AI4SoilHealth

Pipeline for stream of in-situ & EO data with robust & FAIR principles

D4.15

Version 1.1

17th December 2025
Lead Author: Thomas Gumbricht, Stockholm University
Contributors: Fatemeh Hateffard, Hsiang-Ju Fan, Gustaf Hugelius, Stockholm University
Reviewed by: Giulio Genova, OpenGeoHub

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	10.12.2025	- draft version submitted to reviewer
1.1	15.12.2025	- First version submitted to coordinator

Pipeline for stream of in-situ & EO data with robust & FAIR principles

1. Introduction

This deliverable describes the work done in AI4SoilHealth (website) workpackage 4 (WP4) to create data pipelines for transfer of data into databases. The work includes data collection from field sampling, as well as data gathered via Earth Observation (EO). In the context of AI4SoilHealth and this deliverable, the field data primarily refers to pilot-site sampling of soils data for a set of different methods assessed within WP4, done at selected pilot-site locations, in close collaboration with WP6 (see https://ai4soilhealth.eu/pilot-sites/). The EO-data refers to data extracted for given point locations from the A4SoilHealth Soil Health Data Cube created and curated within WP5 (see https://shdc.ai4soilhealth.eu/).

The data pipelines consists of (1) structured data protocols and (2) scripts for transferring data into suitable database formats (json files). These scripts are organised as a Jupyter notebook front calling a set of Python packages for converting Comma Separated Values (CSV) files to structured json files. The databases into which soil data are fed are indented for further use within AI4SoilHealth, primarily to feed into the App, but also for direct accessibility for data users inside and outside the project. To accommodate these two different objectives of data used, the databases are in most cases available in two differently formatted json objects, a flat version more suitable for direct accessibility and a nested version intended for users who wants to convert the data to a relational Structured Query Language (SQL) database. The soil sampling methods and tools assessed in WP4, and for which this deliverable provides protocols and pipelines includes a defined set of in-situ soil health metrics. The deliverable is structured around those methods for which we have been able to develop joint structured sampling protocols and systematically collect data at multiple pilot sites. For in-situ field data, the following variables have been included (in alphabetical order): aggregates, bulk density, environmental DNA (eDNA) , extracellular enzymatic activity of soil microbes, infiltration capacity of soils, Ion Selective Electrode (ISE), Microbiometer data (commercial soil microbial kit), penetrometer data, salinity of soils, spectroscopic data and wet laboratory data for traditional chemicophysical soil properties. For the EO data, scripts have been developed to extract values for a selected sub-set of the multitude of variables described in the Soil Health Data Cube, this selection may be viewed as preliminary pending further development of the AI4SoilHealth App.

This Deliverable report may be seen as complementary to a dedicated AI4SoilHealth data pipeline GitHub page available at: **https://ai4sh.github.io/in-situ_data_docs/docs/**

All the documentation, scripts and data protocols described below are also available there, often with more detail at the webpage. The webpage is also a living portal which will continually be updated and develop further. Below we present the data-pipeline scripts for transferring field data into databases (section 2), brief method descriptions for each field data variable (section 3) and a description of the scripts to extract EO data for individual point locations form the Soil Health Data Cube (Section 4). More detailed information as well as the most recent versions of data-pipeline scripts can be retrieved form the website.

2. Scripts for organizing AI4SoilHealth in-situ data into databases

The script repository can be accessed via GitHub at this URL:

https://github.com/AI4SH/in-situ_data_management

This repository consists of a Jupyter notebook front calling a set of Python packages for converting Comma Separated Values (CSV) files to structured json files. The repository includes a detailed README.md file which explains how to use the Jupyter notebook, the Python packages are documented here. The Jupyter notebook AI4SH in-situ data management compiles in-situ data from soil sampling campaigns into organised json documents. The input data must be organised as CSV files, typically created by first putting the in-situ sampled data together in a spreadsheet and then exporting the spreadsheet as a .csv file.

2.1. Input data requirements and organisation

To run the notebook at least four files must be defined, i) one defining the user project settings (user project file), ii) one defining process parameters (process file), iii) one with the actual data (data source file), and iv) one that defines the columns (header) in the data file (method source file). The latter two are defined as process parameters in the process file, whereas the user project file and the process file are defined directly in the notebook.

User project file
The user project file must be in json format and defines the path to the root folder where all the other files can be found. The user project file also defines the user and if database access is required (not in the present version of the notebook) the login credentials for the given user must be given. Running the notebook without a database connection, the *host*, *db* and *host_netrc_id* values should be set to *null*. See online file for more details and examples.

Process file
A process_file, whether linked directly in the notebook or via a job file (see below) must include the sub_process_id (at present the only defined sub_process_id is import_csv_single-lines) and the full paths to the data source file (data_src_FPN), the method source file (method_src_FPN) and the path to the destination directory (dst_FP) where to save the json formatted output.

