Specifications
Tiered\textsuperscript{1} \textit{GlobalSoilMap.net} products

Release 2.3
[21/9/2012]\textsuperscript{2}

Science Committee

\textsuperscript{1} Tier 1 – point predictions on 3 arc-second grid; Tier 2 - Version 1 plus block predictions on 3 arc-second grid
\textsuperscript{2} These specifications have been modified as a result of an Uncertainty Workshop held at USDA NRCS, Lincoln, Nebraska, August 27-30, 2012
These specifications are expected to remain valid for a period of 2 years from the date of endorsement

Date of Endorsement of v2.2: February 8, 2012
V2.3: to be endorsed
Table of Contents

List of Figures ........................................................................................................................................ 5
List of Tables ........................................................................................................................................ 5
Summary .................................................................................................................................................. 6
1.1 Definition ....................................................................................................................................... 10
1.2 Vertical Dimension - Depth ........................................................................................................... 10
1.3 Horizontal Dimension - Resolution ............................................................................................... 12
1.4 Geo-referencing ............................................................................................................................. 12
1.5 Excluded Non-soil Areas .................................................................................................................. 13
1.6 Tiling of Grid Cell Data ................................................................................................................... 14
1.7 Unique Numbering of Grid Cell Data .............................................................................................. 15
1.8 Volumetric data: voxels .................................................................................................................. 16
2. Soil Properties ................................................................................................................................... 18
2.1 Depth of Soil ................................................................................................................................. 18
2.2 Primary Soil Properties .................................................................................................................. 18
2.3 Derived Soil Properties .................................................................................................................. 19
2.4 Additional Soil Properties ............................................................................................................. 20
2.5 Time (Year) ...................................................................................................................................... 20
3. Uncertainty ......................................................................................................................................... 21
3.1 Definition ........................................................................................................................................ 21
3.2 Estimating uncertainty for the GlobalSoilMap.net predictions of soil property values .............. Error! Bookmark not defined.
3.3 Validation of GlobalSoilMap.net predictions of soil property values .......................................... 22
3.4 Reproducibility of GlobalSoilMap.net predictions of soil property values ................................. 22
3.5 Policy for release of interim GlobalSoilMap.net predictions of soil property values .................. 22
References ............................................................................................................................................... 23
Appendix A: Minimum data set for each 100 m grid cell ................................................................. 25
Appendix B: Some possible approaches for computing weighted mean value by grid cell ............... 26
Appendix C: Correlations of soil properties derived from different soil analytical methods ........... 28
   C1.0 Rationale ..................................................................................................................................... 28
   C1.1 Organic Carbon ......................................................................................................................... 29
   C1.2 pH ............................................................................................................................................. 32
   C1.3 Particle Size Distribution (sand, silt and clay) ............................................................................ 34
   C1.4 Bulk Density ............................................................................................................................. 37
   C1.5 Available Water Capacity ......................................................................................................... 39
   References for Appendix C ............................................................................................................... 43
Appendix D: Equal Area projections for use by each of the GlobalSoilmap.net nodes .................... 49
   Suggested projections expressed in terms of R-code ....................................................................... 50
Appendix E: Background on uncertainty ............................................................................................ 51
   Understanding Uncertainty ................................................................................................................ Error! Bookmark not defined.
Appendix F: North American node tiering concepts ........................................................................... 53
List of Figures

Figure 1. The tile numbering system used by CIAT for 3 arc-second SRTM DEM data .........................15
Figure 2. Illustration of the approach used to number grid cells when dividing by 2, 3 or 5 ...............16
Figure 3. The primary spatial entity is defined as a square grid cell with a horizontal resolution of 3 arc- seconds by 3 arc-seconds. The soil property value reported will be a bulked average for a 100 m by 100 m grid cell located at the center point of the 3 arc-second grid. ........................................................................................................17
Figure 4. Particle size limits used in European countries, Australia and America (Adapted from Nemes et al., 1999a and Minasny and McBratney, 2001) ........................................................................................................34

