SoilGrids1km — global soil information based on automated mapping

Tomislav Hengl [a]Jorge Mendes de Jesus [a] Robert A. MacMillan [b] Niels H. Batjes [a] Gerard B.M. Heuvelink [a,c] Eloi Ribeiro [a] Alessandro Samuel-Rosa [d] Bas Kempen [a] Johan G.B. Leenaars [a] Markus G. Walsh [e] Maria Ruiperez Gonzalez [a]

[a]*ISRIC — World Soil Information, Wageningen, the Netherlands*

[b]*LandMapper Environmental Solutions Inc., Edmonton, Canada*

[c]*Wageningen University, Wageningen, the Netherlands*

[d]*Federal Rural University of Rio de Janeiro, Brazil*

[e]*The Earth Institute, Columbia University, USA / Selian Agricultural Research Inst., Arusha, Tanzania*

Abstract

Background. Soils are widely recognized as a non-renewable natural resource and as biophysical carbon sinks. As such, there is a growing requirement for global soil information, especially in terms of maps of primary and derived soil properties. Although several global soil information systems already exist, these tend to suffer from inconsistencies derived from differing soil mapping concepts across borders and limited spatial detail. *Methodology/Principal Findings.* To address the growing demand for consistent and comprehensive soil data, we present SoilGrids1km — a global 3D soil information system (a stack of soil property and class maps at six standard depths) at 1 km resolution. SoilGrids1km is currently comprised of spatial predictions for a selection of soil properties: soil organic carbon (g kg^{-1}), soil pH, sand, silt and clay fractions (%), bulk density (kg m^{-3}), cation-exchange capacity (cmol+/kg) of the fine earth fraction, coarse fragments (%), soil organic carbon stock (t ha^{-1}), depth to bedrock (cm), World Reference Base soil groups, and USDA Soil Taxonomy suborders; new soil properties and classes will be continuously added. Predictions are based on global models which we fitted, per soil variable, using a compilation of major international soil profile databases (ca. 110,000 soil profiles), and a selection of global environmental covariates representing soil forming factors (ca. 75 covariate layers). Spatial predictions include per pixel uncertainties provided as 90 % prediction intervals. Results of regression modeling indicate that the most useful covariates for modeling soils at the global scale are climatic and biomass indices (based on MODIS images), lithology, and taxonomic mapping units derived from conventional soil survey (Harmonized World Soil Database). Prediction accuracies assessed using 5–fold cross-validation were between 23–51%. *Conclusions/Significance.* SoilGrids1km provide an initial set of examples of soil spatial data for input into global models at a resolution and consistency not previously available. Some of the main limitations of the current version of SoilGrids1km are: (1) weak relationships between soil properties/classes and explanatory variables due to scale mismatches, (2) difficulty to obtain covariates that capture soil forming factors, (3) low sampling density and spatial clustering of soil profile locations, (4) noise due to partially-harmonized soil profile data. However, as the SoilGrids system is highly automated and flexible, increasingly accurate predictions can be generated as new input data become available and new modelling approaches are tested. SoilGrids1km maps are available for download via `http://soilgrids.org` under a Creative Commons Non Commercial license.

Introduction

There is increasing recognition of the urgent need to improve the quality, quantity and spatial detail of information about soils to respond to challenges presented by growing pressures on soils to support a large variety of critical functions [1,2,3,4]. Arrouays et al. [3] argue that existing soils information is not well suited to addressing vital questions related to mapping, monitoring or modelling soil processes that are driven by fluxes or changes in soils of water, nutrients, carbon, solutes or energy. Conventional models of soil variation describe variation in the horizontal dimension using polygons comprising classes of named soils [5]. In the vertical dimension, variation is described in terms of classes of horizons or layers that vary in their properties, thickness and depth. These conceptual models of discrete variation of classes of soil in horizontal and vertical directions are not well suited for use in many of the (global) simulation models and decision making systems currently used to describe and interpret soil functions and processes, such as supporting crop growth modelling, modelling hydrological and climatological processes, soil carbon dynamics or erosion [5,2].

Email address: `tom.hengl@wur.nl` (Tomislav Hengl).

Most modern spatial models that require information about soils as an input need accurate numerical information about continuous variation in soil properties. Models also require input data layers that are complete, consistent and as correct and current as possible. These requirements are not well met by current sources of soils information, especially sources of global extent.

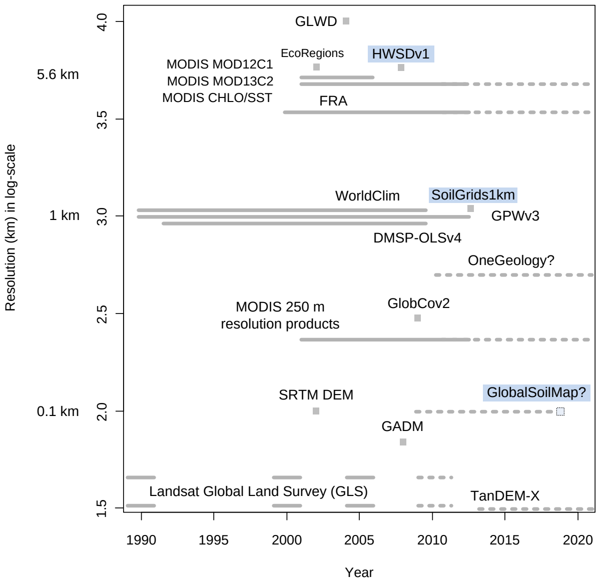


Fig. 1. **Spatial resolution and temporal coverage / publication time of some widely used global environmental data layers (global soil layers have been highlighted)**: GLWD — Global Lakes and Wetlands Database, HWSD — Harmonized World Soil Database, MOD12C1 — MODIS Land Cover Type Yearly L3, MOD13C2 — Vegetation Indices Monthly L3, CHLO/SST — MODIS Aqua Level-3 annual Chlorophyll / mid-IR Sea Surface Temperature, FRA — Forest Resources Assessment, GPW — Gridded Population of the World, DMSP-OLS — Nighttime Lights Time Series, GlobCov — Land Cover classes based on the MERIS FR images, GADM — Global Administrative Areas, TanDEM-X — Germany's topographic radar mission. Key agenda setters in the terms of production and dissemination of remote sensing and thematic environmental layers at the beginning of the 21st century include: NASA's MODIS (Moderate-resolution Imaging Spectroradiometer) and Landsat products — in terms of thematic content and usability [6,7,8], and Germany's TanDEM-X new global 12 m resolution DEM with ±2 m vertical accuracy [9]. Based on information retrieved on February 15th 2014.

Soil is probably one of the least well described thematic layers at the global scale, and existing global soil maps are often of undocumented or unknown accuracy [5]. At the moment, only coarse scale soil maps of the world are available at an effective resolution of about $\sim$20 km [10]. The most commonly used global soil maps include [5,2]: Harmonized World Soil Database (HWSD) [11], USGS-produced soil property maps [1] and ISRIC-WISE based soil property maps [12].

While widely used and cited, these various coarse resolution soil maps tend to suffer from artefacts due to use of different soil mapping concepts between countries and regions, from variation in the underlying soil mapping scale (usually between 1:0.5M to 1:5M) and from differences in reliability of source data within and between continents [5,2]. They can also not easily be updated with new information and often lack any measure of uncertainty, which is assumed to be significant. In summary, currently available global soil maps are not comparable in level of detail, spatial accuracy and usability with other global environmental layers such as global land cover and climatic products (Figure 1).

In this paper, we present and describe SoilGrids1km — a global 3D soil information system at 1 km resolution — as a first response to the need for new, consistent and coherent, global soil information. SoilGrids1km was produced using the Global Soil Information Facilities (GSIF), which was recently developed at ISRIC as a framework and platform to support widespread, open collaboration in the assembly, collation and production of global soil information.

Materials and Methods

Global Soil Information Facilities

ISRIC — World Soil Information has a mandate to serve the international community with information about the world's soil resources to help address major global issues. Over the last four years, in collaboration with a growing number of international partners and with direct support from the Bill and Melinda Gates Foundation (AfSIS project [2]), ISRIC has been developing a cyberinfrastructure called Global Soil Information Facilities (GSIF).

GSIF has a particular emphasis on supporting the assembly and collation of geo-registered soil profile descriptions with associated analytical data, and on supporting the production of new maps of 3D continuous soil properties and soil classes at global to regional scales. GSIF consists of several components: data portals for assembling and hosting soil profile data and covariate data, software for global soil data analysis and mapping, and facilities for documenting data and methods and for automating workflows.

One of these components is "SoilGrids" — an automated system for global soil mapping. SoilGrids is an implementation of model-based geostatistics [13,14] for the purpose of predicting soil properties (in 2D or 3D) and soil classes for a global soil mask (see Figure 3c) using automated mapping. Automated mapping is the computer-aided generation of maps from point observations and covariate layers, with minimal human intervention, so that map updating is easy. In the context of geostatistical mapping, automated mapping implies that model fitting, prediction and visualization

[1] http://soils.usda.gov/use/worldsoils/mapindex/

[2] http://africasoils.net

are run using fully automated and reproducible workflows [15,14]. The current implementation of SoilGrids focuses on producing predictions at 1 km spatial resolution and for a selection of soil properties and classes of interest to modelers and to international organizations such as FAO, Intergovernmental Panel on Climate Change (IPCC), the Consultative Group on International Agricultural Research (CGIAR) and similar.

We have imagined GSIF as a crowd-sourcing system, largely inspired by systems such as OpenStreetMap, Geo-wiki [16] and the R Open Source environment for statistical computing [17]. In this context, GSIF follows the "Agile" approach to software / IT development [18] meaning that we support rapid development, integration of soil field data, output validation, and rapid publishing of results. A new development cycle with new outputs (in principle of improved accuracy) is implemented in succession within an automated processing framework until the desired target specifications have been reached.

Input data for SoilGrids1km

The main input data sources for SoilGrids1km are global compilations of publicly available (shared) soil profile data and environmental layers at 1 km resolution; both are freely accessible via portals[3] . The main sources of soil profile data used to produce the first version of SoilGrids1km are: the USA National Cooperative Soil Survey Soil Characterization database[4] and profiles from the USA National Soil Information System[5] , LUCAS Topsoil Survey database [19], Africa Soil Profiles database [20], Mexican National soil profile database [21], Brazilian national soil profile database [22], Chinese soil profile database [23], and the soil profile archive from the Canadian Soil Information System [24]. Other significant sources of profile data used are: ISRIC-WISE [25], SOTER [26], SPADE [27], and Russian soil reference profiles [28].