If there are metadata entries that are exactly identical for all records (lines) in the data source file (e.g. the person responsible for the field sampling, the analysis or the sample logistics), then this metadata can be stated in the process file instead of being repeated for every record (line) in the data source file, as in the example below. You can add any number of processes in the array *process* in the *process file*. If there are relatively few data files to process, the best option is to put these few processes in a single process file and link that process file directly from the notebook. See online file for more details and examples.

Job file
If you have a multitude of files to process, or if you need to set varying common records (e.g. different persons responsible for the field sampling, the analysis or the sample logistics), an alternative is to use a job file that links to a set of sequential process files.

In a job file you can link to a sequence of process files either as an array directly in the job file or by linking to a pilot file (simple text file) that lists the process files to execute. List the process files to run as an array directly in the job file or link to a _pilot_file in the job_file. A *pilot file* is a simple text files that lists the *process_files* to run and where blank lines or lines starting with hashtag are ignored.

Data source file

All the actual data from the in-situ sampled soil analysis must be in the data source file. In addition, all metadata that are unique to each sample (e.g the sample point id) must also be listed in the data source file. The column headers in the data source file can, in principle, have any name, the actual content transferred to the json output files is defined in the method source file. To make life easier it is, however, better to set the header column in the data source file to the parameter name used in the json output.

Required metadata to generate the json output include:

- pilot_country,
- pilot_site,
- sample_id,
- min_depth,
- max_depth,
- sample_date,
- sample_analysis_date,
- user_email_sampling,
- user_email_logistic, and
- user_email_analysis.

Optional metadata, assigned default values (in parenthesis) if missing in the data source file, include:

- subsample (a),
- replicate (0),
- sample_preparation__name (None),
- sample_preservation__name (None),
- sample_transport__name (None),
- sample_storage__name (None),
- transport_duration_h (0), and
- comment ().

Metadata that is common across all samples listed in a data source file can instead be added as parameters in the process_file, and typically include:

- user_email_sampling,
- user_email_logistic, and
- user_email_analysis.

But also other metadata can be put in the process file instead of the data source file. If your data source file contains data only for a single site (or pilot_site) you can omit the parameters for pilot_country and pilot_site from the data source file and add them as parameters in the process_file. Similarly, if all samples are taken the same date you can also put the sample_date parameter in the process_file and omit it from the data source file. As noted above, the actual data on soil properties must be listed in the data source file. As the actual data can be in many different units, derived from a variety of methods and measured with different instruments (inlcuding different makes and models), the parameters to include in the data source file are not predefined. But they must be defined in the method source file; defined regarding three properties:

- method,
- equipment, and
- unit.

In comparison with the tools, instruments and methods included in the original AI4SH proposal, the following in-situ methods are required for completing the list of soil descriptors appearing both in the proposed EU soil monitoring law and preliminary defined by AI4SH WP 3:

- Soil volumetric sampling and analysis, and
- Soil sampling representing topsoil and subsoil.

The sampling, according to the proposed EU soil monitoring law, must also follow a stratified and random framework that generates a Coefficient of Variance (CoV) for key descriptors that is below 5 %. The EU soil law explicitly states that the sampling should follow the Bethel[1] (1989) algorithm. A framework implementing the EU soil law requirement has been established; the statistical outcome from applying this framework is pending the results from the field trial.

Other WPs have requested in-field observations to be collected as part of the pilot site (in-situ) sampling, for variables that are not possible (or difficult) to observe from extracted soil samples or using satellite images and other spatial datasets:

- land use and cover,
- erosion,
- land management history, and
- soil aggregate structure.

While site characteristics of (present) land use and cover and erosion can be determined by an experienced soil expert in the field, land management history requires interviews or records provided by the owner/manager on a field to parcel scale. Precise and objective soil aggregate structure requires a complex laboratory analysis. While experienced soil surveyors may estimate soil aggregation “in-field”, via visual and sensory examination of soil, it requires substantial knowledge and inter-calibration between surveyors. There is some potential for in-field indirect estimation of soil aggregate structure by shear stress testing, but requires more detailed scrutiny. The value of rapid in-situ determination of management history and

[1]

soil aggregate structure is undisputable, but also difficult to obtain at any degree of certainty. These variables are tentatively included in the site protocol but as non-compulsory to fill out.

Method source file

A method source file, in contrast to the data source file, must have exactly 6 columns, all with the exact headers:

- "header",
- "parameter",
- "unit",
- "method",
- "equipment", and
- "url".

Entries in the header column must correspond to column headers (exact spelling) in the data source file. You might thus edit the records in the header column in the method source file. The second column in the method source file, parameter, is the object name that will appear in the json output file. It should not be changed. The columns for unit, method and equipment should only be filled for actual records (lines) where the header column corresponds to data. For metadata these fields should all be set to "None".