List of Tables

Table 1. Depth intervals for which soil property values and uncertainty will be provided .................11
Table 2. Summary of maximum rooting depth by biome (after Canadell et al., 1996) ......................12
Table 3. Hierarchical numbering system used to assign unique IDs to each nested grid cell ..........17
Table 4. Specifications for properties related to reporting depth of soil ........................................18
Table 5. Specifications for primary soil properties ........................................................................18
Table 6. Specifications for derived soil properties .......................................................................19
Table 7. Specifications for secondary soil properties ....................................................................19
Table 8. Illustration of the minimum data required for each grid cell ........................................25
Table 9. Regression equations for harmonizing values of organic carbon to a reference standard ....30
Table 10. Regression equations for harmonizing values of organic carbon to a reference standard ....31
Table 11 Example regression equations for converting values of pH between different methods ....33
Table 12. Differences between the International, USDA and ISO/FAO particle size classifications .......35
Specifications for *GlobalSoilMap.net* products

### Executive Summary

### Summary

This document sets out the specifications for the *GlobalSoilMap.net* project products (tier 1 and beyond). The specifications do not prescribe how the products must be made; only what they need to conform to in order to permit collation and presentation of final standardized products.

The specifications focus on five aspects (Table 0):
1. The spatial entity,
2. The soil properties to be predicted (and the date associated with their prediction)
3. The uncertainties for each soil property
4. The age of the data or information used to estimate the predicted properties
5. The validation measure to be used and reported.

### Table 0

<p>| Agreement on the initial specifications was achieved at the <em>GlobalSoilMap.net</em> node meeting in Seoul, Korea on October 25-26, 2009. |
| Agreement on version (2.1) of the specifications was achieved in June, 2011. |</p>
<table>
<thead>
<tr>
<th>Tier</th>
<th>Spatial entity</th>
<th>Grid</th>
<th>Properties</th>
<th>Uncertainty</th>
<th>Date stamping</th>
<th>Validation measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘point’</td>
<td>3 arc second by 3 arc second by standard depths (Table 1)</td>
<td>Point estimates of all properties in Tables 4, 5 &amp; 6 at all depths</td>
<td>Upper and lower 90% PI for all properties at all depths</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>(includes all Tier 1 product)</td>
<td>100 m by 100m block by depth increment</td>
<td>3 arc second by 3 arc second by standard depths (Table 1)</td>
<td>Block average of all properties in Tables 4, 5 &amp; 6 at all depths</td>
<td>Year of data producing estimates for all properties at all depths</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>(includes all Tier 2 product)</td>
<td></td>
<td></td>
<td>Marginal probability distribution for each and every xyz-point</td>
<td></td>
<td>National/Regional RMSE, etc. for point and block predictions for all properties at all depths, by independent (probability) sampling</td>
</tr>
<tr>
<td>4 [A, AA, AAA]</td>
<td>(includes all Tier 3 product)</td>
<td></td>
<td></td>
<td>Complete spatial-multivariate probability distribution (joint probability distribution for all soil properties and xyz-points)</td>
<td></td>
<td>Thresholds of RMSE and % coverage of true values are met for each prediction</td>
</tr>
</tbody>
</table>

5 accuracy thresholds for each attribute (and each depth) will be specified in tabular form, linking average prediction interval width, and the percentage of independent validation samples that fall within the prediction intervals (for A, AA, AAA etc)
Please also refer to Appendices E, F and G

The primary spatial entity that will be delivered is a single point value representing an area of no more than 2 m by 2 m horizontal dimensions located at the centre point of the same global grid of 3 arc-seconds by 3 arc-seconds. The secondary spatial entity for delivery is a volumetric grid cell of specified horizontal and vertical dimensions and location. Each grid cell represents an area of 100 m by 100 m horizontal dimensions located at the centre point of a global grid of 3 arc-seconds by 3 arc-seconds. The point estimate of the primary spatial entity may be necessary to produce estimates for the secondary spatial entity. In the vertical dimension, predictions of soil property values and their associated uncertainties will be made to 2 m (if possible) with data reported for 6 depth intervals of 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm and 100-200 cm.

