



AI4SoilHealth

Soil Health Data Cube v1 D5.2

Version 1.0
30 June 2024

Lead Authors: R. Minarik (OpenGeoHub), X. Tian (OpenGeoHub), R. Simoes (OpenGeoHub), T. Hengl (OpenGeoHub), Serkan Isik (OpenGeoHub), L. Parente (OpenGeoHub),

Participating authors: D. Consoli (OpenGeoHub) & Y-F. Ho (OpenGeoHub)

Reviewed by: Lucas Gomes, AU; Mogens H. Greve, AU

Action Number: 101086179

Action Acronym: AI4SoilHealth

Action title: Accelerating collection and use of soil health information using AI technology to support the Soil Deal for Europe and the EU Soil Observatory



HISTORY OF CHANGES

Version	Publication date	Changes
1.0	30.06.2024	<ul style="list-style-type: none">Initial version





Soil Health Data Cube v1

Project ID: <https://cordis.europa.eu/project/id/101086179>

Summary: Soil Health Data Cube for pan-EU (SHDC4EU) V1 is now available (70% complete); majority of layers are available via our Zenodo community (<https://zenodo.org/communities/ai4soilhealth/>), STAC.EcoDataCube.eu and/or our project Github (<https://github.com/AI4SoilHealth/>). The SHDC4EU currently includes 90% of the base layers and 20% of the predicted soil layers (current focus is on soil properties: soil organic carbon, soil pH, texture fractions, bulk density, soil WRB types and similar). The remaining predictions are gradually being uploaded and added to the back-end and front-end. Due to the technical complexities connected with the import of dozens of point data sets coming from national legacy soil monitoring projects, we are continuously performing quality checks and this has resulted in some delays. In the coming months (until 1st of September) we plan to finish ALL remaining layers and submit / finish two (3) scientific publications explaining the general steps and accuracy assessment of soil type and soil property mapping. We also plan to publish the technical specifications for the AI4SoilHealth Soil Health data cube via the Github manual shown below. The final 100% V1 will be then presented during the M18 reporting period.

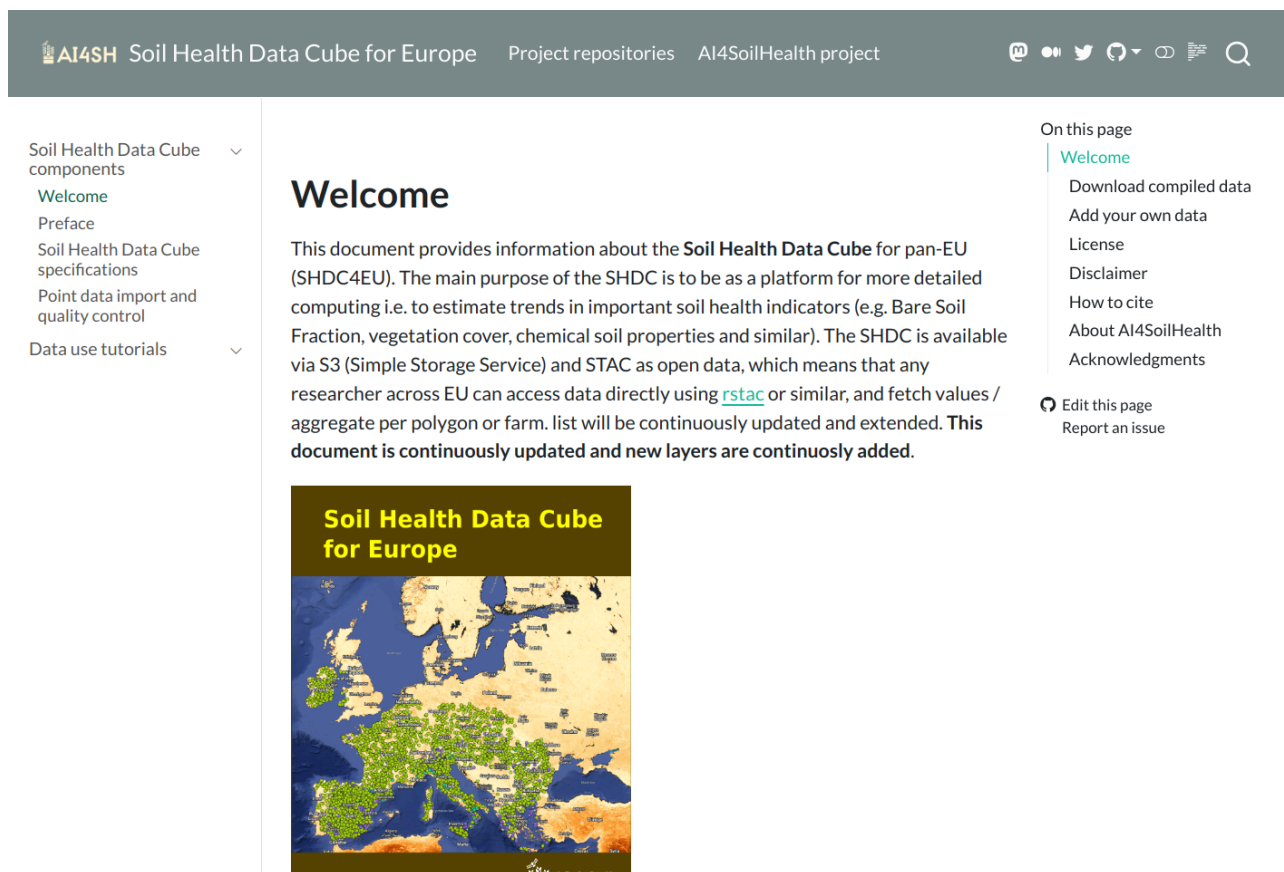


Figure 1: The Soil Health Data Cube is documented via <https://shdc.ai4soilhealth.eu> which will be used as the central place for all documentation including data access and use tutorials.

Three (3) publications are in preparation that describe production steps and results of cross-validation include:



- Tian, X. et al. **“Time-series of Landsat-based bi-monthly and annual spectral indices for continental Europe for 2000--2022”**, ESSD submitted; preprint: <https://doi.org/10.21203/rs.3.rs-4251113/v1>
- Tian, X. et al. **“Soil carbon mapping at 30-m for 2000–2022 using spatiotemporal Machine Learning”**, IN PROCESS, to be submitted to PeerJ,
- Minarik, R. et al. **“Soil type map of Europe at 30-m spatial resolution based on WRB classification”**, IN PROCESS, to be submitted to PeerJ

In principle, all steps and properties of produced components in the Soil Health Data Cube will be documented in the peer-review publications. The process of peer-review, however, can sometimes be long and this could potentially delay releasing all internal data. To avoid unnecessary delays, we plan to publish and register all preprints as soon as they are submitted, then mention clearly to users that the data is under construction and should be only used for testing purposes (until the publications pass peer-review and no more revisions are required).

1. Introduction

Soil Health Data Cube for pan-EU (SHDC4EU) has been previously described in the D5.1: Soil Health Data Cube technical specifications (OGH, M9). Task: 1.1, 2.2, 5.1, 6.4. The main purpose of the SHDC is to serve as a platform for more detailed computing i.e. to estimate trends in important soil health indicators (e.g. Bare Soil Fraction, vegetation cover, chemical soil properties and similar). The SHDC is available via S3 (Simple Storage Service) and STAC as open data, which means that any researcher across EU can access data directly using `rstac`¹ or similar seamless software, and fetch values / aggregate per polygon or farm. SHDC will be fully documented via <https://shdc.aii4soilhealth.eu> and will be continuously updated with new layers, new examples and worked out computational notebooks.