The compilation of points shown in Figure 2 is possibly the largest collection of soil ground-truth data in the world. It can be compared, for example, to a compilation of meteorological station data used to generate the WorldClim images [29]. A large part of the soil profile data used to generate SoilGrids1km can be accessed via the WorldSoilProfiles.org data portal, however some data sets such as LUCAS [19] have strict data use policies and can only be obtained from the original data provider.

As covariates for SoilGrids1km we used a selection of GIS layers (75): mainly MODIS images, but also climate surfaces [29], Global Lithological Map (GLiM) [30], HWSD mapping units [11], and SRTM DEM-derived surfaces. These layers (apart from the GLiM) are all available via the WorldGrids.org data portal. The actual number of covariates used during the analyses is different for each soil variable as these are iteratively selected for each soil attribute, based on their statistical significance to help predict the specific attribute.

Before model fitting, the original covariates were converted to principal components ($n = 95$) to reduce data overlap and help remove noise and artefacts [7]. Number of components is larger than the number of original covariates because covariates such as lithology and land form classes are converted to indicators before the principal component analysis.

Soil mask map

We make no spatial predictions for global land cover categories that represent non-active soil areas, such as: artificial surfaces and associated areas (>50 % of pixel covered with urban areas), bare rock areas, water bodies [31], shifting sands, permanent snow and ice. The global mask map of soils with vegetation cover and world deserts is shown in Figure 3c.

The soil mask map was derived using the long term MODIS LAI images (MOD15A2), MODIS land cover product (MOD12Q1) [6], and global water mask [31] products. We distinguish three classes in the soil mask:

(1) soils with vegetation cover — pixels with MODIS LAI > 0 for at least one month in the last 12+ years (2000–2011),
(2) urban areas — equal to the MODIS land cover product *"Urban and built-up"* class,
(3) bare soil areas — areas without any biological activity but classified as *"Barren or sparsely vegetated"* in the MODIS land cover product.

Spatial prediction models

Two groups of spatial prediction models were implemented:

(1) 2D or 3D regression and/or regression-kriging [32,33] combined with splines for numerical properties as implemented in the GSIF package for R. Here, the regression part is fitted using either:
    - Multiple linear regression [34] (for predicting pH, sand, silt and clay percentages and bulk density),
    - General Linear Models (GLM's) with log-link function [35,36] (for predicting organic carbon content and CEC),
    - Zero-inflated models [37] (for predicting coarse fragments and depth to bedrock; Figure 4),
(2) Multinomial logistic regression (as implemented in the nnet package for R) for predicting distribution of soil classes [36].

As a general framework for mapping soil properties and classes we use the regression-kriging method commonly

[3] http://soilprofiles and http://worldgrids.org
[4] http://ncsslabdatamart.sc.egov.usda.gov/
[5] http://soils.usda.gov/technical/nasis/

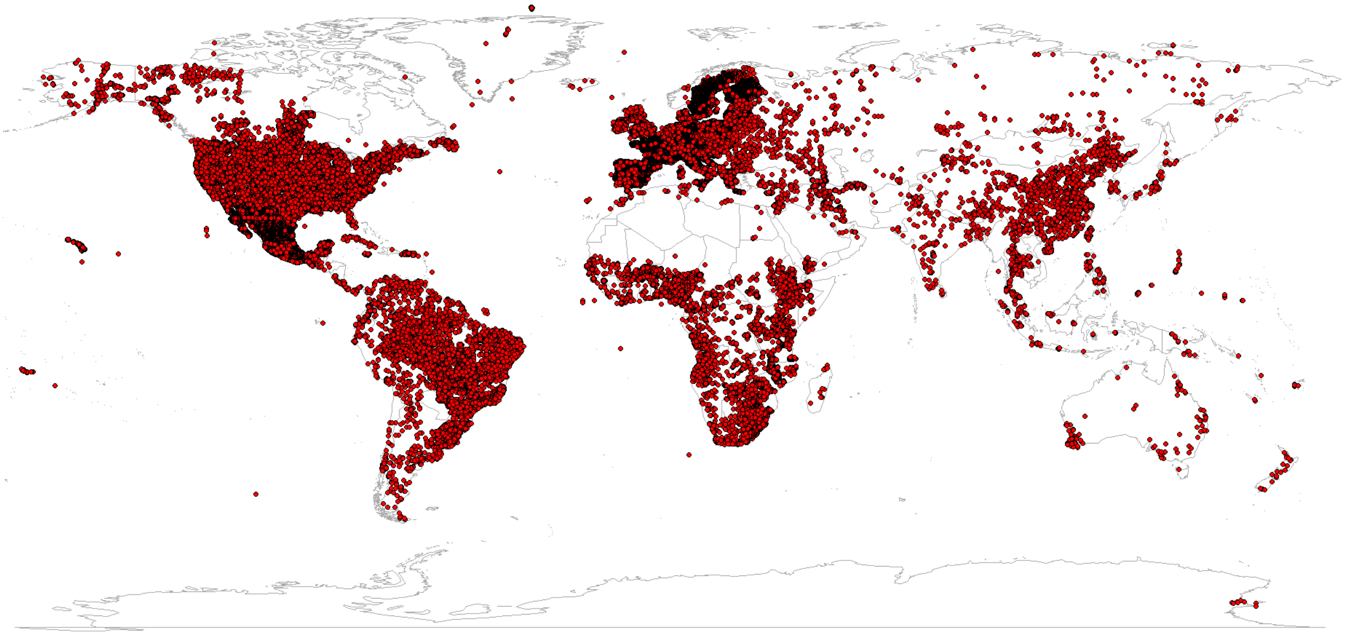


Fig. 2. **World distribution of soil profiles used to generate the SoilGrids1km product (about 110,000 points)**. Courtesy of various national and international agencies (see: Acknowledgments).

used in geostatistical mapping of soil properties [32,33,38]. We extend the existing 2D regression-kriging method to 3D space i.e. to predict values at voxels (Figure 4 right). In addition, we combine regression with splines, so that relationships between the soil property and covariates as well as soil-depth are modelled simultaneously:

$$\hat{z}(\mathbf{s}_0, d_0) = \sum_{j=0}^{p} \hat{\beta}_j \cdot X_j(\mathbf{s}_0, d_0) + \hat{\mathbf{g}}(d_0) + \\ + \sum_{i=1}^{n} \lambda_i(\mathbf{s}_0, d_0) \cdot e(\mathbf{s}_i, d_i) \quad (1)$$

where $\hat{z}$ is the predicted soil property, $\mathbf{s}_i$ are geographical coordinates, d_i is depth expressed in meters below land surface. Note that $\hat{\beta}_j \cdot X_j$ and d_0, $\hat{\mathbf{g}}(d_0)$ are the trend part of the model, where $X_j(\mathbf{s}_0, d_0)$ are covariates at the target location $\mathbf{s}_0$ and depth d_0, $\hat{\mathbf{g}}(d_0)$ is the predicted vertical trend, modelled by a spline function, and $e(\mathbf{s}_i, d_i)$ are residuals interpolated using 3D kriging using kriging weights $\lambda_i(\mathbf{s}_0, d_0)$. Because all covariates in the current version of SoilGrids1km are in fact 2D (i.e. values available at surface or for topsoil only), we copy the values of covariates for all depths in the regression matrix, which is a simplification. With the increasing availability of gamma radiometrics and similar, we anticipate that also 3D covariates will be used more in the near future with values differing per depth, although many covariates (e.g. elevation) will always remain 2D by definition.

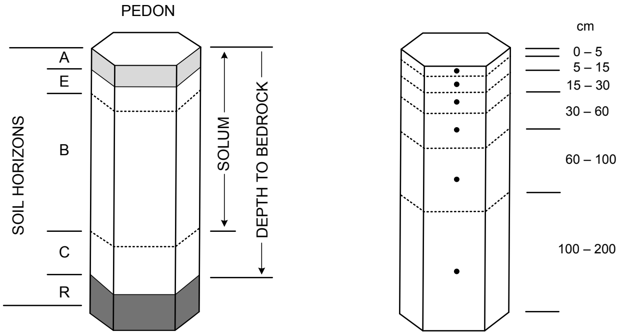


Fig. 4. **Standard stratification and designation of a soil profile: (left) soil horizons, solum thickness and depth to bedrock ('R' layer), and (right) six standard depths used in the *GlobalSoilMap* project [3]**.

3D regression and/or regression-kriging can be considered novel approaches to modeling soil variation. For comparison, the GlobalSoilMap project[6] proposes that soil-depth spline functions and spatial prediction functions should be fitted separately [40,3]. This spatial prediction system can be considered 2.5D because 2D models need to be fitted for each standard depth, i.e. each depth is modelled using a separate model that includes different combinations of covariates and in which data from predictions at one depth do not influence predictions at another. In the case of 3D modelling, a single model (Eq.1) is used for predicting in both X, Y and d for any property or class of interest, and fitting of the regression equation and residuals occurs at the same