Twelve soil properties will be predicted at each location. These are: (1) total profile depth (cm), (2) plant exploitable (effective) soil depth (cm), (3) organic carbon (g/kg), (4) pH (x10), (5) sand (g/kg), (6) silt (g/kg), (7) clay (g/kg), (8) gravel (m$^3$ m$^{-3}$), (9) ECEC (cmolc/kg), (10) bulk density of the fine earth (< 2 mm) fraction (excludes gravel) (Mg/m$^3$), (11) bulk density of the whole soil in situ (includes gravel) (Mg/m$^3$) and (12) available water capacity (mm). Additional soil properties including, for example, EC (dS/m) may be predicted at the discretion of the nodes but these are not mandatory.

Each soil property will have an estimate of the uncertainty associated with the prediction for each depth (for properties reported by depth) for each grid location. Uncertainty here is defined as the 90% prediction interval (PI), which is the range in values within which the true value at any point prediction location is expected to be found 9 times out of 10 (90%). Methods of estimating uncertainty are not specified here but are outlined in separate appendices.

Predictions of both soil property values and their associated uncertainty will be produced and reported at point support (primary product) and subsequently averaged or aggregated to 100 m x 100 m block support (secondary product). Selection of prediction and aggregation methods is the responsibility of each node and is likely to vary from node to node and through time.

Each node will be responsible for fully documenting the procedures used to generate all products they submit. This documentation will be specified as metadata attached to each product. A template for recording metadata about how all outputs are produced will be developed and added as an addendum to these specifications. The metadata documentation will specify all inputs used and the exact procedures applied to the inputs to produce all outputs.
In the longer term, future versions of these specifications will evolve towards defining complete workflows that specify all inputs and procedures required to produce all outputs, in addition to specifying the outputs. Future versions of these specifications will aim at achieving complete reproducibility of outputs by requiring all nodes to completely specify all of their inputs and procedures and have all inputs and procedures available within the node so that outputs can be regenerated or updated automatically. This level of automation and reproducibility is not considered feasible to obtain at present but it is a goal that future tiers of the product will specify complete reproducibility.

Validation

For Tier 1 products for any contiguous region some minimal statement of accuracy is required.

An appropriate measure for each property at each depth increment is the root mean square error of the point predictions. This can be achieved by cross validation for point-based methods, and true validation for soil-class map based methods. In the latter case at least, on average, one observation point per 10 000 square kilometres may be required and probably at least 50 points are required to obtain an estimate.

(In subsequent tiers it is anticipated a richer set of validation criteria will be used, including mean error, and the percentage of the map area that fall within the uncertainty limits. Ultimately, sampling will be required to produce such quality estimates)
1. Spatial entity

1.1 Definition

Two separate spatial entities are defined. Only the primary spatial entity has to be delivered in tier 1 product.

The primary spatial entity is meant to represent point support. It is operationally defined as a volume of soil to a depth of 2 m (or bedrock if bedrock occurs within less than 2 m) for a rectangular area with small and irregular horizontal dimensions of less than 2 m by 2 m (effectively a pedon at point support) located at the centre of a defined global grid of 3 arc-seconds by 3 arc-seconds (approximately 93 m x 93 m at the equator).

The secondary spatial entity for delivery in subsequent tiers is defined as a volume of soil to a depth of 2 m (or bedrock if bedrock occurs within less than 2 m) for a rectangular area with regular, fixed horizontal dimensions of 100 m by 100 m located at the centre of a defined global grid of 3 arc-seconds by 3 arc-seconds (approximately 93 m x 93 m at the equator).

1.2 Vertical Dimension - Depth

The depth of soil for which data will be reported (2 m) reflects arguments presented in the USDA-NRCS Soil Survey Manual (Soil Survey Division Staff, 1993) (Chapter 2 page 3): “for purposes of most soil surveys, a practical lower limit of a pedon is bedrock or a depth of about 2 m, whichever is shallower. A depth of 2 m provides a good sample of major soil horizons, even in thick soil. It includes much of the volume of soil penetrated by plant roots, and it permits reliable observations of soil properties”.

Soil survey reports commonly do not report soil property data for depths greater than 1.0 – 1.5 m. Therefore, it will often be necessary to infer values below 1.0 m and estimate the values for uncertainty of predictions for soil properties at lower depths. It is recognized that uncertainty will increase with depth and that uncertainty will be high for greater depths for most areas.