In the recent decade, there has been increasing interest in mapping spatial distribution of dynamic soil properties such as soil organic carbon (SOC) and at finer resolutions (see especially the National Academies of Sciences, Engineering, and Medicine report: “Exploring a Dynamic Soil Information System: Proceedings of a Workshop”). Some most recent global maps of soil organic carbon SOC at 1 km and 250 m are provided by FAO and ITPS (2018) and de Sousa et al. (2020). At continental level, de Brogniez et al. (2015) and Yigini and Panagos (2016) produced detailed SOC maps for Europe. SOC has been mapped at even finer spatial resolutions in countries such as Denmark (Adhikari et al., 2014), France (Mulder et al., 2016; Chen et al., 2018), Hungary (Szatmari et al., 2019) and Switzerland (Stumpf et al., 2018), just to mention a few most advanced national initiatives. Although stitching high resolution national carbon maps (the FAO approach) is technically possible, it has proven to lead to significant differences at borders and values in general can differ just because different laboratory standards and different sampling designs are used. For example, a country that makes all their soil laboratory data on sampling and monitoring agricultural land, might significantly underestimate national SOC stock as it would completely miss various pools of SOC in wetlands and forests. In order to produce unbiased pan-EU estimates of SOC changes at highest possible spatial resolution one needs global unbiased predictive mapping models which can account for large spatial

¹ <https://cran.r-project.org/web/packages/rstac/>



clusterings of training points. Soil carbon information is not critically missing in the projects such as AI4SoilHealth where most of modelling is done at 30-m spatial resolution and assuming dynamic changes of land use, vegetation, climate etc. Ugbaje et al., (2024) developed spacetime predictions of SOC stocks for Australia at a 90 m spatial resolution covering 1990 and 2018. With the Soil Health Data Cube we are doing a somewhat similar thing — mapping soil properties through space time so we can also analyse trends in values 2000 to 2025.

2. SHDC4EU current base layers

2.1 Pan-EU Landmask

[Three Pan-EU land masks](#) designed for different specific applications in the production of soil health data cube:

- Land mask: with values differentiating land, ocean, and inland water
- NUT-3 code map: with values differentiating administrative area at nut-3 level
- ISO-3166 country code map: with values differentiating countries according to ISO-3166 standard

The jupyter notebooks and bash files that are used to produce masks, merge tiles, reproject crs, resample to another resolution are openly available at [GitHub Repository: SoilHealthDataCube](#).

All the land-masks are aligned with the standard spatial/temporal resolution and sizes indicated/recommended by the AI4SoilHealth project, Work Package - 5. The coverage of these maps closely match the data coverage of <https://land.copernicus.eu/pan-european> i.e. the official selection of countries listed here: https://lanEEA39d.copernicus.eu/portal_vocabularies/geotags/eea39.

These masks are created by [Xuemeng](#), [Yu-Feng](#), and [Martijn](#) from [OpenGeoHub](#).

2.2 Landsat-based Spectral Indices Data Cube

This data cube offers a time-series of Landsat-based spectral indices maps across continental Europe—including Ukraine, the UK, and Turkey—from 2000 to 2022. To ensure high-quality data for modelling, a comprehensive preprocessing workflow is employed, including temporal aggregation, gap-filling using the Seasonally Weighted Average Generalization (SWAG) method, and land masking. The processed data is organised into four tiers of predictors: bimonthly Landsat bands, spectral indices, annual aggregated indices, and long-term temporal feature indices. These predictors encompass various thematic groups, covering:

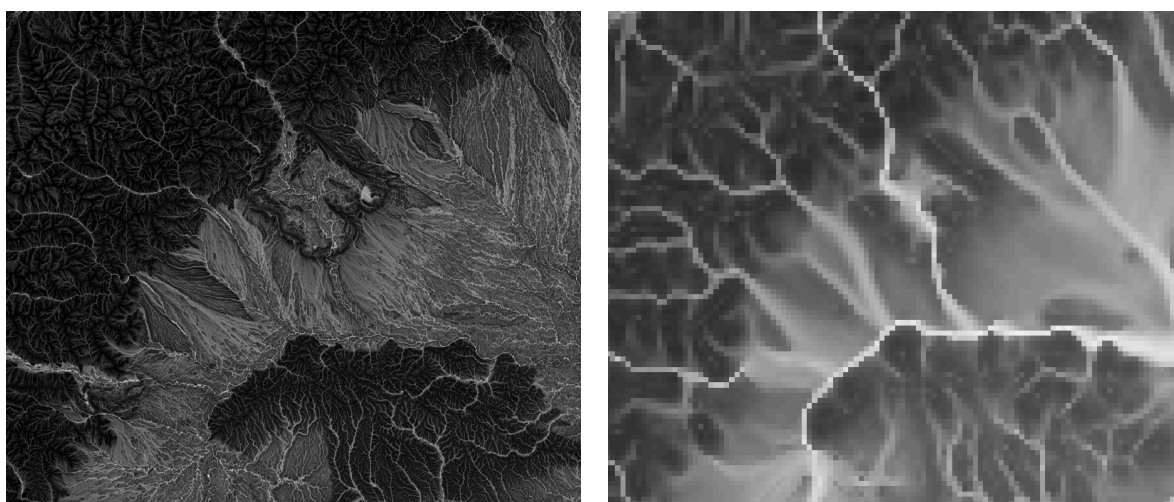
- Vegetation index: Normalized Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR).
- Soil exposure: Bare Soil Fraction (BSF).
- Tillage and soil sealing: Normalized Difference Tillage Index (NDTI) and minimum Normalized Difference Tillage Index (minNDTI).
- Crop patterns: Number of Seasons (NOS) and Crop Duration Ratio (CDR).
- Water dynamics: Normalized Difference Snow Index (NDSI) and Normalized Difference Water Index (NDWI)

All the indices are aligned with the standard spatial/temporal resolution and sizes indicated/recommended by AI4SoilHealth project, Work Package - 5. The code used for index calculation, map production, visualisation, plausibility check, and model experiment of the Landsat-based spectral indices data cube is openly available in [Github Repository: SoilHealthDataCube](https://doi.org/10.21203/rs.3.rs-4251113/v1). The paper is submitted to ESSD (<https://doi.org/10.21203/rs.3.rs-4251113/v1>).

2.3 Digital terrain model of Europe

We prepared the ensemble digital terrain model (DTM) based on the global version (<https://doi.org/10.5281/zenodo.7634679>). We resampled the baseline 30 m DTM into the [octaves](#) (a set of hierarchical layers doubling the spatial resolution to reveal more complex terrain features) in resolutions of 30, 60, 120, 240, 480 and 960 m. The DTMs were hydrologically corrected and the set of 16 features was calculated (Table 1). The effect and the importance of the fingers spatial resolution for digital soil mapping is illustrated in Figure 2.

Because of the big data size, the layers are still uploaded to Zenodo (<https://doi.org/10.5281/zenodo.12608805>). It will be finished in July, 2024.



Topographic wetness index in 60 m. Topographic wetness index in 960 m.

Figure 2. The effect of “scale” i.e. spatial resolutions using the example of TWI in Po valley, northern Italy.

Table 1: The list of morphological and hydrological features included in the SHDC4EU.

Number	Name	URL
1	Slope in percent	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_morphometry_0.html
2	Minimum curvature	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_morphometry_0.html
3	Maximum curvature	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_morphometry_0.html
4	Hillshade	https://gdal.org/programs/gdaldem.html
5	Valley depth	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_channels_7.html



6	Negative openness	https://saga-gis.sourceforge.io/saga_tool_doc/2.2.4/ta_lighting_5.html
7	Positive openness	https://saga-gis.sourceforge.io/saga_tool_doc/2.2.4/ta_lighting_5.html
8	Northernness	https://grass.osgeo.org/grass82/manuals/addons/r.northernness.easterness.html
9	Easternness	https://grass.osgeo.org/grass82/manuals/addons/r.northernness.easterness.html
10	Sink removal	https://saga-gis.sourceforge.io/saga_tool_doc/2.1.4/ta_preprocessor_2.html
11	Flow accumulation (Tot catch area)	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_hydrology_0.html
12	Catchment area	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_hydrology_19.html
13	TWI	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_hydrology_20.html
14	LS factor	https://saga-gis.sourceforge.io/saga_tool_doc/7.3.0/ta_hydrology_22.html
16	Terrain classification (12)	https://saga-gis.sourceforge.io/saga_tool_doc/9.3.1/ta_morphometry_25.html
17	Geomorphon classes (10)	https://grass.osgeo.org/grass83/manuals/r.geomorphon.html

2.4 Lithology map of Europe

The lithology map for Europe is generated from a 1:1 million scale soil parent material map produced by the European Geological Data Infrastructure (EGDI). The original data from EGDI contains missing lithology information in some regions such as Switzerland, the Balkan regions, and Baltic countries. The gaps in the lithology map, including Ukraine and Turkey, were filled by training a random forest classifier model based on parameters derived from DTM (given in Section 2.3) and soil regions map from Die Bundesanstalt für Geowissenschaften und Rohstoffe (BGR).