For more in-depth explanations, including examples illustrated with real AI4SH in-situ data, we refer to the GitHub page.

3. In-situ data variables and method descriptions

This section presents brief method descriptions relevant to data pipelines for each variable. For more detailed descriptions and insights into how the variable protocols tie to the pipelines and different database formats, we refer to the website.

3.1. Data formats for databases

The AI4SH data for all the variables below are available in two differently formatted json objects, a flat version more suitable for direct accessibility and a nested version intended for users who wants to convert the data to a relational SQL database. The flat versions do not include any arrays whereas the nested versions are generally structured with nested arrays. Please refer to the website for more information on the flat and nested json objects, respectively, for each of the variables.

3.2. Aggregates

Aggregates, clumps of smaller soil particles glued together by biological processes, are key components for e.g. soil hydraulic properties, nutrient recycling and pool sizes. Stable soil aggregates also contribute to resistance towards wind and water erosion, and are linked to improved water capture, infiltration and storage. The smartphone apps Slakes and Moulder (a renamed version of Slakes) uses image analysis of dried, pea sized soil aggregates before and after soaking. Aggregates that resist dispersion after rewetting and *slake* (or *mould*, i.e. disintegrate) less, are more stable compared to those that break apart. A higher resistance indicate a healthier soil. Both apps are available on Apple's App store as well as Google play for Android.

Short method description
From a dried soil sample three pea sized aggregates are selected and put in a shallow dish. A smartphone with the app installed is mounted above the dish and an image taken of the dish and the three aggregates. The dish is then gently filled with water, enough to just cover the aggregates. The user can adjust the app's identification of the aggregates, and another images is captured. After 10 minutes a third image is taken and the app calculates an aggregate stability index (table 1).

Table 1. Aggregate stability recorded with each observation with the Slakes/Moulder apps.

Property	Indicator	AI4SH extended naming	Unit
Aggregate stability	aggregate-stability-index	slakes_aggregate-stability-index	unitless
Aggregate stability	aggregate-stability-index	moulder_aggregate-stability-index	unitless

3.3. eDNA

Environmental DNA (eDNA) analysis is a complex laboratory bound chain of processes called metabarcoding. Also the sampling in the field requires special equipment and the sample has to be preserved with a shielding liquid preserving the DNA during storage and transport. The raw eDNA result is a dataset of nucleotide sequences, that is then compared to database libraries (bioinformatic treatment) for determining both the organisms (or organism groups) and the processes that occur in the soil. With growing DNA sequence libraries and reduced costs for the metabarcoding, progressively more and more information can be obtained from eDNA analysis. eDNA metabarcoding is thus increasingly becoming an important method for soil health characterisation.

Short method description

The field sampling requires care to avoid DNA contamination from e.g. human DNA. Once extracted only a very small sample (milligram) is required. The sample must be preserved to prevent any deterioration or metabolic changes in the DNA nucleotides and that requires a DNA shield to be added to the sample. Once the DNA shield is added, the sample can be stored and transported.

The pipeline of processes that compose eDNA metabarcoding in the laboratory include:

- extraction,
- amplification,
- purification,
- sequencing, and
- bioinformatic treatment (database library mining).

The eDNA analyzed as part of AI4SH focused on three taxonomic domains:

- Bacteria (target gene/region: 16S),
- Archaea (target gene/region: 16S), and
- Eukarya (target gene/region: ITS).

For the Eukarya the analyses were restricted to fungi.

For each group the richness (total number of species - recorded as *alpha* and *chao1* indexes), diversity (recorded as alpha *Shannon* and alpha *Simpson* indexes) and evenness (*pielo* index) are recorded. Group related soil functions included for example chemoheterotrophy, human pathogens, plant pathogens and nitrogen fixation. Table 1 is a complete list of eDNA derived organism groups and functions recorded with each observation with the applied metabarcoding pipelines.

Table 1. eDNA derived organism groups and functions recorded with each observation with the applied metabarcoding pipelines.

Property	Indicator	AI4SH extended naming	Unit
Prokaryotes			
alpha observed richness	Prokaryotes alpha richness	metabarcoding_Prokaryotes-alpha-observed-richness	unitless