[6] http://globalsoilmap.net

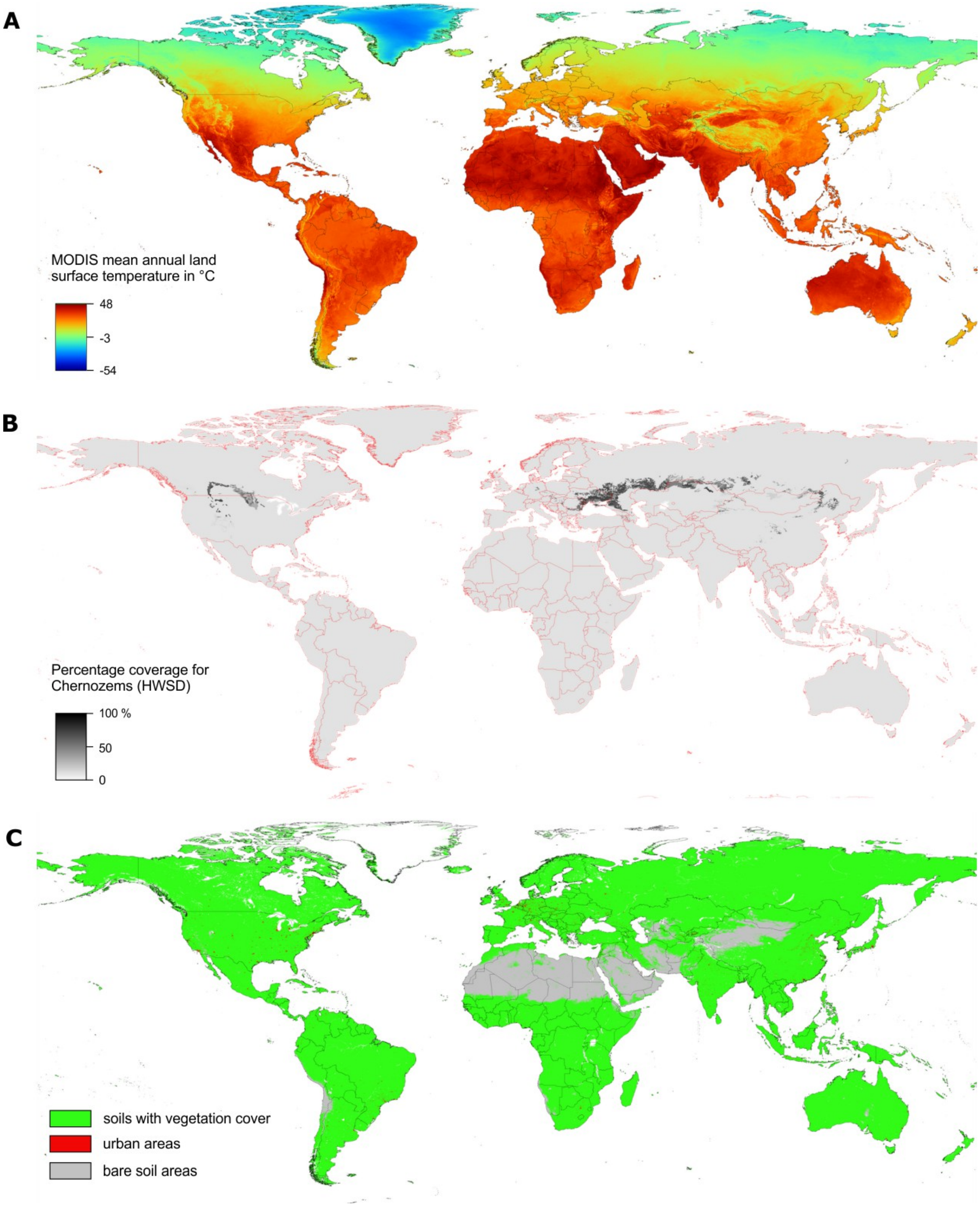

A
MODIS mean annual land surface temperature in °C
48
-3
-54
B
Percentage coverage for Chernozems (HWSD)
100 %
50
0
C
🟩 soils with vegetation cover
🟥 urban areas
⬜ bare soil areas

Fig. 3. **Examples of input layers used to generate SoilGrids1km**: (a) long-term day-time MODIS land surface temperature, (b) percent cover Chernozems (based on the HWSD data set), and (c) global soil mask map. The spatial prediction domain of SoilGrids1km are the areas with vegetation cover and urban areas, while bare soil areas have been masked out. See text for more explanation.

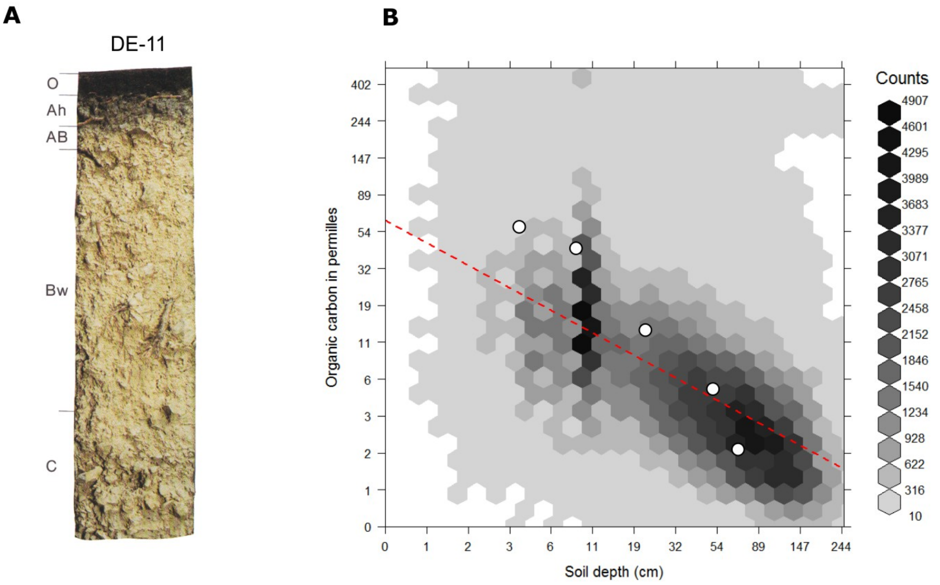


Fig. 5. **Individual soil profile from the ISRIC soil monolith collection (a) and globally fitted regression model for predicting soil organic carbon using depth only (b)**. The individual profile horizons are described by Mokma and Buurman [39]. Adjusted R-square for the model on the right is 0.363. Open circles show measured values for the profile on the left.

time as part of a single step. Another advantage of using a full 3D spatial prediction system, in comparison to the 2.5D, is also that it allows for producing spatial predictions and confidence intervals at any 3D location and not only at standard depths.

For each soil property, we have evaluated which version of the model in Eq.(1) would be most applicable. For example, initial tests showed that, for some soil properties e.g. soil organic carbon content and bulk density, the soil-depth relationship ($\hat{\mathbf{g}}(d_0)$) can often be better modelled using a log-log relationship. Consider for example:

$$\widehat{\texttt{ORC}}(d) = \exp(\tau_0 + \tau_1 \cdot \log(d)) \tag{2}$$

where $\widehat{\texttt{ORC}}(d)$ is the predicted soil organic carbon content at depth d and τ_1 is the rate of decrease with depth. The model fitted using the global compilation of soil profiles (Figure 5b) has $\tau_0 = 4.1517$ (standard error 0.005326) and $\tau_1 = -0.60934$ (standard error 0.00145). This model explains 36 % of the variation in the log-transformed ORC, which is a significant portion. This illustrates that any global soil property model can significantly profit from including depth into the statistical modelling. For other soil properties that do not show a monotonic vertical trend, higher order splines implemented via the ns function in the package splines [35] have been used to account for complex, nonlinear relationships.

Further, soil covariate layers (X_j) used to produce SoilGrids1km were selected to represent the CLORPT model originally presented by Jenny [41,38]:

$$S = f(cl, o, r, p, t) \tag{3}$$

where S stands for soil (properties and classes), cl for climate, o for organisms (including humans), r is relief, p is parent material or geology and t is time. Most of the cl, o, r, p, t covariates are now publicly available and can be obtained at low cost thanks to NASA's/USGS Earth Observation projects such as MODIS and SRTM. We have also included soil class information (WRB reference groups) extracted from the HWSD (Figure 3b). These are basically traditional soil polygon delineations, comparable to other categorical covariates e.g. land cover classes or geological units.

The 3D regression function used for modelling changes of the of soil organic carbon content in 3D was thus (in R syntax):

```
R> formulaString = (ORCDRC + 1) ~ PC1 + PC2 + . . . + PC95
   + ns(altitude, df = 2)
R> glm(formula = formulaString, family =
   gaussian(link = log), data = rmatrix)
```

where ORCDRC is the organic carbon content, PC1 to PC95 are the principal components derived from some 75 covariate layers representing Jenny's soil forming factors, altitude is depth in meters from the soil surface, rmatrix is the regression matrix with values of target variable and predictors, ns is the natural spline function and df = 2 sets the number of allowed breakpoints (in this case two breakpoints to allow for curvilinear relationship).

Soil classes are useful *'carriers of soil information'* [42], hence for SoilGrids1km we also provide global predictions for standard soil classes classified according to the two most widely used international soil classification systems:

- FAO's World Reference Base (WRB) — with focus on mapping soil groups e.g. Chernozem, Luvisols, Gleysols and similar. The current system [43] defines 32 reference soil groups.
- United States Department of Agriculture (USDA) Soil Taxonomy — with focus on mapping the soil suborders. The current system [44] defines 67 soil suborders (subdivision of 12 orders: Alfisols, Andisols, Aridisols, Entisols, Gelisols, Histosols, Inceptisols, Mollisols, Oxisols, Spodosols, Ultisols and Vertisols).

Models for predicting WRB soil groups and USDA soil orders were fitted using the nnet package (fits multinomial log-linear models via neural networks) using the default settings of 100 maximum iterations [36]. Soil classes are modeled as 2D variables i.e. the model does not include depth component, e.g.:

```
R> formulaString = TAXGWRB ~ PC1 + PC2 + ... + PC95
R> nnet::multinom(formula = formulaString,
     data = rmatrix, MaxNWts = 7000)
```

where TAXGWRB is the field observed WRB soil group, nnet::multinom is the function to fit a multinomial logistic regression and MaxNWts sets the maximum allowable number of weights high enough for such a large regression data (regression model with ca. 100 covariates).

Note that all predictions in the initial version of SoilGrids1km were made using regression modelling alone. 3D kriging on a sphere at almost one billion locations (130 million pixels times 6 depths) was beyond our technical capacities in 2013/2014. Efforts to use full 3D regression-kriging to produce the first version of SoilGrids1km were abandoned in response to two main issues. Firstly, the computational load to undertake global kriging was too demanding for the processing resources and time we initially had at our disposal. We are working to both increase our processing power and to make the global kriging algorithms more efficient so we can run them globally for subsequent versions of SoilGrids1km. Secondly, there are very large areas of the world (e.g. Russia, northern Canada) that presently have almost no point profile data. These areas lack a sufficient number and density of point observations to successfully compute residuals, which can then be kriged (otherwise kriging leads to serious artifacts). Since we were unable to produce residuals for large parts of the world, we decided not to try to krige residuals globally at first, at least until we obtain enough new point data to support computing and kriging residuals for all major portions of the globe. A full implementation of the 3D regression-kriging model built for SoilGrids has been run successfully at the continental level in Africa but, for the present (February 2014), we have not been able to apply full 3D regression-kriging globally. As soon as these technical limitations are solved, future versions of SoilGrids1km will likely also include a 3D kriging component.