By generating 1 million random points, geographically balanced over the whole pan-EU land area, each class in the map was covered properly. Classes whose number of samples is less than 10 were discarded from the model training. The hyperparameter tuning of the model was carried out via a Bayesian approach with a criteria to maximize accuracy of 5k-fold cross validation.. The tuned random forest model achieved an accuracy of 47% (Kappa=0.43) for the testing data, 20% of the generated sample points.

The lithology of Turkey, on the other hand, was digitised from the available geology map produced by the General Directorate of Mineral Research and Exploration (MTA). The available raster map was post-processed and classified as 20 lithology classes using the k-means algorithm. These classes were harmonized with the classes in the EGDI lithology map.

The updated lithology map of Europe is uploaded to the [Zenodo community](#).

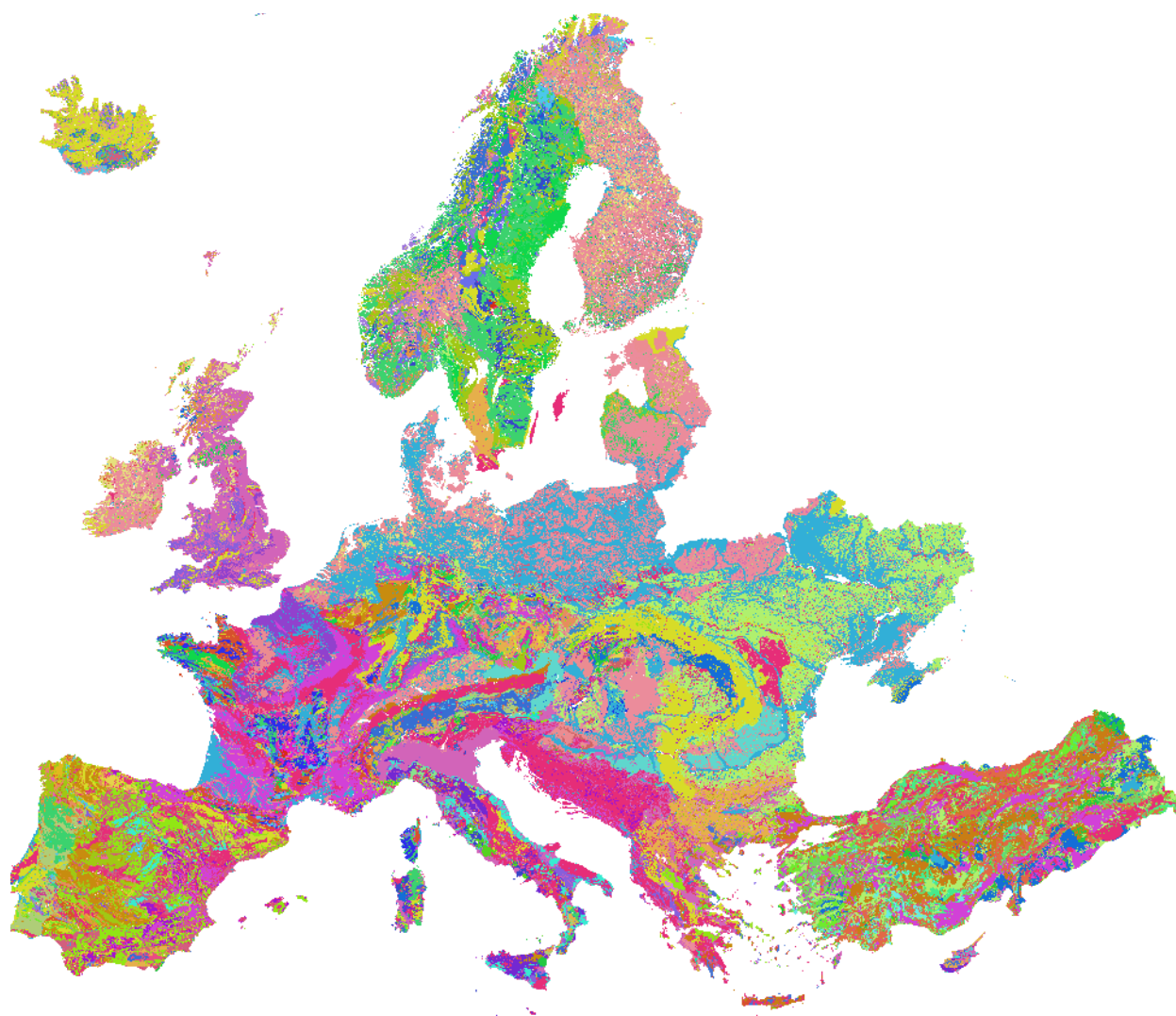


Figure 3. The updated lithology map of Europe now matches the AI4SoilHealth land-mask.

3. SHDC4EU current predicted soil layers

Table 2: List of target soil variables in the V1 of the SHDC4EU. Dynamic soil properties cover the period 2000–2025+ and can be used to update to more recent years.

Title	Unit	Standard code	2D/3D	2D+T	3D+T	Depth intervals	Temporal support	Status
Soil organic carbon density	kg/m3	ISO 10694. 1995		✓	✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process



Soil carbon stock change	t/ha/yr	ISO 10694. 1996		✓	✓	0-20, 20-50, 50-100	long-term	Models fitted / production in process
Soil pH in a suspension of soil in water	[-]	ISO 10390. 1994		✓	✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Soil pH measured in a CaCl ₂ solution	[-]	ISO 10390. 1994			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Soil total nitrogen content	dg/kg	ISO 11261:1995			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Soil bulk density	t/m ³	Adapted ISO 11272:2017			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Soil texture fractions (sand, silt, clay)	g/g	ISO:11277			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Soil WRB subgroup	factor	WRB2022	✓			0-200	long-term	Models fitted / production in process
Depth to bedrock	cm		✓			0-200	5-year	Models fitted / production in process
Extractable potassium content	mg/kg	USDA-NRCS, 2004			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Carbonates content CaCO ₃	g/g	ISO 10693:1995			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Extractable Phosphorus content (Olsen)	mg/kg	ISO 11263. 1994			✓	0-20, 20-50, 50-100	5-year	Models fitted / production in process
Monthly Gross Primary Productivity	kg/ha/yr			✓		NA	bi-monthly	Production completed
Fraction of bare soil	m ² /m ²			✓		NA	annual	Production completed



3.1 Soil type predictions WRB for pan-EU

3.1.1. Harmonisation of the soil profiles

In order to collate the most comprehensive pan-EU training dataset, we collated all available international, national and regional datasets from project partners (marked partner), national institutions (marked institut) and publicly available datasets (Table 3). OpenGeoHub has signed a data protection agreement in the case of some datasets, therefore the point data are not sharable.