Property	Indicator	AI4SH extended naming	Unit
alpha chao1 estimated richness	Prokaryotes chao1 richness	metabarcoding_Prokaryotes-alpha-chao1-estimated-richness	unitless
alpha dominance	Prokaryotes alpha dominance	metabarcoding_Prokaryotes-alpha-dominance	unitless
alpha pielou e	Prokaryotes Pielou evenness	metabarcoding_Prokaryotes-alpha-pielou-e	unitless
alpha shannon	Prokaryotes alpha Shannon diversity	metabarcoding_Prokaryotes-alpha-shannon	unitless
alpha simpson	Prokaryotes alpha Simpson diversity	metabarcoding_Prokaryotes-alpha-simpson	unitless
functional prediction chemoheterotrophy	Prokaryotes chemoheterotrophy	metabarcoding_Prokaryotes-functional-prediction-chemoheterotrophy	unitless
functional prediction human pathogens all	human pathogens	metabarcoding_Prokaryotes-functional-prediction-human-pathogens-all	unitless
functional prediction nitrogen fixation	Nitrogen fixation	metabarcoding_Prokaryotes-functional-prediction-nitrogen-fixation	unitless
Fungi			
alpha observed richness	Fungi alpha richness	metabarcoding_Fungi-alpha-observed-richness	unitless
alpha chao1 estimated richness	Fungi chao1 richness	metabarcoding_Fungi-alpha-chao1-estimated-richness	unitless
alpha dominance	Fungi alpha dominance	metabarcoding_Fungi-alpha-dominance	uitless
alpha pielou e	Fungi Pielou evenness	metabarcoding_Fungi-alpha-pielou-e	unitless

Property	Indicator	AI4SH extended naming	Unit
alpha shannon	Fungi alpha shannon diversity	metabarcoding_Fungi-alpha-shannon	unitless
alpha simpson	Fungi alpha simpson diversity	metabarcoding_Fungi-alpha-simpson	unitless
funtional prediction Ectomycorrhizal fungi	Ectomycorrhizal fungi function	metabarcoding_Fungi-funtional-prediction-Ectomycorrhizal-fungi	unitless
funtional prediction Arbuscular mycorrhizal fungi	Arbuscular mycorrhizal fungi function	metabarcoding_Fungi-funtional-prediction-Arbuscular-mycorrhizal-fungi	unitless
funtional prediction fungal saprotrophs	Saprotrophic fungi function	metabarcoding_Fungi-funtional-prediction-fungal-saprotrophs	unitless
funtional prediction fungal plant pathogens	Plant pathogenic fungi function	metabarcoding_Fungi-funtional-prediction-fungal-plant-pathogens	unitless

3.4. Enzymatic activity

All biological processes and functions are governed by enzymes and their activities. Assessing the activity rates of soil extracellular enzymes has traditionally been an expensive analysis restricted to the laboratory and specialist personnel.

The Swiss start-up Digit Soil, a partner of the AI4SH (AI4SoilHealth) project, has developed a rapid method for quantifying five key soil enzymatic activities. The Soil Enzymatic Activity Reader (SEAR) and reaction plates developed by Digit Soil form a stand-alone system that can be used in an ordinary home or office (laboratory-independent) and is applicable for use by, for instance, citizen scientists. This move makes the complex measurement of soil functional diversity and activity simple and inexpensive.

3.5. Short method description

The Soil Enzymatic Activity Reader (SEAR), developed by Digit Soil, uses a combination of reaction plates and fluorescence spectroscopy Fetzer et al., 2025. The method requires fresh samples, which must be either analysed directly from the field or stored and transported at a low temperature (a few degrees Celsius) for no more than a few days before analysis. Reaction plates, designed with 5 x 5 wells (which include three replicates for each enzyme plus internal calibration wells), are pressed onto the soil contained in a soil tray. The tray is then inserted into a spectrometer that uses targeted UV light to excite the reactants in the reaction plate. A digital camera registers the resulting signal, which is later computed to determine the enzymatic activity rate over a 40-minute period. The enzymatic activity rates of five key enzymes are then reported. The enzymatic activities measured are summarised in table 1.

Table 1. Enzymatic activities recorded with each observation with the Soil Enzymatic Activity Reader (SEAR).

Property	Indicator	AI4SH extended naming	Unit
Chitinase/β-glucosaminidase	GLA	digit-soil-sear_GLA	pmol*min^-1
β-Glucosidase	GLS	digit-soil-sear_GLS	pmol*min^-1
phosphatases (Phosphomonesterases)	PHO	digit-soil-sear_PHO	pmol*min^-1
β-Xylosidase	XYL	digit-soil-sear_XYL	pmol*min^-1
Leucine-aminopeptidase	LEU	digit-soil-sear_LEU	pmol*min^-1

3.6. Infiltration

Infiltration, the capacity of soil to swallow water, is a prerequisite for soils and their ecosystems to function. Reduced infiltration capacity, due e.g. to compaction or water repellant conditions arising from drought, makes soils more prone to erosion and less resilient to droughts and storms. Low infiltration capacity also exacerbate flash floods and downstream flooding. As part of AI4SH a fairly simple singe-ring infiltration method (denoted as Beerkan method) is used that allows to link the infiltration rates to various soil hydraulic properties

3.7. Short method description

Single ring infiltration is a published scientific method for estimating soil hydraulic properties (Lassabatere et al., 2006). The ring can easily be constructed from a food tin can (~8-10 cm in diamter) and inserted ~1 cm into the soil. Small, fixed volumes of water (typically 50 to 100 ml) are sequentially poured into the ring and the time it takes for the water to infiltrate is recorded. The pouring of water continuous until the time it takes for it to infiltrate is stable. To calculate the hydraulic properties (table 1), data on bulk density and soil moisture before and after the infiltration test are needed - e.g. obtained using the soil cylinder bulk density and volumetric soil moisture content method.