Quality control

Resulting spatial predictions in SoilGrids1km are evaluated using two groups of methods:

- Cross-validation: We used 5–fold cross-validation to estimate the average mapping accuracy for each target variable. For continuous soil properties, we evaluate the amount of variation explained by the models [45]; and for soil classes we evaluate the map purity (i.e. proportion of observations correctly classified) and kappa statistic.
- Visual checking and overlay analysis: Because there is a large amount of spatial data, we have requested users to visually explore maps and look for artefacts and inconsistencies. Inconsistencies and artefacts in maps can be continuously reported through a Global Soil Information mailing list.

To derive amount of variation explained by the models for numeric variables we first derive Root Mean Square Error [46]:

$$RMSE = \sqrt{\frac{1}{l} \cdot \sum_{i=1}^{l} [\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)]^2} \tag{4}$$

where l is the number of validation points. Amount of the variation explained by the model is then:

$$\Sigma_{\%} = \left[1 - \frac{SSE}{SSTO}\right] = \left[1 - \frac{RMSE^2}{\sigma_z^2}\right] [0 - 100\%] \tag{5}$$

where SSE is the sum of squares for residuals at cross-validation points (i.e. $RMSE^2 \cdot n$), and $SSTO$ is the total sum of squares.

Derivation of secondary soil properties: soil organic carbon stock

The SoilGrids1km output maps can be further used for estimation of secondary soil properties which are typically not measured directly in the field and need to be derived from primary soil properties. For instance, consider estimation of the global carbon stock (in t ha^{-1}). This secondary soil property can be derived from a number of primary soil properties [47]:

$$\mathtt{OCS}\ [\mathrm{kg\,m^{-2}}] = \frac{\mathtt{ORC}}{1000}\ [\mathrm{kg\,kg^{-1}}] \cdot \frac{\mathtt{HOT}}{100}\ [\mathrm{m}] \cdot \mathtt{BLD}\ [\mathrm{kg\,m^{-3}}] \cdot \\ \cdot \frac{100 - \mathtt{CRF}\ [\%]}{100} \tag{6}$$

where OCS is soil organic carbon stock, ORC is soil organic carbon mass fraction in permilles, HOT is horizon thickness in cm, BLD is soil bulk density in kg m^{-3} and CRF is volumetric fraction of coarse fragments (>2 mm) in percent (see also Figure 6).

The propagated error of the soil organic carbon stock (Eq.6) can be estimated using the Taylor series method [48]:

$$\sigma_{\mathtt{OCS}} = \frac{1}{10,000,000} \cdot \mathtt{HOT} \cdot \\ \cdot \left(\mathtt{BLD}^2 \cdot (100 - \mathtt{CRF})^2 \cdot \sigma^2_{\mathtt{ORC}} + \right. \\ + \sigma^2_{\mathtt{BLD}} \cdot (100 - \mathtt{CRF})^2 \cdot \mathtt{ORC}^2 + \\ \left. + \mathtt{BLD}^2 \cdot \sigma^2_{\mathtt{CRF}} \cdot \mathtt{ORC}^2 \right)^{-\frac{1}{2}} \tag{7}$$

where $\sigma_{\mathtt{ORC}}$, $\sigma_{\mathtt{BLD}}$ and $\sigma_{\mathtt{CRF}}$ are standard deviations of the predicted soil organic carbon content, bulk density and coarse fragments, respectively. Note that we first predict OCS values for all depths / horizons, then aggregate values for the whole profile (0–2 m). We further use a map of predicted depth to bedrock to remove all predictions outside the effective soil depth (areas where soil is shallower than 2 m).

A more robust way to estimate the propagated uncertainty of deriving OCS would be to use geostatistical simulations (e.g. derive standard error from a large number of realizations $\gg$100) that incorporate spatial and vertical correlations. Because we are dealing with massive data sets, running geostatistical simulations for millions of pixels was not yet considered as an option.

Software implementation

SoilGrids1km predictions are generated via the GSIF package for R, which makes use of a large number of other basic and contributed packages — gstat, raster, rgdal and other R packages for spatial analysis [49]. GSIF package for R contains most of the functions required to produce SoilGrids, and will remain the main platform in the future to obtain global model parameters and access SoilGrids through an API.

As previously mentioned, the target resolution of SoilGrids1km is relatively coarse, nevertheless, the computational intensity and memory required to produce SoilGrids1km is high: one run of SoilGrids1km takes about 12–16 hours on a 12–core HP Z420 workstation with 64 GiB RAM running on a Windows 7 64-bit system. Note also that since we produce predictions at six depths and uncertainty for each depth, the quantity of GeoTIFF maps produced is in the order of 250 × 912MiB ≈ 250 GiB. To deal with processing such large data sets we used a combination of tiling and parallel processing, as implemented via the snowfall package for R [50], to maximize the CPU usage and minimize the time required to produce predictions.

The spatial prediction process consists of four main steps:

(1) preparation of gridded covariates (principal component analysis),
(2) preparation of point data,
(3) model fitting and
(4) spatial prediction and construction of GeoTiffs.

From the steps listed above, spatial prediction require the longest computing time, which is often in the order of 20 or more hours using the computer specification listed above. As a rule of thumb, we look for mapping frameworks that can generate outputs within 48 hrs. If the whole process from model fitting to prediction and export of maps to GeoTiffs consumes $\gg$48 hrs of computing, we consider the system to be impractical for routine operational use.

Results

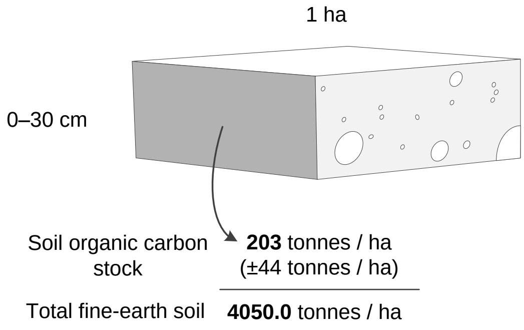
Model fitting

The results of model fitting (Table 1) indicate that the distribution of soil organic carbon content is mainly controlled by climatic conditions, i.e. monthly temperatures and rainfall [51], while the distribution of texture fractions (sand, silt and clay) is mainly controlled by topography and lithology. These key predictors agree with expectations based on existing knowledge. The regression models account for between ca. 20–50 % of observed variability in the target variables (Table 1). Detailed model parameters can be obtained from the SoilGrids1km homepage[7] .

Figure 7 illustrates two examples of spatial predictions for soil organic carbon content and pH. As mentioned previously, soil organic carbon clearly decreases with depth (see also the soil-depth curves shown in Figure 8). Areas mapped as having elevated values of organic carbon are typically associated with cooler and wetter climate regimes and boreal-tundra type vegetation [51,52,53,54]. Note that several soil variables have skewed distributions hence also the output predictions are skewed, so that we use log-transformed legends to maximize contrast in the map (Figure 7).

Figure 8 shows predicted values for organic carbon and pH (mean value and confidence intervals) for the same location

[7] http://soilgrids.org



Bulk density (BLD):	1500 kg / m^3 (s.d. = ±100)
Organic carbon (ORC):	50‰ (s.d. = ±10)
Coarse fragments (CRF):	10% (s.d. = ±5)
Total volume of the block (HOT):	30 cm (· 1 ha)
Soil organic carbon stock (OCS):	203 tonnes / ha (±44)

$$\begin{aligned}\mathbf{OCS} &= \mathrm{ORC}/1000 \cdot \mathrm{BLD} \cdot (100\text{-}\mathrm{CRF})/100 \cdot \mathrm{HOT}/100\\ &= 1/10{,}000{,}000 \cdot \mathrm{ORC} \cdot \mathrm{BLD} \cdot (100\text{-}\mathrm{CRF}) \cdot \mathrm{HOT}\\ &= 1/10{,}000{,}000 \cdot 50 \cdot 1500\ \mathrm{kg}/\mathrm{m}^3 \cdot (100\text{-}10) \cdot 30\ \mathrm{cm}\\ &= 20.25\ \mathrm{kg}/\mathrm{m}^2 = 203\ \text{tonnes / ha}\end{aligned}$$

$$\begin{aligned}\mathbf{OCS.sd} &= 1/10{,}000{,}000 \cdot \mathrm{HOT} \cdot \mathrm{sqrt}(\ \mathrm{BLD}^2 \cdot (100 - \mathrm{CRF})^2 \cdot \mathrm{ORC.sd}^2 +\\ &+ \mathrm{BLD.sd}^2 \cdot (100 - \mathrm{CRF})^2 \cdot \mathrm{ORC}^2 + \mathrm{BLD}^2 \cdot \mathrm{CRF.sd}^2 \cdot \mathrm{ORC}^2\ )\\ &= 4.4\ \mathrm{kg}/\mathrm{m}^2 = 44.1\ \text{tonnes / ha}\end{aligned}$$

Fig. 6. **Soil organic carbon stock calculus scheme**. Example of how total soil organic carbon stock (OCS) and its propagated error can be estimated for a given volume of soil using organic carbon content (ORC), bulk density (BLD), thickness of horizon (HOT), and percentage of coarse fragments (CRF). See text for more detail.

shown in Figure 5. The prediction intervals are rather wide (see also Figure 11), which is connected to the fact that the models explain only 23–51 % of the variation. However, it is important to note that these are global maps of predictions made using relatively coarse resolution covariates. We assume that is unlikely that any effort to map the distribution of soils at a resolution of 1 km could explain a much larger proportion of the total variation in soil properties, as much of this variation occurs over distances less than 1 km [55].

Also note that SoilGrids1km predictions are not capable of representing abrupt changes in values through depth e.g. due to buried horizons, textural heterogeneity or similar. Because we have used linear or close to linear models (plus smoothing splines) to predict values of targeted soil properties and not e.g. regression-trees, these models have smoothed out a significant amount of the variability in the point data, so that it is not realistic to expect abrupt changes in soil properties; at least not vertically (as illustrated previously in Figure 8).

Figure 9 (with a zoom in on Italy) shows that the SoilGrids1km predictions exhibit an order of magnitude greater spatial detail than previous global soil information products e.g. HWSD. This is mainly because a large stack of fine resolution remote sensing based covariate layers has been used to generate SoilGrids1km, and many of these have shown to be significantly correlated with soil properties and classes. Spatial classification accuracy for mapped soil classes, when evaluated using kappa statistics (Table 1), shows a somewhat better match between what was observed on the ground for the USDA classification system (ground-truth classification available for 16,212 profiles) than for the WRB system (classification available for 37,015 profiles).