Table 3: The overview of training data

Country	Redistribution	Original Soil Classification	Num of points	Basic correlations to WRB2022
Denmark (partner)	Restricted	FAO1974	1214	WISE30Sec , HWSD 2.0
Germany BZE (partner)	Restricted	KA5	2861	https://doi.org/10.4324/9781849774352 ; KA5
Belgium (Vlaanderen)	Restricted	Belgic system	933	10.13140/2.1.4381.4089
Netherlands	Yes	Dutch system	2061 (200,000)	https://doi.org/10.4324/9781849774352
Slovenia (institut)	Restricted	Slovenian system	1800	https://doi.org/10.1007/978-94-017-8585-3
Croatia (Partner)	Restricted	Croatian system	2055	https://doi.org/10.1007/978-94-007-5815-5
Portugal	Yes	WRB various	2889	WRB2022
Italy (partner)	Restricted	WRB2015	2576	WRB2022
GeoCradle	Yes	WRB various	300	WRB2022
Hungary (partner)	Restricted	WRB2015	65	WRB2022
Estonia (institut)	Restricted	WRB2015	123	WRB2022
SOTER	Restricted	FAO1990	662	HWSD 2.0
WOSIS	Yes	WRB various/FAO	1205	HWSD 2.0
European Soil DB v2 raster	Restricted	WRB various	1000	WRB2022

Three types of the soil profiles were presented in the training dataset. The [conversion tables](#) are publicly available to comment and adjust.

- The soil profiles in older WRB classifications.** These points were converted to the WRB 2022 with no semantic changes. The principal qualifiers were moved to the supplementary qualifiers and vice versa. Because the majority of the points (85 %) contained only one or more principal qualifiers, the supplementary qualifiers were excluded from the final legend. The only exception was the Haplic principal qualifier for which the most important qualifier was preserved. The points were assigned to quality codes 6 or 5 (Table 3).
- The soil profiles in national or FAO classifications.** These points were correlated to WRB using already published conversion tables with forwarding corrections to the WRB 2022. The references are provided in table 1. The points were assigned to quality codes 4 or 3.

3. **The soil profiles have only WRB reference soil group (RSG) assignments.** These points (< 5% of total points) were assigned with the two most frequent principal qualifiers for the specific RSG. Therefore 2 alternatives for the points were used in the training step. The points were assigned to quality codes 2 or 1.

Table 4: WRB weights reflecting the quality of the training soil profiles

Weight	Code	Quality class
1	6	Original or simplified original WRB classification used (no semantic changes)
0.9	5	older WRB classification converted to WRB2022 (no semantic changes), i. e. principal qualifiers moved to supplementary qualifiers
0.8	4	national classification converted to WRB2022 principal qualifiers and RSG using already published national conversion tables
0.6	3	national or FAO classification converted to WRB2022 principal qualifiers and RSG using already published generic conversion tables (Krasilnikov et. al., HSWD, ISRIC reports)
0.4	2	national classification to WRB2022 using already published conversion tables for RSG, most frequent primary classifier added based on the other national resources
0.1	1	national classification to WRB2022 using already published conversion for RSG, most frequent primary classifier of WOSIS/HWSD_2.0 for the Europe

The special cases were WOSIS database, [European Soil DB v2 raster](#) (ESDB) and Dutch dataset. We used only WOSIS points having WRB or FAO full classification. Moreover, some regions of Europe (Spain, France and Scandinavia) were underrepresented in the training dataset. Therefore we randomly sampled 1000 points from WRBFL (WRB Full Legend). We have also selected a representative sample (2061) from the original 200 000 Dutch profiles [using doubly balanced sampling](#) creating a stratified sample in the geographical and thematic domain. The distribution of the pont over Europe is in Figure 4.

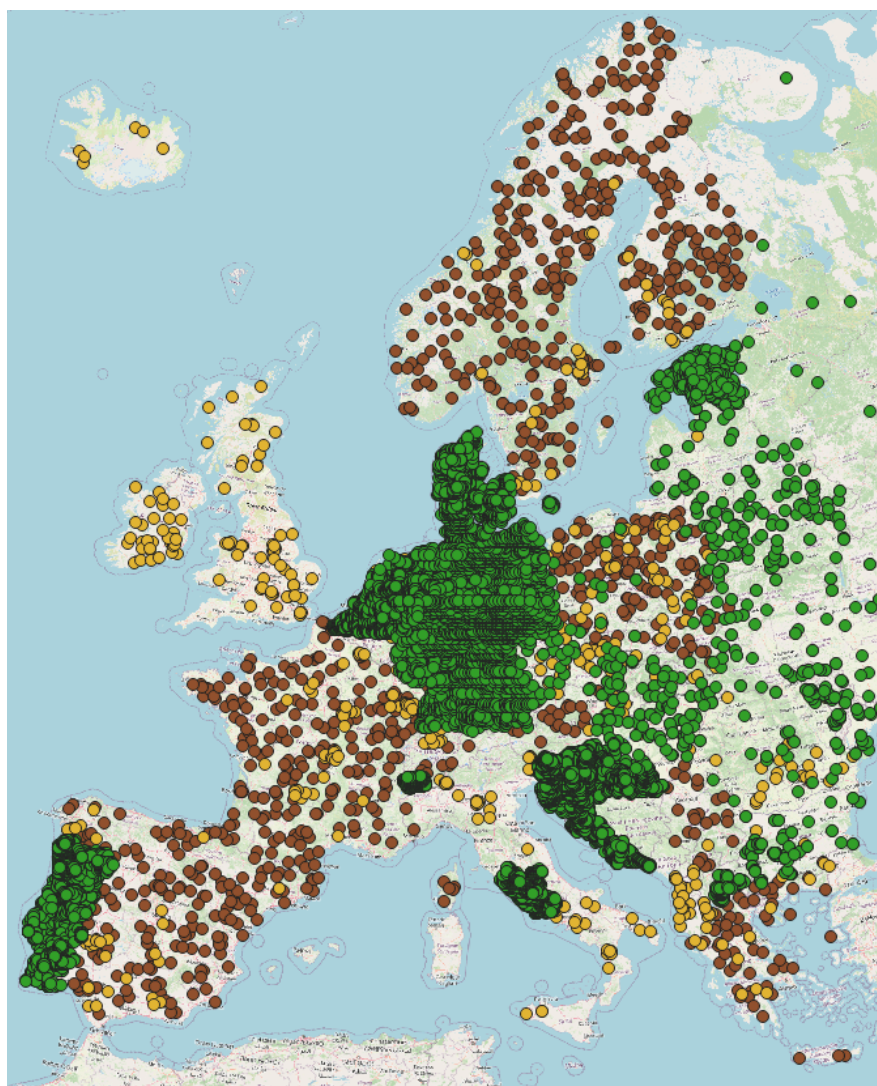


Figure 4. Spatial distribution of all available soil types point data. Highest quality national data are highlighted in green. WOSIS points are highlighted in orange. Synthetic ESDB points extracted from the map are highlighted in brown.

The data harmonisation resulted in 19 745 unique points. The total number of training points was 22 155, because for points with known RSG only, two most frequent principal qualifiers were added based on the national records or the WOSIS database. The harmonisation resulted in 363 soil type classes consisting of the main principal qualifier and RSG. For modelling, we used soil type classes having more than 10 points. It resulted in 21 583 training points and 156 classes.

3.1.2. Modelling pipeline

We adapted the automated modelling pipeline described in sections 3.2.3 and 3.2.4 for the classification task with imbalanced categories using a Random Forest (RF) model in python. We trained the long term 2D model predicting probabilities of the soil types to incorporate the natural fuzzy aspect of the WRB classification and the soil classifications in general. Moreover we used a weighted modelling approach when

the weight from the Table 4 representing the quality was assigned to each point. Feature selection and hyperparameter tuning was performed on the 20% of the training data selected by stratified random sampling using predefined blocks to avoid local spatial correlation. The model hyperparameter tuning was done by using RandomSearch featuring spatial block CV, sample and weights from Table 4. The final model was trained on all samples and the predictive power of the model was compared to the prior baseline classifier from scikit-learn that returned probabilities equal to the empirical class prior distribution based on the spatially blocked 5-fold cross validation. The predictive power was compared using the log loss score, log_loss ratio and weighted F1 (details in [10.7717/peerj.13573](https://doi.org/10.7717/peerj.13573)).