Depth is measured from the soil surface. For mineral soils, the soil surface is the top of the mineral soil. For organic soils (or mineral soils with an O horizon), the top of any surface horizon identified as an O horizon is considered the soil surface. The soil surface is the top of the part of the O horizon that is at least slightly decomposed. Fresh leaf or needle fall that has not undergone observable decomposition is excluded when determining soil depth. For soils with a cover of 80 percent or more rock fragments on the surface, the depth is measured from the surface of the rock fragments. (Soil Survey Division Staff, 1993: Chapter 3 page 4).
A value will be predicted for 12 soil properties, and for the uncertainty associated with this prediction, for six fixed depth intervals (Table 1). The values reported for these six depth intervals act as coefficients for a spline function that will provide a mechanism for reporting continuous variation with depth for all properties at all grid cells.

In addition, a value, and associated uncertainty, will be predicted for each of depth to bedrock or consolidated material and for plant extractable (effective) soil depth (depth to restricting layer).

Table 1. Depth intervals for which soil property values and uncertainty will be provided

<table>
<thead>
<tr>
<th>No.</th>
<th>Depth Interval</th>
<th>Lower 2.5 Percentile of mean</th>
<th>Estimated Value of Soil Property</th>
<th>Upper 97.5 Percentile of mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 - 5 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5 – 15 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15 – 30 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>30 – 60 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>60-100 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>100-200 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Depth to rock</td>
<td></td>
<td>Depth to rock in cm</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Effective Depth</td>
<td></td>
<td>Effective Depth in cm</td>
<td></td>
</tr>
</tbody>
</table>

The value reported for each depth interval for each soil property will represent the mean value of that property over the specified depth interval within an area of 100 m x 100 m horizontal extent centred at each 3 arc-second by 3 arc-second grid cell. The soil properties to be predicted are listed and described in section 2.

Total profile depth is depth to a lithic or paralithic contact in cm as defined below.

Depth to bedrock.—This refers to the depth to fixed rock. Hard and soft bedrock are distinguished. Hard bedrock is usually indurated but may be strongly cemented, and excavation difficulty would be very high or higher. Soft bedrock meets the consistence requirements for paralithic contact (Soil Survey Division Staff, 1993, Chapter 6 page 13).

Plant Exploitable (Effective) Depth is defined as: “The lower limit of soil is normally the lower limit of biologic activity, which generally coincides with the common rooting depth of native perennial plants” (Soil Survey Staff, 1975; Soil Survey Division Staff, 1993, Chapter 1 page 5).

We can define root depth by either the evidence of the roots themselves, or on the presence of barriers to root extension. The first option requires rules for root abundance to define the lower limit, or inferences on the depth of native roots from soil morphology. Depths may differ between biomes as given in Table 2. Although logical, the approach is complex. GSM soil specifications need to be capable of consistent application across the globe, and not reliant on
complex accessory data. The second option defines the depth of a relatively-easy-rooting zone from the soil surface to a root boundary. The boundary is defined by one or more morphological barriers. These barriers include: massive rock, jointed rock, clean sand, pan, high density material (bulk density >1.85), extremely gravelly or densely packed gravel, and chemical toxicity.

The extremely deep rooting ability of some tree species in arid land, where roots penetrate to great depth in jointed rock is noted and would not be recognised in option two. The choice of option needs to consider the application of the data. If it is to map ecosystem behaviour across different biomes then option one is favoured. If it is to explore opportunities for regional or global food production then the agronomic depth provided on option 2 is favoured. The second option is proposed for these specifications because it is more likely to provide a consistent global soil information product and is likely to be more widely used. Where option one is required it is suggested that it be mapped as an additional layer by nodes.