Table 1. Hydraulic soil properties recorded with each single ring infiltration observation.

Property	Indicator	AI4SH extended naming	Unit
infiltration rate	infiltration	single-ring-infiltration_final-infiltration-rate	m*day^-1
porosity	porosity	single-ring-infiltration_porosity	vol*vol^-1
sorptivity	sorptivity	single-ring-infiltration_sorptivity	m*day^-1/2
saturated hydraulic conductivity	hydraulic conductivity	single-ring-infiltration_saturated-hydraulic-conductivity	m*day^-1
field capacity	field capacity	single-ring-infiltration_field-capacity	vol*vol^-1
plant avaiable water	plant available water	single-ring-infiltration_plant-available-water	vol*vol^-1
saturated water content	saturated water content	single-ring-infiltration_saturated-water-content	vol*vol^-1
hg water pressure head scale parameter	hg water pressure head	single-ring-infiltration_hg-water-pressure-head-scale-parameter	m
n shape parameter of retention curve	retention curve n	single-ring-infiltration_n-shape-parameter-of-retention-curve	unitless
m shape parameter of retention curve	retention curve m	single-ring-infiltration_n-shape-parameter-of-retention-curve	unitless

3.8. Ion Selective Electrode

The ionic composition, together with water, to a large extent defines the non-biological living conditions in a soil ecosystem. Key ions include hydrogen (H+, usually expressed as pH), sodium (Na+), chloride (Cl-), potassium (K+), calcium (Ca2+), magnesium (Mg2+) and different nitrogen species like nitrate (NO3-) and ammonium (NH4+). These ions can all be determined using designated Ion Selective Electrodes (ISEs). ISEs measure the activity of specific ions, usually in a solution, by converting the electrical potential across a membrane that is only permeable to a specific ion. The concentration difference between a reference solution inside the ISE and the concentration in the solution causes a potential that is null (0) when the concentration across the membrane is equal. The larger the concentration difference the larger the absolute value of the potential. An ISE measures this potential, and from 2 or more solutions with known concentrations of the target ion, can be used to estimate the concentration of a particular ion in a solution with unknown concentrations of this particular ion.

Short method description

ISEs are sensitive instruments and require carefulness when handled and calibrated. The most widely used ISE type is for measuring pH, and pH electrodes are also cheaper and more stable compared to other ISEs. In AI4SH, pH ISEs for both direct observation in the field, and for soil diluted in distilled water (at a ratio 1:5) were tested (table 1).

Table 1. pH recorded Ion Selective Electrode (ISE) observations.

Property	Indicator	AI4SH extended naming	Unit
Hydrogen ion concentration	pH(soil)	xspectre-ise-ph-solid_ph(soil)	pH
Hydrogen ion concentration	pH(water)	xspectre-ise-ph-liquid_ph(water)	pH

The ISE for direct observation in the field has a more robust tip that can be inserted directly into soft soils, or after pricking a small hole with a pencil or similarly tipped tool in more sturdy soils. It is despite this a sensitive and breakable tool and was only tested at a single pilot site. Ordinary ISEs are built for sampling solutions, and was tested at several pilot sites as part of the AI4SH in-situ sampling. Following the standard procedure when measuring pH, the AI4SH solution applied method diluted soil samples in distilled water at a ratio 1:5 and then measured the pH with a standard pH ISE. For calibration, pH buffers of pH=4.01 and pH=10.01 were used. The microcontroller unit for running and recording the pH ISE observations in AI4SH is constructed by the Swedish startup Xspectre.

3.9. Microbiometer

Microbiometer is a commercial kit for quick estimation of soil microbial carbon biomass and the ratio between bacteria and fungi. Within AI4SH, there has not yet been any rigorous testing of this method in comparisons to more established laboratory methods. The method is still included here pending more careful assessment of the method for support of soil property or soil health assessment.

Short method description

The commercial Microbiometer test mixes a small (1 ml) soil sample with a reagent solution whereafter the mixture is rested for 20 minutes while the reaction is completed and the solution settles. Three drops of the supernatant are extracted using a provided pipette and applied to a likewise provided test card. The test card is then scanned with a smartphone using the Microbiometer app. If the colour change of the drop area on the test card is too weak, additional three drops are added and the test card re-scanned. From the colour intensity of the drop area, the app reports the soil's microbial biomass carbon content and fungal-to-bacterial ratio (table 1). The test takes approximately 30 minutes in total. The app and the test card were updated during 2025 and thus the AI4SH reported results relate to both the *classic* and *pro* versions of the Microbiometer.