The result showed a general improvement of 16–20 % of the log_loss for the RF model predicting principal qualifiers + RSG and RSG compared to the baseline model (Table 5).

Table 5: The global accuracy assessment

Model	WRB Level	N of classes	Precision	Recall	Weighted F1	Log Loss	Log loss ratio
RF	Principal Q.	156	0.15	0.20	0.15	3.61	0.2
Baseline	Principal Q.	156	0.15	0.20	0.15	4.5	NA
RF	RSG	26	0.33	0.32	0.32	2.2	0.16
Baseline	RSG	26	0.05	0.23	0.08	2.63	NA

The probabilities of belonging to the classes were predicted for the pixels. The probabilities were also used to inform about the per pixel using 90 % confidence interval. Figure 5 shows the detailed example of the classification in the Alps where the Swiss pilot site is. The visual accuracy assessment is in line with the accuracy report of the classification (Table 6). For the reporting purposes we aggregated the probabilities of the predicted classes to the level of RSG. The soil types presented in the map are highlighted in yellow.

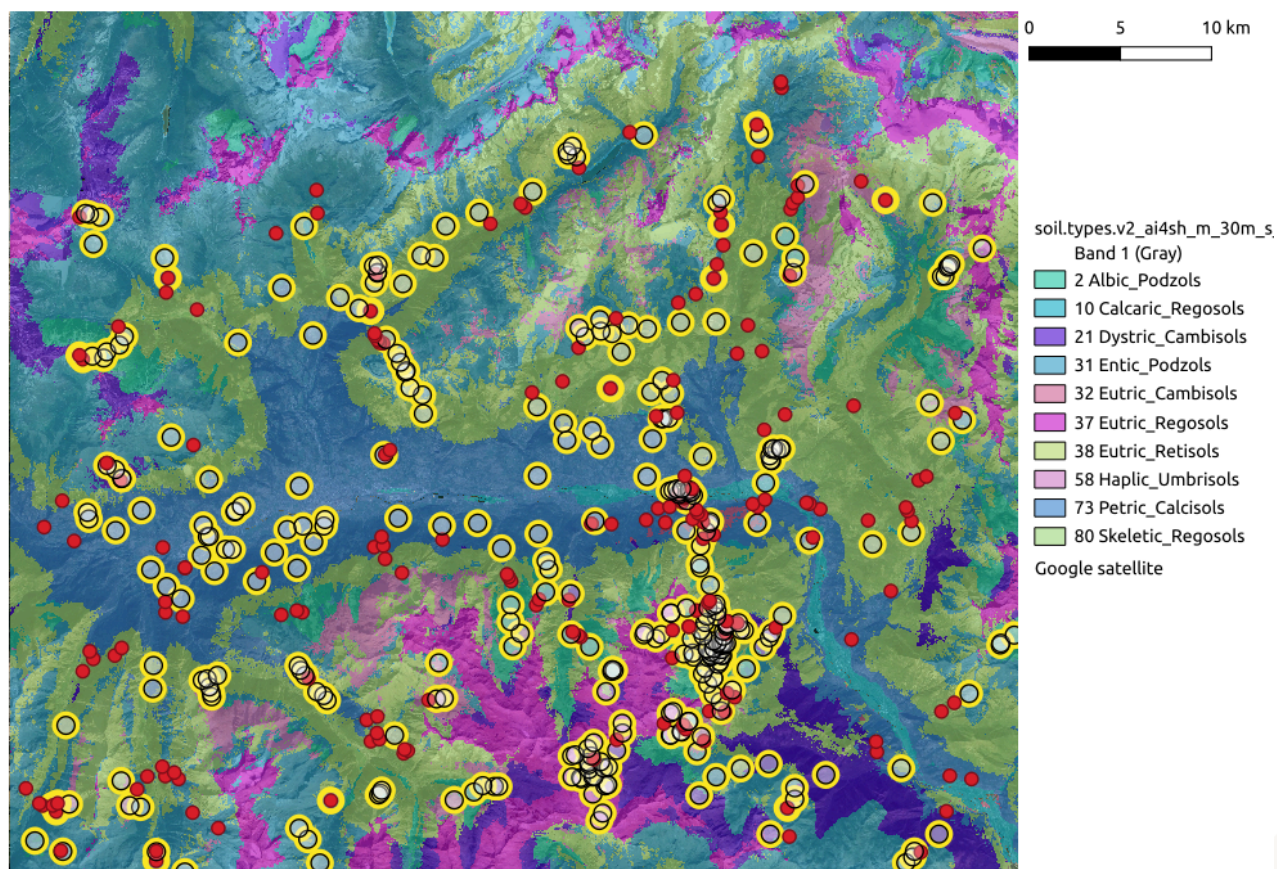


Figure 5. The detailed classification of the Swiss pilot site region in the Alps on the background of the most probable soil type. The black points are accurately classified points The red points are misclassified in the level of the principal qualifiers. The yellow halo marks the accurately classified points on the level of RSG.

Table 6: The aggregated probabilities of the predicted classes to the level of RSG.

	precision	recall	f1-score RF	f1-score Baseline	support
Acrisols	0.12	0.09	0.1	0	142
Alisols	0.05	0.06	0.05	0	31
Andosols	0	0	0	0	53
Anthrosols	0.26	0.32	0.28	0	1121
Arenosols	0.21	0.13	0.16	0	464
Calcisols	0.13	0.25	0.17	0	251
Cambisols	0.52	0.35	0.42	0.38	4996
Chernozems	0.3	0.69	0.42	0	281
Ferrasols	0	0	0	0	25



Fluvisols	0.3	0.29	0.3	0	936
Gleysols	0.32	0.25	0.28	0	1690
Histosols	0.22	0.31	0.26	0	678
Kastanozems	0.06	0.26	0.1	0	53
Leptosols	0.26	0.33	0.29	0	1539
Luvisols	0.35	0.27	0.3	0	2442
Phaeozem	0	0	0	0	33
Phaeozems	0.26	0.3	0.28	0	1454
Planosols	0.25	0.14	0.18	0	276
Podzols	0.43	0.58	0.49	0	2430
Regosols	0.22	0.15	0.18	0	1577
Retisols	0.16	0.24	0.19	0	241
Solonchaks	0	0	0	0	12
Solonetz	0	0	0	0	27
Stagnosols	0.18	0.28	0.22	0	759
Umbrisols	0.03	0.02	0.02	0	348
Vertisols	0.12	0.2	0.15	0	270
accuracy	0.32	0.32	0.32	0.22	
macro avg	0.18	0.21	0.19	0.01	22129
weighted avg	0.33	0.32	0.32	0.08	22129

Despite the significant delay of 2 months caused by negotiating the conditions of the data sharing agreement and the time consuming harmonisation, the resulting delay is minimal compared to the plan (the map published on June 30). The training dataset was consolidated, the final model was trained and validated on the selected pilot sites with acceptable level of accuracy. Now the prediction of probabilities is running over Europe and the map will be published in July. For now, the users can use the proof of concept version containing 1km maps published in the Zenodo community (<https://doi.org/10.5281/zenodo.7820797>).

3.2 Soil property predictions for pan-EU

3.2.1 Soil property training data preparation

The first edition of the Soil Health Data Cube integrates soil point measurements from 22 sources across 37 countries, covering 11 key soil properties: soil organic carbon (SOC), nitrogen (N), carbonate (CaCO₃), soil texture (silt, sand, and clay), cation exchange capacity (CEC), electrical conductivity (EC), pH in water, pH in CaCl₂ solution, bulk density, extractable phosphorus (P), and extractable potassium (K). The spatial distribution of the harmonized soil data is shown in Fig. 6.