Table 2. Summary of maximum rooting depth by biome (after Canadell et al., 1996)

<table>
<thead>
<tr>
<th>Biome</th>
<th>N</th>
<th>Mean maximum rooting depth (m)</th>
<th>Highest value for rooting depth (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boreal Forest</td>
<td>6</td>
<td>2.0 ± 0.3</td>
<td>3.3</td>
</tr>
<tr>
<td>Cropland</td>
<td>17</td>
<td>2.1 ± 0.2</td>
<td>3.7</td>
</tr>
<tr>
<td>Desert</td>
<td>22</td>
<td>9.5 ± 2.4</td>
<td>53</td>
</tr>
<tr>
<td>Sclerophyllous shrubland and forest</td>
<td>57</td>
<td>5.2 ± 0.8</td>
<td>40</td>
</tr>
<tr>
<td>Temperate coniferous forest</td>
<td>17</td>
<td>3.9 ± 0.4</td>
<td>7.5</td>
</tr>
<tr>
<td>Temperate deciduous forest</td>
<td>19</td>
<td>2.9 ± 0.2</td>
<td>4.4</td>
</tr>
<tr>
<td>Temperate grassland</td>
<td>82</td>
<td>2.6 ± 0.2</td>
<td>6.3</td>
</tr>
<tr>
<td>Tropical deciduous forest</td>
<td>5</td>
<td>3.7 ± 0.5</td>
<td>4.7</td>
</tr>
<tr>
<td>Tropical evergreen forest</td>
<td>5</td>
<td>7.3 ± 2.8</td>
<td>18</td>
</tr>
<tr>
<td>Tropical savanna</td>
<td>15</td>
<td>15.0 ± 5.4</td>
<td>68</td>
</tr>
<tr>
<td>Tundra</td>
<td>8</td>
<td>0.5 ± 0.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

1.3 Horizontal Dimension - Resolution.

The primary spatial entity is located at the nodes of the 3 arc-second by 3 arc-second grid exactly matching the global SRTM DEM data set and extended to the poles. (The global SRTM DEM data set provides the source for several key covariates used in predicting soil properties.) The secondary spatial entity will have regular, equal horizontal dimensions for an area of 100 m by 100 m located at the centre of a defined global grid of 3 arc-seconds by 3 arc-seconds. Adopting a fixed grid resolution of 100 m by 100 m to define the horizontal extent of the reporting area for each block estimate means that all values for the entire world will describe an area of equal extent. Storing the value for this area estimate at the centre point of a global 3 by 3 arc-second grid facilitates global data compilation.

1.4 Geo-referencing
The spatial location of each grid cell will be given in geographic coordinates (lat/long) using WGS84 for the horizontal datum and EGM96 for the vertical datum. The following georeferencing information will apply for each grid point location:

1. Projection: Geographic  
2. Horizontal Datum: WGS84  
3. Vertical Datum: EGM96  
4. Latitude: reported in decimal degrees  
5. Longitude: reported in decimal degrees  
6. Date associated with the value estimate: Year

Only final grid maps of predicted soil properties need to be delivered in geographic coordinates at a resolution of 3 arc-seconds. Use of geographic coordinates is specified in order to facilitate seamless compilation of global data sets, with no gaps, offsets, duplication or edge matching issues.

For actual preparation of the predictions, at the node level, it is expected that nodes will work using some regional projection and datum. For example, nodes may choose to use a regular (e.g. Albers or Lambert) Equal Area projection for node wide compilation and processing of data. Nodes may also work at different grid resolutions (e.g. 25 m x 25 m, 30m x 30 m, 50m x 50 m, 90m x 90 m or 100 m x 100 m). It is only required that each node have procedures for re-projecting and re-calculating their original projected point data to deliver final, 100 m x 100 m block averages in geographic coordinates at a 3 arc-second resolution. Tools and procedures to facilitate coordinate conversions have been obtained and will be provided to project participants.

1.5 Excluded Non-soil Areas

Predictions of soil properties will not be made for grid cells that are considered to be occupied wholly or dominantly (> 50%) by non-soil materials, including permanent water and ice, bare rock and permanently sealed surfaces (urban areas and pavements). No attempt will be made to specify the types or proportions of non-soil materials in a grid cell.