Table 1. Microbiometer recorded observations with each observation.

Property	Indicator	AI4SH extended naming	Unit
Microbial biomass	microbial-C	microbiometer_Microbial-C	ug*g^-1
Fungi fraction	fungi-fraction	microbiometer_fungi-fraction	%
Baceteria fraction	bacteria-fraction	microbiometer_bacteria-fraction	%

3.10. Penetrometer

Physically robust and microchip controlled penetrometers for analysing physicochemical soil properties directly in-situ have become available over the past few years. Using multiple steel pins, developments in microelectronics and signal interpretation these simple instruments can now separate water and salt content and also detect specific ions, like hydrogen (pH) and those derived from nitrogen, phosphorus and potassium (NPK) – the key soil macronutrients. Soil penetrometers that can operate at the power levels of mobile phones have emerged only recently and are as yet scientifically unproven. The AI4SH project tested the 5-pin penetrometer model NPKPHCTH-S from ComWinTop store on AliExpress operating at 5 volt and mounted with a microcontroller and usb/Bluetooth connection, developed by the startup Xsepctre.

Note the following issues related to the penetrometer data when applying it for further analysis:

- the penetrometer data is only available for a subset of pilot sites,
- in most cases each sample was analysed using more than 1 subsample (thus there are more than one record per sample), and
- for the Finnish site (Jokioinen), 3 different copies of the same penetrometer model where used to allow evaluating bias and consistency.

When considering this data, it is useful to consider the difference between what is a subsample, replicate and repetition. The difference between a subsample and a replicate is that if the same physical volume of soil is used for the analysis it is a replicate, but if the analysis is applied to a separate physical volume it is a subsample. Thus all destructive analysis methods can per definition not be replicates. For the soil penetrometers to apply a replicate would mean letting it stay in the exact same position for two sequential observations. As each penetrometer observation is already based on 6 repetition and the recorded data is stated as mean value and standard deviation (see example below), one more observation in the exact same position renders no new information. Thus, for the penetrometer analysis done as part of AI4SH, the position was typically shifted to different sides in the dug central pit, and each observation recorded as a subsample.

Short method description

The penetrometer is simply pushed into the soil, either from the top or vertically in a pit. The observation is started from the device (computer or smartphone) connected to the microcontroller unit. In the setup used in AI4SH, each observation is repeated 6 times and the user must approve of the results displayed as mean and standard deviation before the observation is recorded. The penetrometer used in the AI4SH project simultaneously records 9 soil parameters with each observation, table 1.

Table 1. Soil properties recorded with each observation with the ComWinTop NPKPHCTH-S 5 steel pin penetrometer.

Property	Indicator	AI4SH extended naming	Unit
temperature	temperature	xspectre-penetrometer_temperature	C
Soil moisture	soil-moisture-volumetric-content	xspectre-penetrometer_soil-moisture-volumetric-content	vol*vol^-1
pH	pH(soil)	xspectre-penetrometer_ph(soil)	pH
Nitrogen	N	xspectre-penetrometer_nitrogen	mg*l--^-1
Phosphorus	P	xspectre-penetrometer_phosphorus	mg*l^-1
Potassium	K	xspectre-penetrometer_potassium	mg*l^-1
Electric conductivity	electrical-conductivity	xspectre-penetrometer_electrical-conductivity	us*cm^-1
Salinity	salinity	xspectre-penetrometer_salinity	mg*l^-1
Total Dissolved Solid	total-dissolved-solids	xspectre-penetrometer_total-dissolved-solids	mg*l^-1

3.11. Salinity

Salinisation is recognised as one of the largest threats to soil health in Europe and elsewhere. A simple way to test soil salinity is to dissolve a small amount of a soil sample is distilled water and estimate the salinity from the resistance between two (or more) electrodes.

Short method description

Dissolved salts decrease the resistance to electric currents, whether in (wet) soil or a solution. Measuring the resistance across two electrodes is a simple way to determine the resistance and by calibrating the observed resistance against standard solutions, the salinity of a sample can be determined (table 1). In AI4SH a simple bi-pin electrode operated from a microcontroller built by the Swedish startup Xspectre is used as a citizen scientist approach for observing soil salinity. The procedure includes taking a small fraction of soil and dissolving it in distilled water at a ratio 1:5 and then observe the soil salinity. For calibration, two standard solutions, one with a high and one with low salinity are used. Note that the AI4SH project also measured salinity directly in the field with the penetrometermethod.