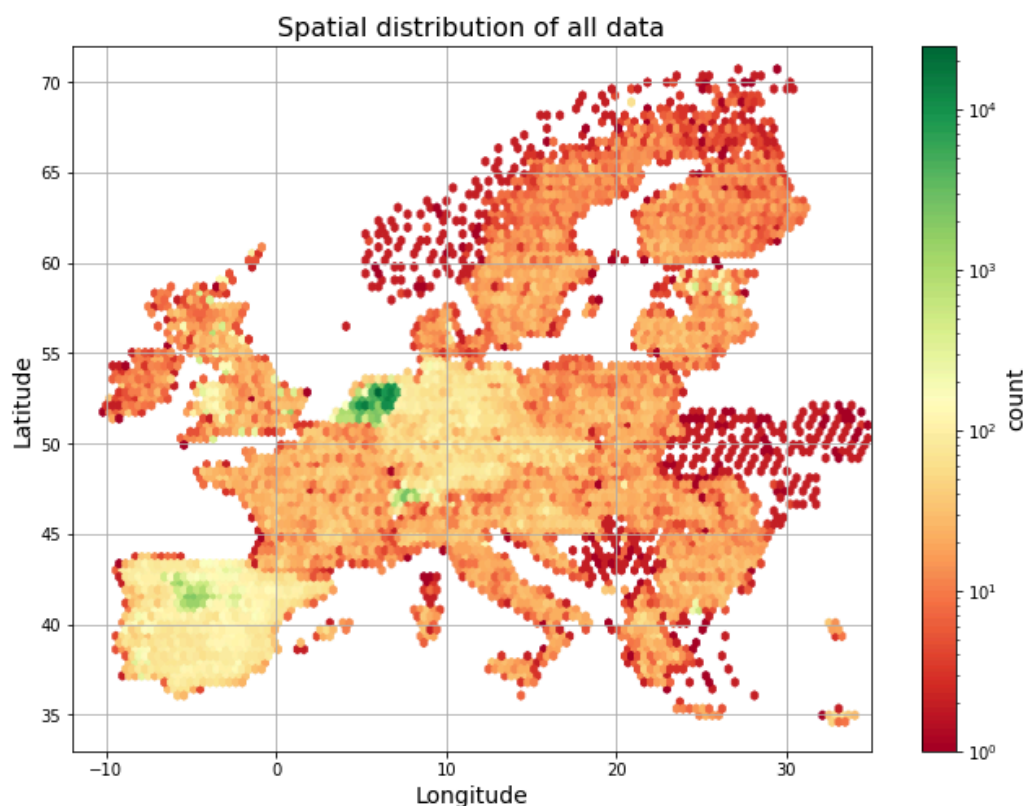


Figure 6. Spatial distribution of all available soil point data.

These data are standardized into a unified dataset that (1) includes comprehensive metadata such as sampling year, depth, and location; (2) essential soil property data for at least one of the above-mentioned 11 properties; (3) and a quality score for each measurement to indicate reliability. The dataset preparation follows a structured, four-stage process, using the LUCAS soil survey as a benchmark to ensure data consistency and accuracy:

1. **Data Collection:** We begin by gathering soil point data from a range of sources, including consortium partnerships, public databases, and various networks. This stage aims to compile a diverse array of data to be representative of the European soil as much as possible.
2. **Data Assembly:** Following collection, we apply an initial filter to keep only soil properties that are relevant to our study (i.e. among the above-mentioned 11 properties). We record each data record's measurement method, the units used, and its source. The data from each source is then organised into a standardised format, and then integrated together.
3. **Data Cleaning:** The assembled data undergoes a thorough cleaning process to ensure that all entries are complete and precise, especially regarding coordinates, depth, and sampling year. Only data samples collected after the year 2000 are kept to maintain consistency with the study's temporal scope.
4. **Property Harmonization:** In this phase, we harmonise 13 different soil properties by standardising the measurements to align with those of the LUCAS reference dataset. This standardisation ensures consistency in measurement methods and units. For units that are unknown, we make assumptions

based on their statistical distribution. Each data point is also assigned a quality flag to reflect the reliability of the corresponding measurement (see Fig. 5).

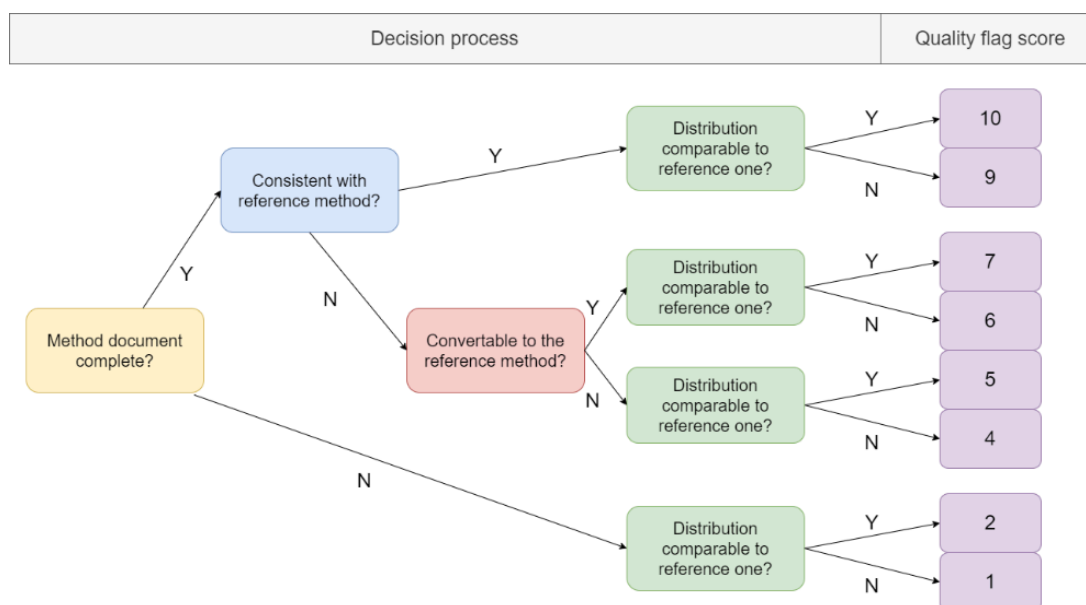


Figure 7. The decision process to assign a quality score for each soil property measurement.

The details could be found in [AI4SH soil data harmonization specifications](#).

3.2.2 Predictor layers preparation

The features we used for model fitting and map production contain around 450 covariate layers.. These layers have been prepared to comply with the technical specifications outlined in D5.1: Soil Health Data Cube, ensuring they are well-suited for integration, cross-comparison, and subsequent map production. The covariate layers include a diverse range of geospatial layers that detail various environmental conditions. Key categories include climate, landsat-based spectral indices, parental material, water cycle, terrain, and human pressure factors. Details could be found in [AI4SH soil health data cube covariates preparation](#).

3.2.3 Pipeline to test and select best predictive model

A standardized pipeline has been developed to automate model development for predicting soil properties. This pipeline enhances model performance through hyper-parameter tuning, feature selection, and cross-validation. The process begins with the input of harmonized soil data, the list of covariate paths, and a defined quality score threshold to ensure data reliability for each property. The pipeline processes inputs to produce calibration and training datasets, trained models, sorted feature importance, performance metrics, and accuracy plots (an example is shown in Fig. 8). This organized output ensures reproducibility and transparency, facilitating detailed model examination. The goal is to identify the best model for each soil property from five candidates: Artificial Neural Network (ANN), Random Forest (RF), LightGBM, and their weighted variants, which use normalized square quality scores as weights. ANN is excluded from weighted regression due to scikit-learn's limitations.