Excluded grid cells will be identified by means of a mask file. The GlobCover Land Cover product, the finest resolution (300 m) and most widely accepted digital database of global land use and land cover currently available will initially be used to create this mask file. The non-soil classes of this land cover database will be used to identify grid cells that are predominantly (> 50%) non-soil. Each grid cell will report a value to identify NONSOIL. A value of 0 will indicate the grid cell is dominantly soil and a value of 1 will identify a grid cell that is dominantly occupied by non-soil materials. The GlobCover categories that are considered to identify non-
soil areas are: Artificial surfaces and associated areas (Urban areas >50%); Bare areas; Water bodies; Permanent snow and ice; and No data (burnt areas, clouds).

1.6 Tiling of Grid Cell Data

The project has adopted a modification of the global 3 arc-second SRTM data as its reference base. All coordinates will be an exact multiple of 3 arc-seconds from a raster origin (lower left corner) located exactly at a whole degree of latitude and longitude (see Figure 1).

A location file will be provided to specify the exact location of every 3 arc-second grid cell on a global basis. This will ensure exact spatial conformity of all 3 arc-second raster data contributed to the GlobalSoilMap.net project with no offsets, gaps, overlaps or duplication of grid cells. Each 3 arc-second grid cell has been assigned a unique ID number. The ID numbers identify a nested hierarchy in which the ID number of any grid cell explicitly identifies both the tile number in which a cell is located and the location of the grid cell within the tile. The unique ID number also identifies a number of coarser resolution grids cells within which any grid cell is nested.

The CIAT SRTM tiling system divides the world between 60° N and 60° S into 5° by 5° tiles numbered from 1 to 72 East-West (starting at 180°) and 1 to 24 North-South (starting at 60° N). Each 5° by 5° tile is identified by combining the EW 5° by 5° tile number (between 1 and 72) with the NS tile number (between 1 and 24) (e.g. 41-10). New tile numbers will be assigned to identify 5° by 5° tiles beyond 60° N and 60° S. Each 5° by 5° tile is further subdivided into 25 tiles of 1° by 1° numbered from 1-1 in the bottom left corner to 5-5 in the top right corner (Figure 1).

In the CIAT tiling approach, the centre of the lower left cell (the cell at the origin of each 1° by 1° tile) is placed exactly at the intersection of a whole degree of latitude and longitude. This places the lower left corner of the grid cell (the true origin of the raster matrix) exactly ½ of a single 3 arc-second grid cell to the south and west of a whole degree of latitude and longitude (dashed grid in Figure 1). In order to maintain all grid cells belonging to a given tile entirely within that tile, these specifications propose that the lower left corner of each grid cell be located exactly at the 1 degree by 1 degree intersection. This places the centre of the lower left grid cell exactly ½ of a single 3 arc-second grid cell to the north and east of a whole degree of latitude and longitude. This results in a minor shift of ½ of a grid cell length to the north and west for the GlobalSoilMap.net grid cells relative to the original source SRTM DEM data set (thicker solid line grid in Figure 1).

Nodes will be encouraged to obtain and process finer resolution imagery and other data sets to improve upon the initial identification of areas of non-soil based on the GlobCover product. For example, the global surface water database (SWDB) could be used to improve identification of areas of permanent surface water.
1.7 Unique Numbering of Grid Cell Data

A unique number is assigned to each 3 arc-second grid cell in the world. The largest integer number that can be stored using 32 bit integers is 4,294,967,295 (10 digits). There are potentially 279,936,000,000 unique grid cells (12 digits) for the entire globe at a resolution of 3 arc-seconds. Therefore, it is not possible to store a single unique integer ID number for every 3 arc-second grid cell in the world using 32 bit integer numbers.

The solution for this project is to break the unique number into two parts and store the two parts separately. The first part identifies the 1° by 1° tile (Tile ID) and the second identifies the unique number of each grid cell within a 1° by 1° tile (1 Deg Cell ID) (Table 1). The two numbers can be concatenated to produce a unique global grid cell ID number that is nested and hierarchical down to the finest resolution of 3 arc-seconds (and even down to 1 arc-second). The concatenated number can be stored in 64 bit integer format and so should be useable once most computers have 64 bit operating systems that support storage of 64 bit integers.
The hierarchical numbering system is designed to both uniquely identify each 3 arc-second grid cell and to explicitly identify the grid cells at several coarse resolutions within which any given finer resolution cell is nested. Coarser resolution grid cells are identified by truncating the ID number from the right. The length of the number provides the information needed to decide whether to truncate the number by 1 or 2 digits. The rule will be to truncate by 2 digits for levels 4 (length 3 digits) and 8 (length 8 digits) and by 1 for all other levels.