Table 1. Bi-pin electrode recorded total dissolved solids (TDS) in solutions of soil dissolved in distilled water at a ratio 1:5.

Property	Indicator	AI4SH extended naming	Unit
Total dissolved solids	total-dissolved-solids	xspectre-gx16-ec_total-dissolved-solids	ppm

3.12. Spectroscopy

Light is progressively used as an analytical tool for determining chemical and physical compositions in astronomy, industrial manufacturing, food and pharmaceuticals processing, and science - including for laboratory soil analysis. Recent developments in microelectronics, light sensing technology and artificial intelligence (AI), have led to both field spectrometers for in-situ soil analysis and to cheaper layperson spectrometers that are also possible to use for soil analysis. Soil spectroscopy is progressively becoming a mature method for determining soil properties. Spectroscopy cannot determine all properties that are traditionally analysed using wet laboratory methods, but many key attributes are already routinely determined by spectroscopy also in commercial soil laboratories.

Short method description

The AI4SH in-situ soil sampling and subsequent spectroscopic analysis looked at soil samples prepared in three different ways:

- undisturbed soil directly in the field,
- mixed and (naturally) wet soil samples, and
- dried and sieved soil samples.

Existing soil spectral libraries, like the Open Soil Spectral Library (OSSL) are built from dried and sieved (2 mm) samples. The potential for determining the chemical and physical properties using soil spectroscopy is thus higher when using spectra from dried and sieved samples. Direct field scanning and spectral scanning of wet samples is more rapid and cost effective, but non-linearly affected by varying water contents, and were also tested as part of AI4SH.

Further, to compare the performance of high-grade laboratory instruments with both commercial handheld soil spectrometers and cheap (layperson applicable) pocket-sized spectrometers, a suite of spectrometers is used. Most instruments were applied to all three preparation levels, but with restricted testing done with laboratory grade instruments directly in the field. The different instruments used in the project are listed in table 1.

Table 1. Spectrometers applied as part of the AI4SH in-situ soil sampling.

spectrometer	spectral range (nm)	nr of bands	applied
FOSS NIRS DS2500	400-2500	4200	ex-situ
LabSpec ASD*	350–2500	2151	in-situ, ex-situ
Neospectra/ProxiScout	1300-2550	257	in-situ, ex-situ
Xspectre_c12880ma	340-780	288	in-situ, ex-situ
Xspectre_c14384ma-01	650-1050	192	in-situ, ex-situ

*Data from the LabSpec instrument are under processing and not reported as of December 2025

Table 2 Spectral reflectance observation records of the 4 different spectrometers applied within AI4SH.

Property	Indicator	AI4SH extended naming	Unit
diffuse spectral reflectance	reflectance	foss-ds2500-scan_reflectance	reflectance
diffuse spectral reflectance	reflectance	neospectra-scan_reflectance	reflectance
diffuse spectral reflectance	reflectance	xspectre-c12880ma_reflectance	reflectance

Property	Indicator	AI4SH extended naming	Unit
diffuse spectral reflectance	reflectance	xspectre-c14384ma-01_reflectance	reflectance

3.13. Wet laboratory

Traditional chemicophysical soil properties have been determined by wet laboratory methods. While traditional methods are restricted to chemical and physical properties, disregarding biological composition and processed, they are still the benchmark against which novel physicochemical methods are tested and evaluated.

Method description

For almost all collected sample within the AI4SH in-situ campaign, 15 soil properties were analysed using traditional wet laboratory methods (table 1). The majority of the samples were analysed at the German Agrolab. The exception is the trial sampling at the Greek pilot site Ktima-Gerovassiliou.

Table 1. Soil properties analysed in samples from most AI4SH pilot sites at the Agrolab facility.

Property	Indicator	AI4SH extended naming	Unit
pH	pH(water)	agrolab_ph(water)	pH
Electrical conductivity	electrical-conductivity	agrolab_electrical-conductivity	us*cm^-1
Total Organic Carbon	total-organic-carbon	agrolab_total-organic-carbon	percent
Total Nitrogen	total-nitrogen	agrolab_total-nitrogen	percent
Calcium	calcium	agrolab_calcium	cmol*kg^-1
Magnesium	magnesium	agrolab_magnesium	cmol*kg^-1
Potassium	potassium	agrolab_potassium	cmol*kg^-1
Sodium	sodium	agrolab_sodium	cmol*kg^-1