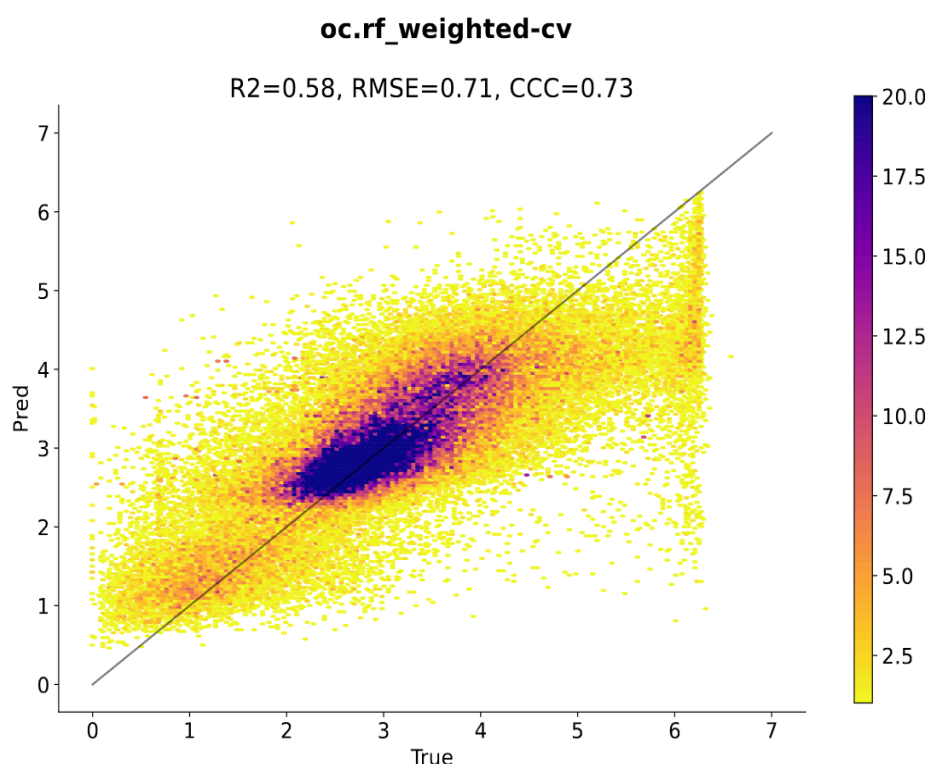


Figure 8. An example accuracy plot of SOC modeling using a weighted Random Forest model. The results shown in the plot were generated from a 5-fold spatially blocked cross-validation.

Firstly, approximately 5000 soil points were selected from LUCAS through stratified random sampling to serve as an independent validation dataset, enabling a robust assessment of the model's generalization capability. The rest of soil data is kept for model training and cross-validation. To maintain representativeness and robustness in the training dataset, approximately 20% of the rest soil property data points are randomly selected in a stratified manner from each spatial block (approximately 120 km grids). This selection ensures a broad geographic representation and enhances the reliability of the models. The selected data points are then utilized for model calibration, which includes both feature selection and hyper-parameter tuning. Hyper-parameter fine tuning is done using `HalvingGridSearch` from *scikit-learn* library. Feature selection is performed using a default Random Forest (RF) model from *scikit-learn*, where only features with above-average importance are consistently voted important across 20 bootstrap runs to be included in the final model. A spatial blocking strategy is used to select 70% of training data for each run to ensure geographically proximate soil points are not selected together, promoting robust feature generalization. The refined feature set and hyperparameters obtained from the calibration phase are further employed to perform a robust validation through a 5-fold spatially blocked cross-validation (CV) method on the remaining 80% of the soil point data. During the cross-validation process, key performance metrics including the coefficient of determination (R^2), Root Mean Square Error (RMSE), Concordance

Correlation Coefficient (CCC), and computation time are recorded for each fold and then averaged.

property	final_model	R2_val	RMSE_val	CCC_val	R2_cv	RMSE_cv	CCC_cv	cv_time (s)	test_time (s)
oc	rf	0.521929	0.631190	0.686796	0.584845	0.709827	0.736924	114.771056	136.970790
ph_h2o	rf	0.733580	0.699773	0.840577	0.603598	0.795962	0.747076	81.725859	58.499077
ph_cacl2	rf_weighted	0.759816	0.695820	0.858062	0.598751	0.838302	0.743977	137.373652	134.724805
bulk_density	rf_weighted	0.279815	0.417525	0.424275	0.356616	0.333586	0.540507	6.947836	7.106372
caco3	lgb_weighted	0.659061	1.287274	0.787558	0.529153	1.427255	0.686390	8.254335	0.615786
CEC	rf_weighted	0.406325	0.544544	0.541953	0.379794	0.594566	0.520019	87.544101	102.824505
EC	rf	0.365355	0.592491	0.512382	0.464276	0.622446	0.621897	269.255990	292.663614
P	rf_weighted	0.289630	0.962262	0.453291	0.301582	0.929909	0.447963	318.177178	372.809715
K	rf	0.424477	0.680985	0.573106	0.369374	0.708407	0.526689	650.003211	753.320741
N	rf_weighted	0.590117	0.420008	0.735299	0.729614	0.451402	0.839741	48.048637	56.602144

Figure 9: Model Performance of the Optimal Model Selected for Predicting Each Property.

The model demonstrating the best overall performance across these metrics is selected as the final model. The best model of each property and their corresponding metrics are shown in Fig. 9. This model is then trained on the complete dataset of soil points using the optimized features and parameters. The fully trained model will be used for soil property map production. Additionally, a quantile regression model with identical parameters and training data estimates prediction intervals to quantify uncertainty. These intervals, representing the probability that true soil values fall within a specified range, will also be produced as map layers to indicate uncertainty.

The data production pipeline is available in the [Github Repository: SoilHealthDataCube](#).

3.2.4 Pipeline to produce continental maps and uncertainties

The selected machine learning models are used to predict the soil for the whole pan-EU. All the features used during the training phase are available as Cloud Optimized GeoTIFF (COG) and stored in local servers to improve the I/O performance. The pan-EU area is divided in tiles to allow parallel computing across several computing nodes and to better fit the memory availability of each machine. The pan-EU area is divided into 693 tiles to allow parallel computing across several computing nodes and to better fit the memory availability of each machine. Each tile comprises 4096 x 4096 pixels with a resolution of 30 meters, covering an area of approximately 122 km by 122 km. For each tile all the features selected by the models are loaded in memory for the corresponding area for the selected year and solid depth. Iteratively, for each soil property, the required subset of features is reordered and adapted to be used as input for the corresponding model. This strategy allows to take advantage of the overlap of features used by different models reducing the overall reading time of input data. The produced tiled predictions are saved as GeoTIFF including the uncertainties represented by the 90 % quantile of the predictions. Predictions at depths of 0 cm, 20 cm, 50 cm, and 100 cm are used to calculate the average properties for the intervals 0-20 cm, 20-50 cm, and 50-100 cm, resulting in one averaged property for each depth range. These maps were computed

for the years 2000, 2005, 2010, 2015, and 2022. After all pan-EU tiles are produced, they are mosaiced, saved again as COG and stored in a local S3-based cloud storage, to allow visual inspection of the produced maps. The computation is performed using the internally developed open-source Python library [scikit-map](#), which uses a C++ backend for most computationally intensive operations. All the code used and developed from the pipeline is open-source and run using Docker containers to improve reproducibility.

Currently, all production ready models of soil properties are fitted and validated. The production is running now which is a delay compared to the plan to finish the production up to the end of June 2024, but the SHDC4EU will be completely finished until 1st of September. The delays were caused by delays in the data licence agreements and harmonization of dozens of different data sources to prepare the most comprehensive dataset of Europe.

3.3 Other soil health indicators for pan-EU

- Gross primary productivity (GPP): GPP estimation was based on Light Use Efficiency (LUE) modelling and considered the harmonized Landsat data cube to produce bi-monthly estimates at 30-m spatial resolution. The accuracy assessment of our GPP product used as reference FLUXNET 2015² and 2,139 data points derived from several flux towers over Europe covering nine (9) land cover types, indicating a R2 of 0.66 and 2.13 gC/m2 of RMSE (See Figure 10). We are currently working on a publication to document all processing and validation steps, and the output maps will be publicly available in the Soil Health Data Cube.
- Bare soil fractions and photosynthetic vegetation fractions: The fractions were computed in 500 m resolutions based on methodology published in Sun et al 2024 (<https://doi.org/10.5194/essd-16-1333-2024>) and Hill & Guerschmann (2022) (<https://doi.org/10.1016/j.agee.2021.107719>).