For many WRB classes our models predicted occurrences in areas that are inconsistent with a strict definition of geographic areas where these classes can occur. The most difficult to map seem to be WRB classes such as Andosols, Solonchaks, Calcisols and Cryosols. These classes are strictly defined (e.g. Andosols are connected with volcanic activities and specific geology) and we need to explore ways to prepare covariates that will prevent prediction of those classes in areas where, by definition, they should not occur. Likewise, USDA suborders are based on soil moisture and climate regimes, for which we did not currently have global covariate maps, and consequently strictly defined classes such as Xerolls (Mollisols in Mediterranean climate; xeric moisture regime) were predicted in Brazil, which probably does not match the definition of the class.

Multinomial logistic regression is a purely data-driven method, so that the overall mapping performance highly depends on representation of environmental conditions by soil samples. All classes that are poorly represented in the environmental space, due to under-sampling, are understandably difficult to map accurately using a purely data-driven model [56]. Nevertheless, the final results of automated extraction of soil classes using multinomial logistic regression are promising, especially for mapping the USDA classes. The mapping accuracy could probably be improved by adding more classification-related covariates and more field observations of soil taxonomy, hopefully through crowd-sourcing, in areas where the accuracy is critically low.

Figure 10 shows derived total soil organic carbon stock based on Eq.(6). According to this map, the total (baseline) amount of soil organic carbon (up to 2 m depth; excluding deserts,

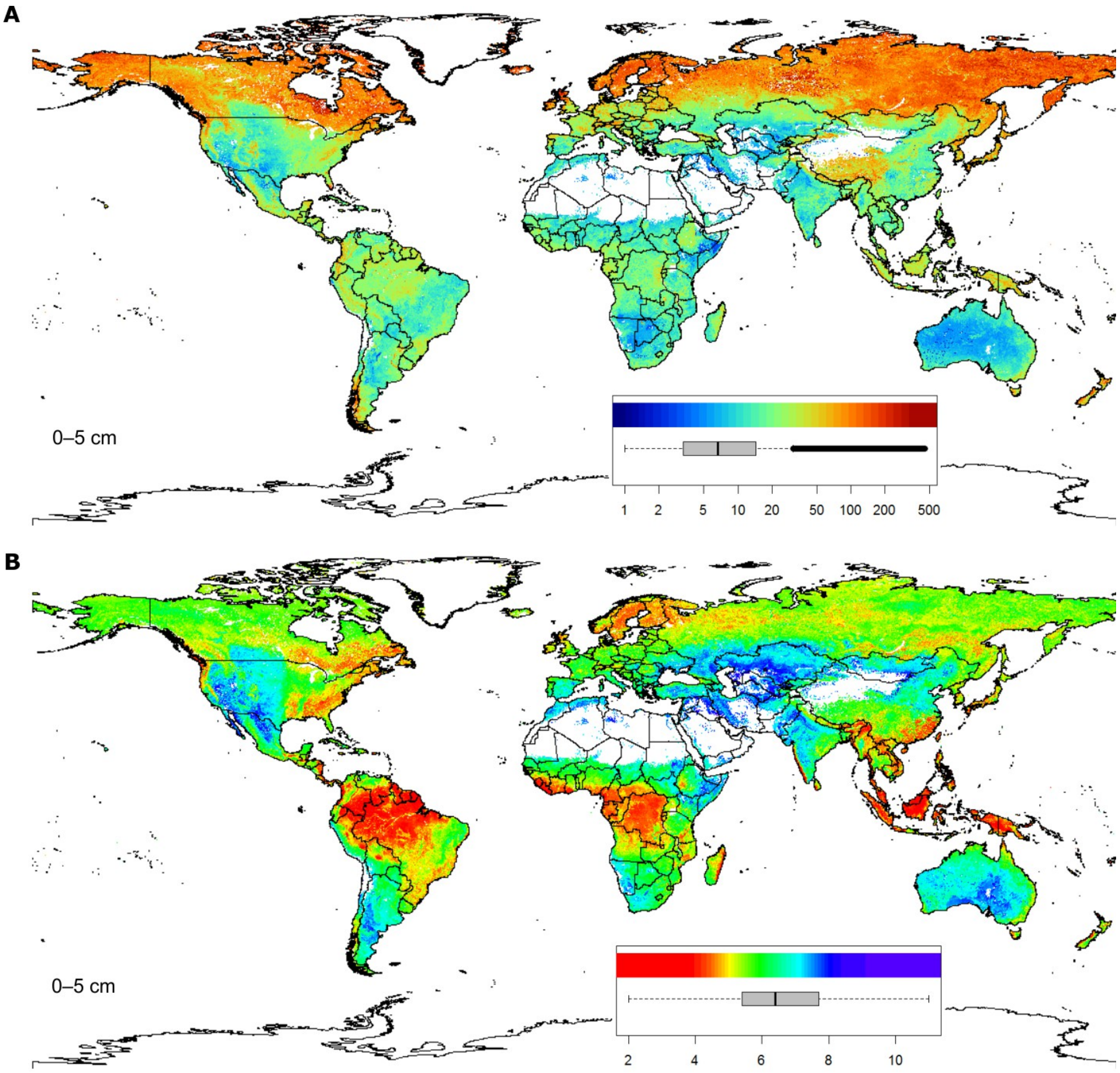


Fig. 7. **Example of SoilGrids1km layers: (A) soil organic carbon content in permille, and (B) soil pH for the topsoil (0–5 centimetres)**. Boxplots show the sampled distribution of the soil property based on the present compilation of global soil profile data.

bare rock areas and ice caps) is about 330 t ha^{-1} on average. The highest concentrations of soil organic carbon are in areas of cooler climate and high rainfall, i.e. northern parts of Canada and Russia seem to be pools for most of the world's soil organic carbon. This largely agrees with results by Hugelius et al. [53] and Scharlemann et al. [57].

The map shown in Figure 10 can be used to supplement maps of total aboveground biomass (see e.g. Ruesch and Gibbs [58] and Scharlemann et al. [57]). Our results also confirm that, overall, the amount of organic carbon below ground is greater than held in biomass above ground [51].

Quality issues

The results of cross-validation are shown in Table 1. The cross-validation results, as expected, largely reflect the model fitting success — properties that can be modeled successfully can also be mapped with higher accuracy. The soil properties that were most difficult to map are soil texture fractions, CEC and WRB soil groups. Although the accuracies of the predictions rarely exceed 50 % of the total variation, all statistical models are significant showing clear spatial patterns (see e.g. Figure 7).

Low cross-validation percentages are common in soil mapping [55,38], i.e. these numbers were not unexpected. Nevertheless, these can be considered promising initial results considering the complexity of harmonization of input point data (see further discussion).

Table 1
Mapping performance of SoilGrids1km — amount of variation explained (from 100 %) or purity/kappa for categorical variables — for eight targeted soil properties and two soil classes distributed via SoilGrids1km. WRB = "World Reference Base"; USDA = "United States Department of Agriculture". Amount of variation explained by the models (Eq.5) i.e. kappa statistics for soil types was determined using 5–fold cross-validation.

Variable name	Type	GSIF code	Units	Range (observed)	Amount of var. explained
Soil organic carbon (dry combustion)	3D	ORCDRC	$g\,kg^{-1}$	0–450	22.9 %
pH index (H_2O solution)	3D	PHIH5X	10^{-1}	2.1–11.0	50.5 %
Sand content (gravimetric)	3D	SNDPPT	$kg\,kg^{-1}$	1–94	23.5 %
Silt content (gravimetric)	3D	SLTPPT	$kg\,kg^{-1}$	2–74	34.9 %
Clay content (gravimetric)	3D	CLYPPT	$kg\,kg^{-1}$	2–68	24.4 %
Coarse fragments (volumetric)	3D	GRAVOL	$cm^3\,cm^{-3}$	0–89	-
Bulk density (fine earth fraction)	3D	BLDVOL	$kg\,m^{-3}$	250–2870	31.8 %
Cation-exchange capacity (fine earth fraction)	3D	CEC	cmol+/kg	0–234	29.4 %
Depth to bedrock	2D	DBR	cm	0–240	-
Soil group (WRB taxonomy)	2D	TAXGWRB	-	-	28.1 % (kappa)
Soil suborder (USDA taxonomy)	2D	TAXOKST	-	-	40.3 % (kappa)

Based on the feedback we received to date from users visiting the project homepage[8], the main limitations of SoilGrids1km are:

(1) problems arising from poor relationships between covariates and dependent variables e.g. covariates can only explain part of the variability, which could possibly improved by using more sophisticated statistical models;
(2) problems arising from high spatial clustering of sampling locations (see Figure 2; observations are too sparse to improve on the regression using a kriging step);
(3) problems associated with using partially-harmonized soil profile data;
(4) problems arising from use of HWSD soil mapping units that are of too coarse scale and often not completely harmonized so that the country borders are still visible (obvious artefact);
(5) limitations in the usability of SoilGrids1km for spatial planning at county or farm scale due to coarse resolution of the maps;
(6) inability to consider and model significant sources of variability e.g. temporal variability due to changes in land use and/or land cover [59];
(7) limitations arising from insufficient use of higher quality and finer resolution conventional soil maps prepared at national to regional scales.

[8] http://soilgrids.org

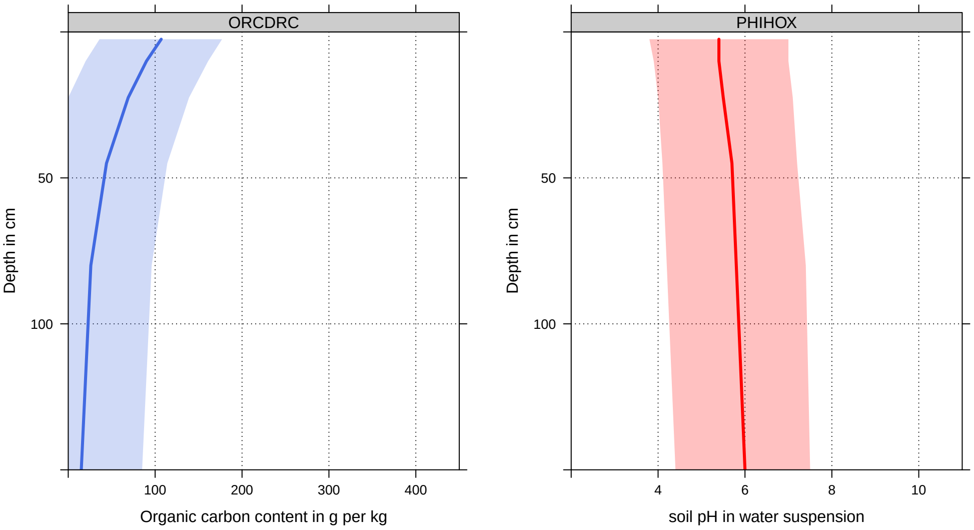


Fig. 8. **SoilGrids1km-derived soil-depth curves for the profile shown in Figure 5**. Location of the profile: 6.3831°E, 50.479167°N. The shaded background indicates the 90 % prediction interval for each depth. ORCDRC = soil organic carbon content in permilles; PHIHOX = soil pH in water suspension. See also Table 1.