Property	Indicator	AI4SH extended naming	Unit
Cation exchange capacity	cation-exchange-capacity	agrolab_cation-exchange-capacity	cmol*kg^-1
Olsen phosphorus	olsen-phosphorus	agrolab_olsen-phosphorus	mg*100g^-1
clay(<0.002mm)	clay(<0.002mm)	agrolab-wet-clay(<0.002mm)	g*100g^-1
fine-silt(0.002-0.02mm)	fine-silt(0.002-0.02mm)	agrolab_fine-silt(0.002-0.02mm)	g*100g^-1
coarse-silt(0.02-0.06mm)	coarse-silt(0.02-0.06mm)	agrolab_coarse-silt(0.02-0.06mm)	g*100g^-1
fine-sand(0.06-0.2mm)	fine-sand(0.06-0.2mm)	agrolab_fine-sand(0.06-0.2mm)	g*100g^-1
coarse-sand(0.2-2.0mm)	coarse-sand(0.2-2.0mm)	agrolab_coarse-sand(0.2-2.0mm)	g*100g^-1

4. Earth Observation (EO) satellite data

Earth Observation (EO) data is progressively becoming more important for understanding patterns and processes at the Earths's surface. Properties related to soil health that can be derived from EO-data include landscape and landform scale patterns, vegetation cover, type and production and soil physicochemical properties. As part of the AI4SoilHealth project, within WP5, a pan-European Soil Health Data Cube with EO (satellite) data and derived static and dynamic properties related to the Earths's terrestrial land surface has been developed under the leadership of OpenGeoHub. This section illustrates how static landscape scale properties derived from the Soil Health Data Cube can be added to the AI4SH in-situ database. The purpose with this is to enable direct comparison between in-situ data and EO data in offline contexts or for specific purposes, such as the AI4SH App. For there purposes a Jupyter Notebook has been developed, and made available at this URL: **https://github.com/AI4SH/EO-data_access**

4.1. Brief Description of Point Data Extraction from the AI4SH Datacube

The Jupyter Notebook *Extract_AI4SH_datacube_points*, extracts static environmental values from the AI4SoilHealth (AI4SH) Datacube for specified in-situ point locations. It reads coordinate points from an Excel file, samples raster layers hosted on the EcoDataCube S3 server, and exports the results to a CSV file. The notebook enables researchers to:

- Load AI4SH in-situ coordinate points from an Excel file
- Convert point data to a GeoDataFrame with proper coordinate reference systems
- Extract values from multiple static raster values (terrain derivatives, crop types, etc.)
- Handle coordinate reprojection automatically when needed
- Save the enriched point data with extracted values

In brief, the **workflow should follow these steps**:

- Environment Setup: Load required Python libraries
- Data Loading: Read in-situ coordinate points from Excel file
- Spatial Data Preparation: Convert points to GeoDataFrame with EPSG:4326 CRS
- Covariate Selection: Define list of raster layer URLs from AI4SH Datacube
- Value Extraction: Loop through each layer and sample values at point locations
- Data Export: Save results to CSV file with all extracted values

Python Package requirements

The Notebook requires the following third party Python packages:

- pandas
- geopandas
- rasterio
- numpy
- openpyxl
- gdal

The terminal command for installing these packages with Anaconda is:

```
conda create -n ai4sh_datacube_access_312 -c conda-forge  pandas geopandas rasterio
numpy openpyxl gdal python=3.12
```

Input Requirements

- Excel file: AI4SH_in-situ_coordinate_points.xlsx containing columns for longitude and latitude
- Directory structure: Excel file should be located in ../AI4SH_point_locations/ relative to notebook

Output

- CSV file: AI4SH_in-situ_points_with_static_values.csv saved to ../AI4SH_point_data/
- Contains original point data plus columns for each extracted value

Data Sources

Static values are accessed from the AI4SoilHealth SoilHealthDataCube via HTTPS, including:

- Terrain derivatives (slope, curvature, hillshade, TWI, etc.)
- Geomorphological features (geomorphons, openness indices)
- Topographic indices (LS-factor, shape index)
- Land cover/crop type data

Dependencies

- Python 3.12
- pandas: Data manipulation and Excel/CSV I/O
- geopandas: Spatial data operations and coordinate transformations
- rasterio: Raster data reading and sampling
- numpy: Numerical operations and handling missing values
- openpyxl: Excel file reading support
- gdal: supports rasterio geo-data processing

Notes

- The notebook handles coordinate reprojection automatically when raster CRS differs from point CRS
- NoData values are converted to NaN for proper handling in pandas
- Progress messages indicate which value is currently being processed
- Internet connection required to access remote raster data from S3 server

Licenses

The data is provided under the following licenses:

- Data License: Creative Commons Attribution license (CC-BY)
- Code License: Massachusetts Institute of Technology License (MIT License)

5. References

Bethel, J. (1989) “Sample Allocation in Multivariate Surveys.” Survey Methodology 15: 47–57.

Fetzer, J., Meller, S., Iven, H., Baur, D., García Rivera, P., Meller, A. & Luster, J. (2025) ‘Novel, laboratory-independent device to measure extracellular enzymatic activity in soils’, Frontiers in Environmental Science, 13, p. 1663635. doi: 10.3389/fenvs.2025.1663635