilean Soil Organic Carbon database, a multi-institutional collaborative effort, Earth System Science Data, 12, 457–468, https://doi.org/10.5194/essd-12-457-2020, 2020.

Poeplau, C., Jacobs, A., Don, A., Vos, C., Schneider, F., Wittnebel, M., Tiemeyer, B., Heidkamp, A., Prietz, R., and Flessa, H.: Stocks of organic carbon in German agricultural soils — Key results of the first comprehensive inventory, Journal of Plant Nutrition and Soil Science, 183, 665–681, https://doi.org/10.1002/jpln.202000113, 2020.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Polidoro, J., Coelho, M., Filho, A. D. C., Lumbreras, J., De Oliveira, A., Vasques, G. d. M., Macario, C. d. N., Victoria, D. d. C., Bhering, S., De Freitas, P., et al.: National program for surveying and interpreting soils in Brazil (PronaSolos): guidelines for implementation, Documentos (INFOTECA-E), Embrapa Solos, Brazilia, BR, https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1135056, 2021.

Potapov, P., Hansen, M. C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., and Ying, Q.: Landsat analysis ready data for global land cover and land cover change mapping, Remote Sensing, 12, 426, https://doi.org/10.3390/rs12030426, 2020.

Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century, Nature Food, 3, 19–28, https://doi.org/10.1038/s43016-021-00429-z, 2022.

Quandt, A., Herrick, J., and Bouvier, I.: LANDPKS: a new mobile tool for sustainable land-use planning and management, in: Proceedings of the 2018 World Bank Conference on Land and Poverty, vol. 22, Washington, DC, USA, 2018.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., and Thompson, J.: Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution, Soil Science Society of America Journal, 82, 186–201, https://doi.org/10.2136/sssaj2017.04.0122, 2018.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., and Thuiller, W.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Safanelli, J. L., Sanderman, J., Bloom, D., Todd-Brown, K., Parente, L. L., Hengl, T., Adam, S., Albinet, F., Ben-Dor, E., Boot, C. M., Bridson, J. H., Chabrillat, S., Deiss, L., Demattê, J. A., Scott Demyan, M., Dercon, G., Doetterl, S., van Egmond, F., Ferguson, R., Garrett, L. G., Haddix, M. L., Haefele, S. M., Heiling, M., Hernandez-Allica, J., Huang, J., Jastrow, J. D., Karyotis, K., Machmuller, M. B., Khesuoe, M., Margenot, A., Matamala, R., Miesel, J. R., Mouazen, A. M., Nagel, P., Patel, S., Qaswar, M., Ramakhanna, S., Resch, C., Robertson, J., Roudier, P., Sabetizade, M., Shabtai, I., Sherif, F., Sinha, N., Six, J., Summerauer, L., Thomas, C. L., Toloza, A., Tomczyk-Wójtowicz, B., Tsakiridis, N. L., van Wesemael, B., Woodings, F., Zalidis, G. C., and Żelazny, W. R.: An interlaboratory comparison of mid-infrared spectra acquisition: Instruments and procedures matter, Geoderma, 440, 116 724, https://doi.org/10.1016/j.geoderma.2023.116724, 2023.

Safanelli, J. L., Hengl, T., Parente, L. L., Minarik, R., Bloom, D. E., Todd-Brown, K., Gholizadeh, A., Mendes, W. d. S., and Sanderman, J.: Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement, PloS one, 20, e0296 545, https://doi.org/10.1371/journal.pone.0296545, 2025.

Saha, K., Rudra Bhowmick, U., Anil Kumar, K., Karthika, K., Das, P., and Lalitha, M.: Application of remote sensing in terrestrial soil organic carbon determination: a review, in: Remote Sensing of Soils, pp. 277–293, Elsevier, ISBN 9780443187735, https://doi.org/10.1016/B978-0-443-18773-5.00004-1, 2024.

Sasmito, S. D., Taillardat, P., Adinugroho, W. C., Krisnawati, H., Novita, N., Fatoyinbo, L., Friess, D. A., Page, S. E., Lovelock, C. E., Murdiyarso, D., Taylor, D., and Lupascu, M.: Half of land use carbon emissions in Southeast Asia can be mitigated through peat swamp forest and mangrove conservation and restoration, Nature Communications, 16, 740, https://doi.org/10.1038/s41467-025-55892-0, 2025.

Scharlemann, J. P., Tanner, E. V., Hiederer, R., and Kapos, V.: Global soil carbon: understanding and managing the largest terrestrial carbon pool, Carbon management, 5, 81–91, https://doi.org/10.4155/cmt.13.77, 2014.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Searchinger, T., James, O., and Dumas, P.: Europe's Land Future, Princeton University, Centre for Policy Research, New Jersey, United States, 2022.

Shamrikova, E., Kondratenok, B., Tumanova, E., Vanchikova, E., Lapteva, E., Zonova, T., Lu-Lyan-Min, E., Davydova, A., Libohova, Z., and Suvannang, N.: Transferability between soil organic matter measurement methods for database harmonization, Geoderma, 412, 115 547, https://doi.org/10.1016/j.geoderma.2021.115547, 2022.

Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L., Ji, D., Ye, A., Yuan, H., Zhang, Q., et al.: A China data set of soil properties for land surface modeling, Journal of Advances in Modeling Earth Systems, 5, 212–224, https://doi.org/10.1002/jame.20026, 2013.

Shaw, C., Hilger, A., Filiatrault, M., and Kurz, W.: A Canadian upland forest soil profile and carbon stocks database, Ecology, 99, 989–989, https://doi.org/10.1002/ecy.2159, 2018.

Smith, P., Soussana, J., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., Van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., and Klumpp, K.: How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal, Global Change Biology, 26, 219–241, https://doi.org/10.1111/gcb.14815, 2020.

Soil Survey Staff: Kellogg Soil Survey Laboratory methods manual. Soil Survey Investigations Report No. 42, Version 6.0, U.S. Department of Agriculture, Natural Resources Conservation Service., https://www.nrcs.usda.gov/resources/guides-and-instructions/kssl-guidance, 2022.

Stanimirova, R., Tarrio, K., Turlej, K., McAvoy, K., Stonebrook, S., Hu, K.-T., Arévalo, P., Bullock, E. L., Zhang, Y., Woodcock, C. E., Olofsson, P., Zhu, Z., Barber, C. P., Souza, C. M., Chen, S., Wang, J. A., Mensah, F., Calderón-Loor, M., Hadjikakou, M., Bryan, B. A., Graesser, J., Beyene, D. L., Mutasha, B., Siame, S., Siampale, A., and Friedl, M. A.: A global land cover training dataset from 1984 to 2020, Scientific Data, 10, 879, https://doi.org/10.1038/s41597-023-02798-5, 2023.

Steinwand, D.: Mapping raster imagery to the Interrupted Goode Homolosine projection, International Journal of Remote Sensing, 15, 3463–3471, https://doi.org/10.1080/01431169408954340, 1994.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., De Brogniez, D., Montanarella, L., Hong, S. Y., Rawlins, B. G., and Field, D. J.: Global soil organic carbon assessment, Global Food Security, 6, 9–16, https://doi.org/10.1016/j.gfs.2015.07.001, 2015.

Stumpf, F., Keller, A., Schmidt, K., Mayr, A., Gubler, A., and Schaepman, M.: Spatio-temporal land use dynamics and soil organic carbon in Swiss agroecosystems, Agriculture, Ecosystems & Environment, 258, 129–142, https://doi.org/10.1016/j.agee.2018.02.012, 2018.

Tian, X., De Bruin, S., Simoes, R., Isik, M. S., Minarik, R., Ho, Y.-F., Şahin, M., Herold, M., Consoli, D., and Hengl, T.: Spatiotemporal prediction of soil organic carbon density for Europe (2000–2022) in 3D+T based on Landsat-based spectral indices time-series, PeerJ, https://doi.org/10.21203/rs.3.rs-5128244/v1, 2024.

Tian, X., Consoli, D., Witjes, M., Schneider, F., Parente, L., Şahin, M., Ho, Y.-F., Minařík, R., and Hengl, T.: Time series of Landsat-based bimonthly and annual spectral indices for continental Europe for 2000–2022, Earth System Science Data, 17, 741–772, https://doi.org/10.5194/essd-17-741-2025, 2025.

Tifafi, M., Guenet, B., and Hatté, C.: Large Differences in Global and Regional Total Soil Carbon Stock Estimates Based on SoilGrids, HWSD, and NCSCD: Intercomparison and Evaluation Based on Field Data From USA, England, Wales, and France, Global Biogeochemical Cycles, 32, 42–56, https://doi.org/10.1002/2017GB005678, 2018.

Turubanova, S., Potapov, P., Hansen, M. C., Li, X., Tyukavina, A., Pickens, A. H., Hernandez-Serna, A., Arranz, A. P., Guerra-Hernandez, J., Senf, C., et al.: Tree canopy extent and height change in Europe, 2001–2021, quantified using Landsat data archive, Remote Sensing of Environment, 298, 113 797, https://doi.org/10.1016/j.rse.2023.113797, 2023.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Ugbemuna Ugbaje, S., Karunaratne, S., Bishop, T., Gregory, L., Searle, R., Coelli, K., and Farrell, M.: Space-time mapping of soil organic carbon stock and its local drivers: Potential for use in carbon accounting, Geoderma, 441, 116 771, https://doi.org/10.1016/j.geoderma.2023.116771, 2024.

United States Department of Agriculture and National Cooperative Soil Survey: National Cooperative Soil Survey Lab Data Mart (Lab Data), https://ncsslabdatamart.sc.egov.usda.gov/querypage.aspx, 2023.

Van Gestel, N., Shi, Z., Van Groenigen, K. J., Osenberg, C. W., Andresen, L. C., Dukes, J. S., Hovenden, M. J., Luo, Y., Michelsen, A., Pendall, E., Reich, P. B., Schuur, E. A. G., and Hungate, B. A.: Predicting soil carbon loss with warming, Nature, 554, E4–E5, https://doi.org/10.1038/nature25745, 2018.