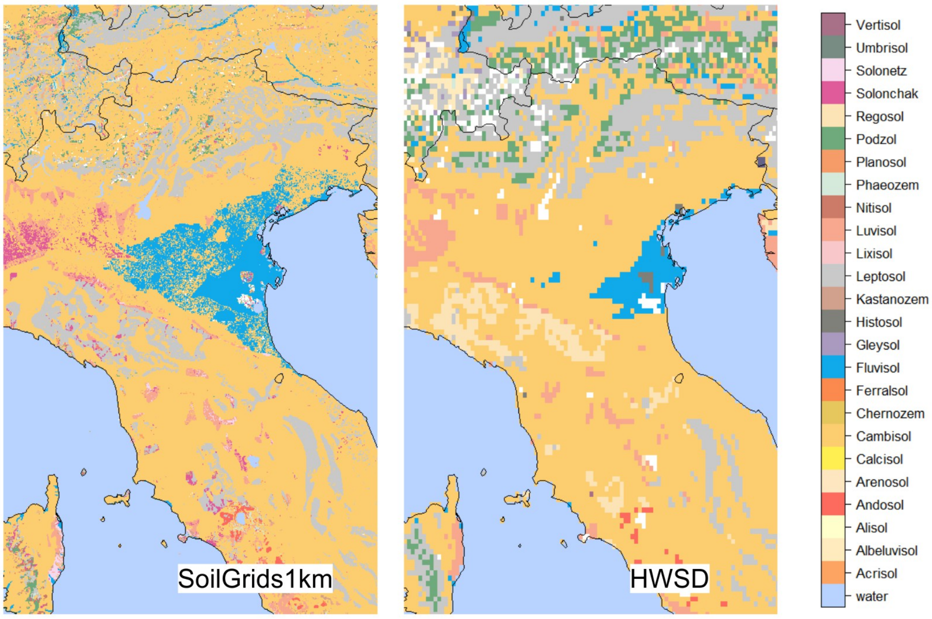


Fig. 9. **Spatial predictions of WRB soil groups for SoilGrids1km (left) and HWSD data set representing conventional soil maps (right)**. A zoom in on North of Italy. White pixels indicate missing values.

Discussion

SoilGrids1km were released on December 5th 2013 (World Soil Day) at the FAO Rome, as a proposed contribution of the Netherlands to the Global Soil Partnership [60]. The system, at the moment, includes predicted values for (Table 1): soil organic carbon ($\mathrm{g\,kg^{-1}}$), soil pH, sand, silt and clay fractions (%), bulk density ($\mathrm{kg\,m^{-3}}$), cation-exchange capacity (cmol+/kg) of the fine earth fraction, coarse fragments (%), soil organic carbon stock ($\mathrm{t\,ha^{-1}}$), depth to bedrock (in cm; see Figure 4), World Reference Base soil groups [43], and USDA Soil Taxonomy suborders [44]. We focussed on generating spatial predictions at six standard depths (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm and 100–200 cm), for which spatially distributed estimates of upper and lower level 90 % prediction intervals are presented. As such, we follow the corresponding specifications of the GlobalSoilMap project [3].

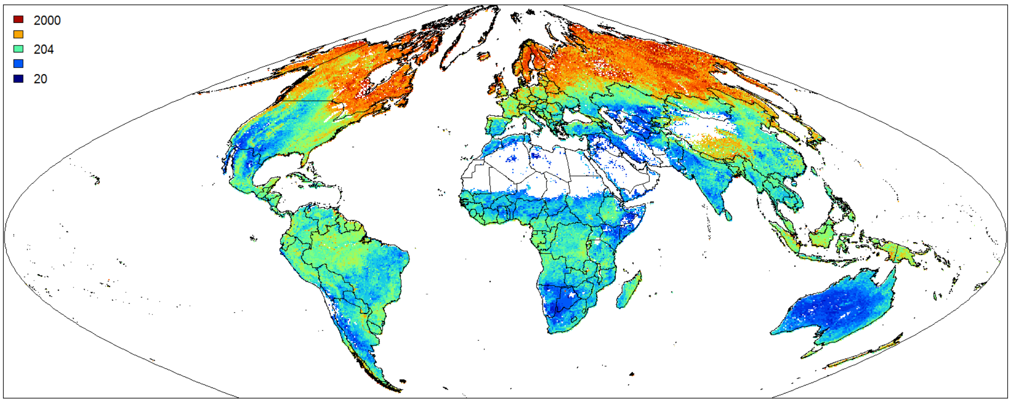


Fig. 10. **Predicted global distribution of the soil organic carbon stock in tonnes per ha for 0–200 centimetres**. Total soil organic carbon stock (here displayed on a log-scale) was estimated as a sum of soil organic carbon stocks for six standard depths and adjusted for the depth to bedrock. Projected in the Sinusoidal equal area projection to give a realistic presentation of areas. Vast deserts (e.g. Sahara or Gobi) can be assumed to contain close to zero organic carbon stock. See also Figure 11.

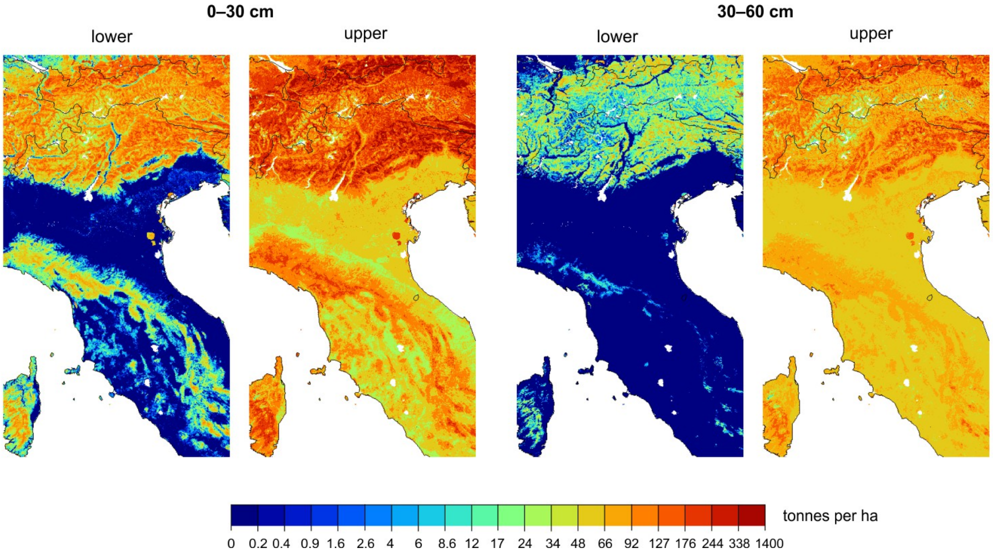
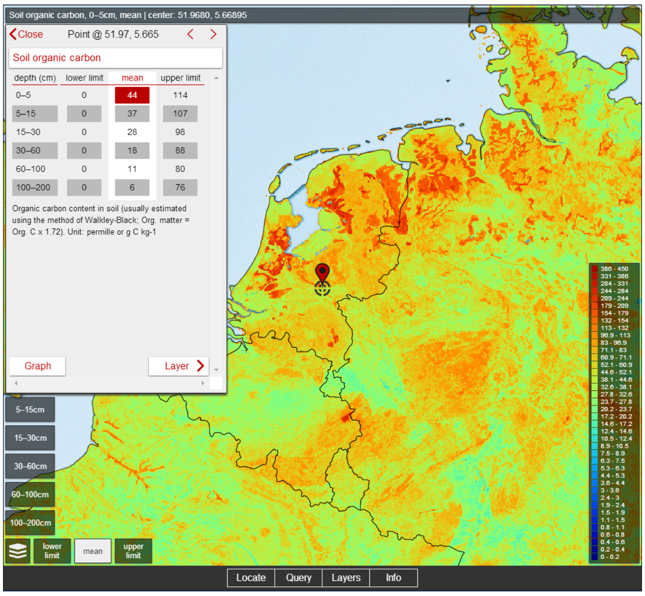


Fig. 11. **Lower and upper confidence limits** (90 % **probability) of estimated soil organic carbon stock (tonnes per ha) for standard depths 0–30 and 30–60 centimeters for the same area as shown in Figure 9**. Derived using the procedure explained in Figure 6.

Initial predictions of soil classes were made at higher (more general) taxonomic levels for both WRB (soil groups) and Soil Taxonomy (suborders). This was done because the available point profile data sets do not provide a sufficient number of locations representative of all of the lower levels of classification in each system. Without a sufficient number of examples for all lower classes, distributed fully across all of the feature space within which each class can occur, it is not possible to successfully predict many of the lower classes defined for either system. Once we have more point observations that encompass the full range of lower level classes across the entire environmental and geographic spectrum of their distribution, we will be able to predict at a more detailed taxonomic level for both classification systems.

The main purpose of SoilGrids1km is to provide initial, fully worked, examples of how complete and consistent global maps of soil properties, and soil classes, can be produced using currently available legacy soil profile data, freely available gridded maps of global covariates and an on-line automated soil mapping system (GSIF). Additionally, we want to use these initial example maps to implement and demonstrate procedures and systems for supporting free and unrestricted access to what we consider to be the best possible current, globally-complete, estimates of soil properties and soil classes. It is hoped that the production, distribution and use of these new, initial, global soil maps will stimulate additional efforts to both improve these maps and to launch new efforts to collect and use new soils information in new soil mapping and monitoring projects. We especially aim at supporting countries in Africa, and large parts of Asia and Latin America, that often have limited infrastructures to produce soil information at fine resolution [5,2]. We think that there is a great potential in using the existing field observa-



Soil organic carbon, 0–5cm, mean | center: 51.9680, 5.66895

Point @ 51.97, 5.665

Soil organic carbon

depth (cm)	lower limit	mean	upper limit
0–5	0	44	114
5–15	0	37	107
15–30	0	28	98
30–60	0	18	88
60–100	0	11	80
100–200	0	6	76

Organic carbon content in soil (usually estimated using the method of Walkley-Black; Org. matter = Org. C x 1.72). Unit: permille or g C kg-1

Fig. 12. **Accessing SoilGrids1km from the SoilInfo app for mobile devices**.

tions and Open Source software to map spatial and spatio-temporal patterns, i.e. without doing any major financial investments.