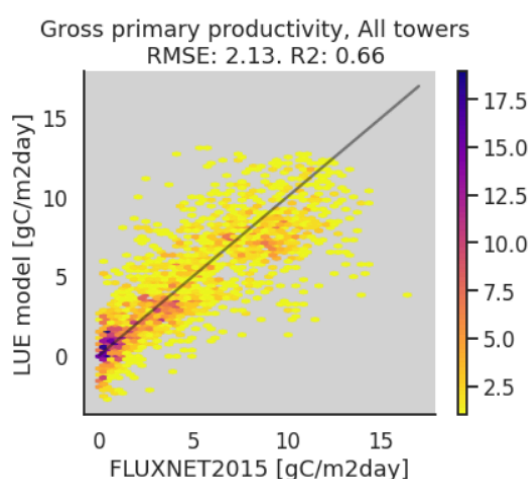


Figure 10: Accuracy assessment of GPP estimates for Europe based on 2,139 in-situ data points derived from FLUXNET 2015.

² <https://fluxnet.org/data/fluxnet2015-dataset>

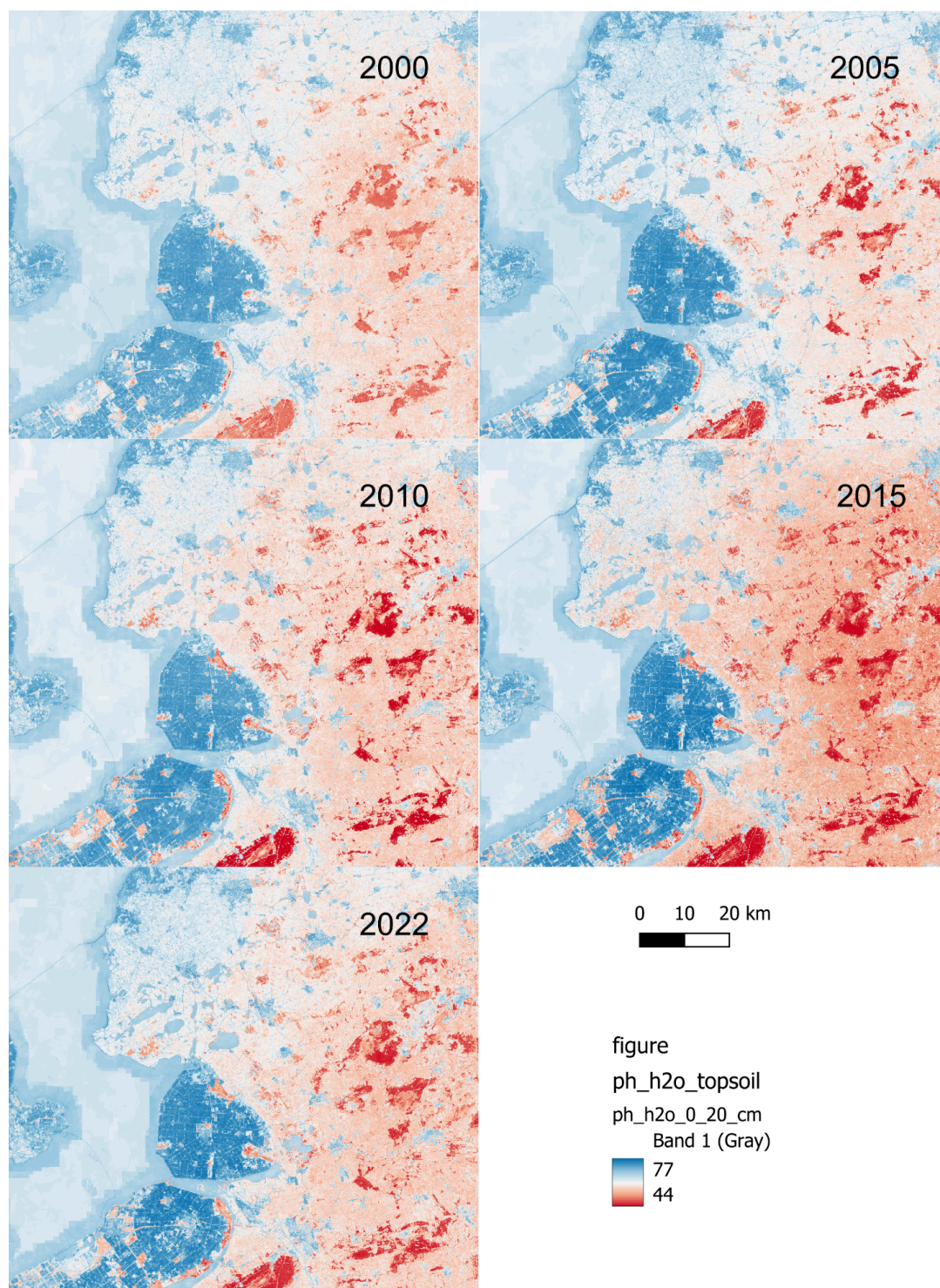


Figure 11. An example Topsoil (0-20 cm) pH in H₂O for central Netherlands. The values are multiplied by 10 to be stored as integers.



4. Summary points

In the coming months (until 1st of September) we plan to finish ALL remaining layers, finish the online guidebook and submit / finish two (2) scientific publications explaining the general steps and accuracy assessment of soil type and soil property mapping. We also plan to publish the technical specifications for the AI4SoilHealth Soil Health data cube via the Github manual. On a positive side we now have the majority of point data for modeling / model training, and this is most likely the most comprehensive collection of legacy soil data ever collected. Also our modeling is based on running spatiotemporal overlays, so that we exactly match environmental conditions at the year of sampling. This type of modelling is referred to as spatiotemporal machine learning and is the basis for producing unbiased predictions of key soil properties and hence also can be used to estimate the trends in properties.

Cited references:

1. Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B., & Greve, M. H. (2014). Digital mapping of soil organic carbon contents and stocks in Denmark. *PloS one*, 9(8), e105519.
2. Chen, S., Martin, M. P., Saby, N. P., Walter, C., Angers, D. A., and Arrouays, D. (2018). Fine resolution map of top-and subsoil carbon sequestration potential in france. *Science of the Total Environment*, 630:389–400.
3. de Sousa, L. M., Poggio, L., Batjes, N. H., Heuvelink, G., Kempen, B., Riberio, E., and Rossiter, D. (2020). Soilgrids 2.0: producing quality-assessed soil information for the globe. *SOIL Discussions*, pages 1–37.
4. FAO and ITPS (2018). Global Soil Organic Carbon map (GSOC) Map. Technical report. rome.
5. Mulder, V., Lacoste, M., Richer-de Forges, A., and Arrouays, D. (2016). GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. *Science of the Total Environment*, 3573:1352–1369.
6. Stumpf, F., Keller, A., Schmidt, K., Mayr, A., Gubler, A., and Schaepman, M. (2018). Spatio-temporal land use dynamics and soil organic carbon in swiss agroecosystems. *Agriculture, ecosystems & environment*, 258:129–142.
7. Szatmári, G., Pirkó, B., Koós, S., Laborczy, A., Bakacsi, Z., Szabó, J., and Pásztor, L. (2019). Spatio-temporal assessment of topsoil organic carbon stock change in hungary. *Soil and Tillage Research*, 195:104410.
8. Yigini, Y. and Panagos, P. (2016). Assessment of soil organic carbon stocks under future climate and land cover changes in europe. *Science of the Total Environment*, 557-558:838–850.
9. Ugbaje, S. U., Karunaratne, S., Bishop, T., Gregory, L., Searle, R., Coelli, K., & Farrell, M. (2024). Space-time mapping of soil organic carbon stock and its local drivers: Potential for use in carbon accounting. *Geoderma*, 441, 116771.