Van Tricht, K., Degerickx, J., Gilliams, S., Zanaga, D., Battude, M., Grosu, A., Brombacher, J., Lesiv, M., Bayas, J. C. L., Karanam, S., et al.: WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping, Earth System Science Data, 15, 5491–5515, https://doi.org/10.5194/essd-15-5491-2023, 2023.

van Wesemael, B., Abdelbaki, A., Ben-Dor, E., Chabrillat, S., d'Angelo, P., Demattê, J. A., Genova, G., Gholizadeh, A., Heiden, U., Karlshoefer, P., Milewski, R., Poggio, L., Sabetizade, M., Sanz, A., Schwind, P., Tsakiridis, N., Tziolas, N., Yagüe, J., and Žížala, D.: A European soil organic carbon monitoring system leveraging Sentinel 2 imagery and the LUCAS soil data base, Geoderma, 452, 117 113, https://doi.org/https://doi.org/10.1016/j.geoderma.2024.117113, 2024.

Venter, Z. S., Hawkins, H.-J., Cramer, M. D., and Mills, A. J.: Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa, Science of the Total Environment, 771, 145 384, https://doi.org/10.1016/j.scitotenv.2021.145384, 2021.

Wadoux, A. M.-C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Science Reviews, 210, 103 359, https://doi.org/10.1016/j.earscirev.2020.103359, 2020.

Wagner, J., Martin, V., Speetjens, N. J., A'Campo, W., Durstewitz, L., Lodi, R., Fritz, M., Tanski, G., Vonk, J. E., Richter, A., Bartsch, A., Lantuit, H., and Hugelius, G.: High resolution mapping shows differences in soil carbon and nitrogen stocks in areas of varying landscape history in Canadian lowland tundra, Geoderma, 438, 116 652, https://doi.org/10.1016/j.geoderma.2023.116652, 2023.

Widyastuti, M. T., Minasny, B., Padarian, J., Maggi, F., Aitkenhead, M., Beucher, A., Connolly, J., Fiantis, D., Kidd, D., Ma, Y., et al.: PEATGRIDS: Mapping thickness and carbon stock of global peatlands via digital soil mapping, Earth System Science Data Discussions, 2024, 1–29, https://doi.org/10.5194/essd-2024-333, 2024.

Wieder, W. R., Pierson, D., Earl, S., Lajtha, K., Baer, S. G., Ballantyne, F., Berhe, A. A., Billings, S. A., Brigham, L. M., Chacon, S. S., Fraterrigo, J., Frey, S. D., Georgiou, K., De Graaff, M.-A., Grandy, A. S., Hartman, M. D., Hobbie, S. E., Johnson, C., Kaye, J., Kyker-Snowman, E., Litvak, M. E., Mack, M. C., Malhotra, A., Moore, J. A. M., Nadelhoffer, K., Rasmussen, C., Silver, W. L., Sulman, B. N., Walker, X., and Weintraub, S.: SoDaH: the SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0, Earth System Science Data, 13, 1843–1854, https://doi.org/10.5194/essd-13-1843-2021, 2021.

Wills, S., Seybold, C., Chiaretti, J., Sequeira, C., and West, L.: Quantifying tacit knowledge about soil organic carbon stocks using soil taxa and official soil series descriptions, Soil Science Society of America Journal, 77, 1711–1723, https://doi.org/10.2136/sssaj2012.0168, 2013.

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., and West, L. T.: Overview of the U.S. Rapid Carbon Assessment Project: Sampling Design, Initial Summary and Uncertainty Estimates, in: Soil Carbon. Progress in Soil Science, edited by Hartemink, A. and McSweeney, K., pp. 95–104, Springer International Publishing, https://doi.org/10.1007/978-3-319-04084-4_10, 2014.

Winkler, M., Plichta, R., Buysse, P., Lohila, A., Spicher, F., Boeckx, P., Wild, J., Feigenwinter, I., Olejnik, J., Risch, A., et al.: Global maps of soil temperature, Global Change Biology, 28, 3110–3144, https://doi.org/10.1111/gcb.16060, 2021.

https://doi.org/10.5194/essd-2025-336
Preprint. Discussion started: 24 June 2025

Xu, J., Morris, P. J., Liu, J., and Holden, J.: PEATMAP: Refining estimates of global peatland distribution based on a meta-analysis, CATENA, 160, 134–140, https://doi.org/10.1016/j.catena.2017.09.010, 2018.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset, Water Resources Research, 55, 5053–5073, https://doi.org/10.1029/2019WR024873, 2019.

Yigini, Y. and Panagos, P.: Assessment of soil organic carbon stocks under future climate and land cover changes in Europe, Science of The Total Environment, 557-558, 838–850, https://doi.org/10.1016/j.scitotenv.2016.03.085, 2016.

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., and Mi, J.: GLC_FCS30: global land-cover product with fine classification system at 30 m using time-series Landsat imagery, Earth System Science Data, 13, 2753–2776, https://doi.org/10.5194/essd-13-2753-2021, 2021.