A number of legitimate concerns exist relative to the initial SoilGrids1km outputs. Probably the most immediate and significant concern has to do with the accuracy and usability of the initial predictions of soil property and class values. We acknowledge that the accuracy of these initial predictions rarely exceeds 50 % of the total variation and, for many properties, is often closer to 20–30 % (Table 1). The results of cross-validation are informative but need to be taken with caution because most of the soil profiles (Figure 2) were not collected using probability sampling, so that the cross-validation results possibly carry the same sampling bias as the original data [61]. Also note that the accuracy of mapping WRB groups is likely lower than the accuracy of mapping USDA soil suborders because over 40 % of the soil profiles that were used for the WRB classification were actually classes translated from national systems. Translation i.e. harmonization of international soil records probably introduces additional noise that can not be solved by regression modelling.

We argue that it is unreasonable to expect any global map of variation in soil properties to explain much more than 50 % of the total observed variation. It is well known that a significant proportion of spatial variation in soil properties occurs over relatively short distances of metres to tens of metres [55,56]. It is therefore unreasonable to expect that a map of global variation in soil properties, portrayed at a spatial resolution of 1 km, will be able to capture and portray the 50 % or more of total variation that occurs at resolutions shorter than 1000 m. Our hope and plan is to gradually improve the accuracy of the predictions by addressing these issues and concerns one by one, in a systematic way (Figure 13). This should be done primarily by working with national and regional soil data agencies, i.e. by adding additional covariates at increasingly finer spatial resolutions and by adding more field/point data from areas that are under-represented.

Although millions of soil profile records have undoubtedly been collected throughout the world, they are often unequally distributed (Figure 2). Likewise, many soil profiles funded by public money are not publicly available or are available in paper format only. Due to unbalanced representation and spatial clustering, predictions in the current version of SoilGrids1km are largely controlled by point data sets available for the USA and Europe. Most of these are from agricultural soils, which inflicts additional bias. Our predictions are therefore likely to exhibit lower accuracy for poorly represented areas such as most of the former Russian Federation, the northern Circumpolar Region, semi-arid and arid areas.

We have also purposely excluded all areas that show no evidence of historical vegetative cover. Our predictions are hence not globally complete. This is a definite drawback for use in global modelling and we acknowledge a need to use either expert judgement or data from other mapping sources to provide alternative predictions for areas with missing values. Again, for deserts and bare rock areas it is perfectly valid to assume a 0 value for soil organic carbon, but it is not as straightforward to estimate soil pH for shifting sand areas for example. For the present, we argue that it is inappropriate to try to make predictions for areas that completely lack vegetative cover e.g. shifting sands of Sahara. These areas have very few to zero point profile observations which can be used to calibrate statistical prediction models. In addition, even if they did have a sufficient number of point profile measurements, the environments of extreme climatic conditions are so different from vegetated ones so that any prediction model is likely to be very different from ones we develop for vegetated areas. We recommend that SoilGrids1km users who require values for the complete land mask fill in the gaps by using expert knowledge or best regional estimates as available from conventional soil mapping (e.g. HWSD, ISRIC-WISE).

It is worth emphasizing that we designed GSIF as a flexible framework with respect to the choice of depths, dimensions (2D or 3D spatial predictions), spatial support size, soil properties and classes and prediction models. Outputs from GSIF are reproducible as a result of use of scripting. Consequently, all maps can be easily updated as new inputs (point and covariate data) become available. We used the GSIF system to generate SoilGrids1km maps for the standard depths defined by the GlobalSoilMap project, but basically one could use the same system for any depth and also for any new property. GSIF is therefore scalable and can be used to produce spatial predictions for virtually any soil property, at any depth and at any spatial or temporal resolution. This, of course, assumes the existence of a sufficient number of point soil observations of appropriate quality and

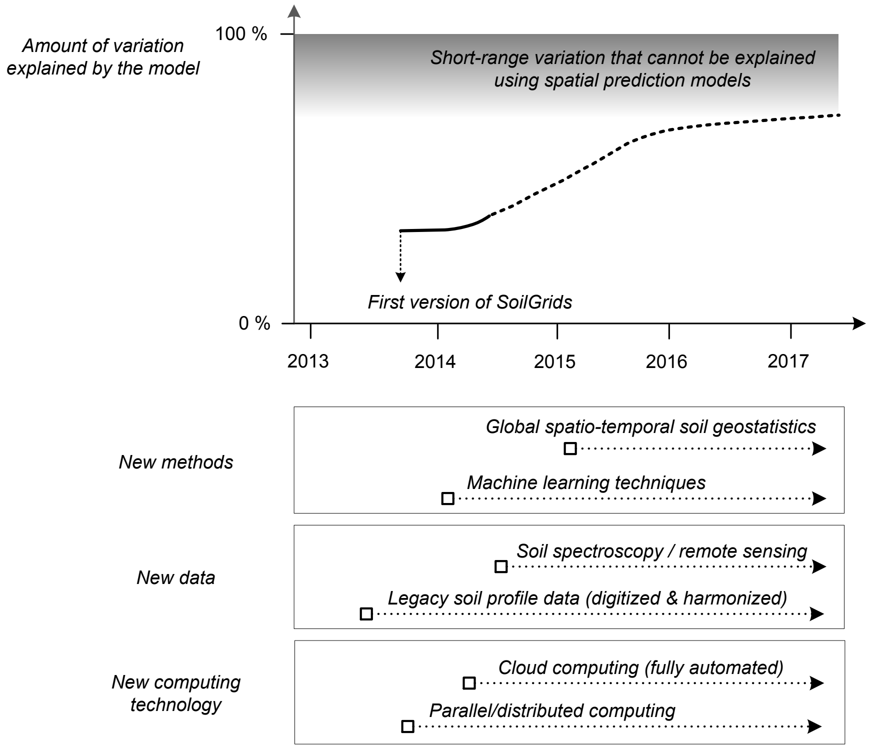

Amount of variation explained by the model
100 %
0 %
Short-range variation that cannot be explained using spatial prediction models
First version of SoilGrids
2013
2014
2015
2016
2017
New methods
Global spatio-temporal soil geostatistics
Machine learning techniques
New data
Soil spectroscopy / remote sensing
Legacy soil profile data (digitized & harmonized)
New computing technology
Cloud computing (fully automated)
Parallel/distributed computing

Fig. 13. **Projected evolution of SoilGrids in the years to come**. We anticipate that the main drivers of success of SoilGrids will be use of machine learning methods for model fitting, development of spatio-temporal geostatistical models, use of new sources of field and remote sensing data and use of faster and more powerful computing capacities. Amount of variation explained by these models will eventually reach a *'natural limit'* (short-range variation that cannot be explained using spatial prediction models), until there is a technological jump in soil remote sensing technology e.g. ground penetrating scanners.

of sufficient covariate layers at sufficiently fine spatial resolution to support modelling at a given spatial resolution.

All methods and models fitted for the purpose of producing SoilGrids1km are available via an Open Source platform (GSIF package for R) and could be adapted for both regional and local mapping. As with input data, the models used to make predictions in GSIF can be improved or replaced in subsequent iterations once better performing models are identified. Prediction models that could be considered in the future include those based on hierarchical Bayes models, regression trees, Random Forests and other machine learning techniques. Regression-trees and similar models could help improve modelling of abrupt changes in values vertically, and Random Forests could help emphasize relative importance of specific covariates. The actual modelling approach used to produce any set of predictions will be reviewed continuously to identify and apply the approach that produces the most correct, consistent and usable outputs.

Because the SoilGrids1km maps can be easily updated (or changed) the process used to produce the map (i.e. SoilGrids system) becomes more important than the map itself. Previously, the map product was seen as more important than the process used to produce it, because any map had to be considered as valid and useful for an extended period, as it took so long, and cost so much, to revise or update the map. Under the GSIF model, the final (or most current) map is no longer the most important output and any system that only provides a final map is considered deficient. We hence argue that it is more important to provide access to all data and models needed to produce (and re-produce) the map than to simply provide the final map itself.

In the future, we hope that GSIF will be used by an increasing number and variety of interested parties, including national and regional soil mapping agencies, commercial consulting agencies, advocacy groups and non-governmental organizations. We envisage GSIF as a platform for cooperation, collaboration, innovation and sharing. It will become so if interested parties decide to participate and contribute as committed partners. The number of soil profiles freely shared by the soil science community is constantly growing and national agencies and other data providers are encouraged to contribute their point data to help improve the prediction accuracy locally for specific countries / regions, for the benefit of the global user community and in support of the global UN conventions.

SoilGrids1km are available for download under a Creative Commons non-Commercial license. SoilGrids1km are also accessible via a Representational State Transfer service[9] and via a mobile phone app “SoilInfo App”[10] (Figure 12).

Acknowledgments

SoilGrids were developed as a part of the Global Soil Information Facilities (GSIF) initiative — tools for collating and serving global soil data developed jointly by ISRIC — World Soil Information and collaborators. ISRIC is a non-profit organization primarily funded by the Dutch government. SoilGrids1km were possible due to generous contributions of point and covariate data by various national and international agencies and organizations to which we are sincerely grateful: USDA National Resources Conservation Service, European Commission’s Joint Research Centre, Africa Soil Information Service project, Mexican Instituto Nacional de Estadística y Geografía / CONABIO, University of São Paulo, Agriculture and Agri-Food Canada, Institute of Soil Science, Chinese Academy of Sciences, Tehran University (National Database of Iran), and others. We are especially thankful for support from the Africa Soil Information Service (AfSIS) project, funded by the Bill and Melinda Gates foundation and the Alliance for a Green Revolution in Africa (AGRA), and to the Food and Agriculture Organization (FAO) of the United Nations for supporting the international soil standards and initiatives developed at ISRIC (SOTER, GSIF). We gratefully acknowledge the significant contributions of Dr. H.I. Reuter, formerly with ISRIC, presently statistical officer at EUROSTAT, in developing the initial concepts and design for GSIF, in helping to conceptualise and assemble the global maps of covariates and in implementing some of the early algorithms.

References

1. Sanchez et al (2009) Digital Soil Map of the World. Science 325: 680-681.
2. Omuto C, Nachtergaele F, Vargas Rojas R (2012) State of the Art Report on Global and Regional Soil Information: Where are we? Where to go? Global Soil Partnership technical report. Rome: FAO, 66 pp.
3. Arrouays D, Grundy MG, Hartemink AE, Hempel JW, Heuvelink GB, et al. (2014) Chapter Three — GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. In: Sparks DL, editor, Soil carbon, Academic Press, volume 125 of *Advances in Agronomy*. pp. 93 - 134. doi:10.1016/B978-0-12-800137-0.00003-0.
4. Viscarra Rossel RA, Webster R, Bui EN, Baldock JA (2014) Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Global Change Biology : n/a–n/a.
5. Grunwald S, Thompson JA, Boettinger JL (2011) Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. Soil Science Society of America Journal 75: 1201–1213.
6. Savtchenko A, Ouzounov D, Ahmad S, Acker J, Leptoukh G, et al. (2004) Terra and Aqua MODIS products available from NASA GES DAAC. Advances in Space Research 34: 710-714.
7. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, et al. (2008) Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. PLoS One 3: e1408.
8. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, et al. (2013) High-resolution global maps of 21st-century forest cover change. Science 342: 850–853.
9. González JH, Bachmann M, Krieger G, Fiedler H (2010) Development of the TanDEM-X calibration concept: analysis of systematic errors. Geoscience and Remote Sensing, IEEE Transactions on 48: 716–726.
10. Hartemink AE, Krasilnikov P, Bockheim J (2013) Soil maps of the world. Geoderma 207/208: 256–267.
11. FAO/IIASA/ISRIC/ISS-CAS/JRC (2012) Harmonized World Soil Database (version 1.2). Rome: FAO.
12. Batjes NH (2012) ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2). Report 2012/01. Wageningen: ISRIC — World Soil Information, 57 pp.
13. Diggle PJ, Ribeiro Jr PJ (2007) Model-based Geostatistics. Springer Series in Statistics. Springer, 288 pp.
14. Brown PE (2014) Model-Based Geostatistics the Easy Way. Journal of Statistical Software ??: ??
15. Pebesma E, Cornford D, Dubois G, Heuvelink GB, Hristopulos D, et al. (2011) Intamap: The design and implementation of an interoperable automated interpolation web service. Computers & Geosciences 37: 343 - 352.
16. Fritz S, McCallum I, Schill C, Perger C, See L, et al. (2012) Geo-Wiki: An online platform for improving global land cover. Environmental Modelling & Software 31: 110–123.
17. R Development Core Team (2009) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 409 pp. ISBN 3-900051-07-0.
18. Shore J (2007) The Art of Agile Development. Theory in practice. O’Reilly Media, 440 pp.
19. Tóth G, Jones A, Montanarella L, editors (2013) LUCAS Topsoil Survey. Methodology, data and results. JRC Technical Reports EUR 26102. Luxembourg: Publications Office of the European Union.
20. Leenaars J (2012) Africa Soil Profiles Database, Version 1.0. A compilation of geo-referenced and standardized legacy soil profile data for Sub Saharan Africa (with dataset). Wageningen, the Netherlands: Africa Soil Information Service (AfSIS) project and ISRIC — World Soil Information, 45 pp. ISRIC report 2012/03.
21. Instituto Nacional de Estadística y Geografía (INEGI) (2000) Conjunto de Datos de Perfiles de Suelos, Escala 1: 250 000 Serie II. (Continuo Nacional). Aguascalientes, Ags. México: INEGI.
22. Cooper M, Mendes LMS, Silva WLC, Sparovek G (2005) A national soil profile database for brazil available to international scientists. Soil Science Society of America Journal 69: 649–652.
23. Shangguan W, Dai Y, Liu B, Zhu A, Duan Q, et al. (2013) A China data set of soil properties for land surface modeling. Journal of Advances in Modeling Earth Systems 5: 212–224.

[9] http://rest.soilgrids.org
[10] http://soilinfo-app.org

24. MacDonald KB, Valentine KWG (1992) CanSIS/NSDB. A general description. Ottawa: Centre for Land and Biological Resources Research, Research Branch, Agriculture Canada.

25. Batjes NH (2009) Harmonized soil profile data for applications at global and continental scales: Updates to the WISE database. Soil Use and Management 25: 124-127.

26. Van Engelen V, Dijkshoorn J, editors (2012) Global and National Soils and Terrain Digital Databases (SOTER), Procedures Manual, version 2.0. ISRIC Report 2012/04. Wageningen, the Netherlands: ISRIC - World Soil Information, 192 pp.

27. Hollis JM, Jones RJA, Marshall CJ, Holden A, Van de Veen JR, et al. (2006) SPADE-2: The soil profile analytical database for Europe, version 1.0. Luxembourg: Office for official publications of the European Communities. EUR22127EN.

28. Stolbovoi V, McCallum I (2002) Land Resources of Russia (CD-ROM). Vienna: IIASA and RAS.

29. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

30. Hartmann J, Moosdorf N (2012) The new global lithological map database GLiM: A representation of rock properties at the Earth surface. Geochemistry, Geophysics, Geosystems 13: n/a–n/a.

31. Carroll M, Townshend J, DiMiceli C, Noojipady P, Sohlberg RA (2009) A new global raster water mask at 250 m resolution. International Journal of Digital Earth 2: 291-308.

32. Odeh I, McBratney AB, Chittleborough D (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67: 215–226.

33. Hengl T, Heuvelink G, Rossiter DG (2007) About regression-kriging: from equations to case studies. Computers & Geosciences 33: 1301–1315.

34. Kutner M, Neter J, Nachtsheim C, Li W (2005) Applied Linear Statistical Models. Operations and decision sciences series. McGraw-Hill Irwin.

35. Hastie TJ (1992) Statistical Models in S, Wadsworth & Brooks/Cole, chapter Generalized additive models. pp. 249-307.

36. Venables WN, Ripley BD (2002) Modern applied statistics with S. New York: Springer-Verlag, 4th edition, 481 pp.

37. Agarwal DK, Gelfand AE, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. Environmental and Ecological statistics 9: 341–355.

38. McBratney AB, Minasny B, MacMillan RA, Carré F (2011) Digital soil mapping. In: Li H, Sumner M, editors, Handbook of Soil Science, CRC Press, volume 37. pp. 1-45.

39. Mokma DL, Buurman P (1982) Podzols and podzolization in temperate regions. ISM monograph. Wageningen: International Soil Museum, 126 pp.

40. Minasny B, McBratney AB (2010) Methodologies for Global Soil Mapping. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, editors, Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, Springer, volume 2 of *Progress in Soil Science*, chapter 34. pp. 429–436.

41. Jenny H (1994) Factors of soil formation: a system of quantitative pedology. Dover books on Earth sciences. Dover Publications.

42. Bouma J, Batjes NH, Groot JJR (1998) Exploring land quality effects on world food supply. Geoderma 86: 43–59.

43. IUSS Working Group WRB (2006) World reference base for soil resources 2006: a framework for international classification, correlation and communication. World soil resources reports No. 103. Rome: Food and Agriculture Organization of the United Nations.

44. US Department of Agriculture (2010) Keys to Soil Taxonomy. U.S. Government Printing Office, 11th edition.

45. Hengl T, Nikolić M, MacMillan RA (2013) Mapping efficiency and information content. International Journal of Applied Earth Observation and Geoinformation 22: 127-138.

46. Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103: 3–26.

47. Nelson DW, Sommers L (1982) Total carbon, organic carbon, and organic matter. In: Page A, Miller R, Keeney D, editors, Methods of soil analysis, Part 2, Madison, WI: ASA and SSSA, Agron. Monogr. 9. 2nd edition, pp. 539–579.

48. Heuvelink G (1998) Error propagation in environmental modelling with GIS. London, UK: Taylor & Francis.

49. Bivand R, Pebesma E, Rubio V (2013) Applied Spatial Data Analysis with R. Use R Series. Heidelberg: Springer, 2nd edition, 401 pp.

50. Knaus J, Porzelius C, Binder H, Schwarzer G (2009) Easier parallel computing in R with snowfall and sfCluster. The R Journal 1: 54–59.

51. Jobbágy EG, Jackson RB (2000) The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecological applications 10: 423–436.

52. Todd-Brown KEO, Randerson JT, Post WM, Hoffman FM, Tarnocai C, et al. (2013) Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. Biogeosciences 10: 1717–1736.

53. Hugelius G, Strauss J, Zubrzycki S, Harden JW, Schuur E, et al. (2014) Improved estimates show large circumpolar stocks of permafrost carbon while quantifying substantial uncertainty ranges and identifying remaining data gaps. Biogeosciences Discussions 11: 4771–4822.

54. Minasny B, McBratney AB, Malone BP, Lacoste M, Walter C (2014) Quantitatively Predicting Soil Carbon Across Landscapes. In: Hartemink AE, McSweeney K, editors, Soil Carbon, Springer International Publishing, Progress in Soil Science. pp. 45-57. doi: 10.1007/978-3-319-04084-4_5.

55. Heuvelink GBM, Webster R (2001) Modelling soil variation: past, present, and future. Geoderma 100: 269-301.

56. Antonić O, Pernar N, Jelaska SD (2003) Spatial distribution of main forest soil groups in Croatia as a function of basic pedogenetic factors. Ecological modelling 170: 363–371.

57. Scharlemann JPW, Tanner EVJ, Hiederer R, Kapos V (2014) Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Management 5: 81–91.

58. Ruesch A, Gibbs HK (2008) New IPCC Tier-1 global biomass carbon map for the year 2000. Oak Ridge National Laboratory, Tennessee: Carbon Dioxide Information Analysis Center.

59. Verburg PH, Neumann K, Nol L (2011) Challenges in using land use and land cover data for global change studies. Global Change Biology 17: 974–989.

60. Montanarella L, Vargas R (2012) Global governance of soil resources as a necessary condition for sustainable development. Current Opinion in Environmental Sustainability 4: 559–564.

61. Brus D, Kempen B, Heuvelink G (2011) Sampling for validation of digital soil maps. European Journal of Soil Science 62: 394–407.