The 3-arc-second grid cells are explicitly nested within coarser resolution grid cells of 6 seconds (~180 m), 30 seconds (~900 m), 1 minute (~1,800 m), 3 minutes (~5,400 m), 6 minutes (~10,800 m), and 30 minutes (~54,000 m). This is done to facilitate generalization to coarser resolutions. The numbering scheme is based on dividing coarser resolution grid cells by a factor of 2, 3 or 5 with numbers assigned to identify grid cell locations within coarser cells as illustrated in Figure 2.

### Figure 2. Illustration of the approach used to number grid cells when dividing by 2, 3 or 5

1.8 Volumetric data:

The primary data intended for Tier 1 product reported for each node is the point predictions. The secondary data intended for Tier 2 product reported for each grid cell will represent a mean value for each property at each of the six specified depth intervals averaged over a horizontal area of 100 m by 100 m. The reporting area is therefore a volumetric pixel. Single values will be reported for total profile depth and plant exploitable (effective) depth, averaged over the entire extent of a grid cell, as these measures apply to a whole site and not to the six specified depths.
Figure 3. The primary spatial entity is a point estimate at the nodes of the 3 arc-second by 3 arc-second grid (Tier 1 product). The secondary spatial entity is defined as a square grid cell with a horizontal resolution of 3 arc-seconds by 3 arc-seconds (Tier 2 product). The soil property value reported will be an average for a 100 m by 100 m grid cell located at the center point of the 3 arc-second grid; the uncertainty reported will be the 95% prediction interval of that average. Please note the diagram above is a little misleading – the 100 m by 100 m blocks will overlap variably depending on latitude.

The advantage of reporting an average value for each depth interval of each grid cell emphasises that it is not feasible to describe or predict the variation in a soil property that occurs at distances shorter than the 100 m (3 arc-second) grid.

Variation in soil properties is known to occur over distances shorter than the resolution at which the predictions are being made (Figure 3). This proportion of the variation is effectively noise and is unexplainable at this resolution using the covariates available to support predictions. Prediction of a value at a single point within a grid cell is likely to produce a result that can fall anywhere within the full range of variation that can occur within the grid cell (Figure 3). Therefore, the best estimate of the value of a soil property within a grid cell that exhibits a large range of variation is the mean value for the entire cell.

The mean value for each soil property at each depth within each grid cell can be computed using several different approaches. Appendix B describes some possible approaches for computing mean values for soil properties for a depth interval of a 3 arc-second grid.
2. Soil Properties

2.1 Depth of Soil

In order to estimate soil properties at specific depth intervals, there is first a need to provide an estimate of the total depth of the soil within each grid cell. The project will estimate the following important depths for each grid cell.

Table 4. Specifications for properties related to reporting depth of soil

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Units</th>
<th>Precision^7</th>
<th>Reference</th>
<th>Description of Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Depth to Rock</td>
<td>cm</td>
<td>N3.0</td>
<td>Soil Survey Division Staff, 1993 Chapter 1 page 5</td>
<td>Depth in cm to a lithic or paralithic contact as defined in USDA Soil Survey Manual. If depth is &lt; 200 cm record actual depth in cm. If depth is &gt; 200 cm record actual depth if known. If not known exactly, record depth as 999 cm.</td>
</tr>
<tr>
<td>2</td>
<td>Plant Exploitable (Effective) Depth</td>
<td>cm</td>
<td>N3.0</td>
<td>Soil Survey Division Staff, 1993 Chapter 3 page 60</td>
<td>Effective depth in cm as defined in the USDA Soil Survey Manual. The lower limit of soil is normally the lower limit of biologic activity, which generally coincides with the common rooting depth of native perennial plants. This depth is where root penetration is strongly inhibited because of physical (including soil moisture or temperature) and/or chemical characteristics.</td>
</tr>
</tbody>
</table>

2.2 Primary Soil Properties – see Appendix x

GlobalSoilMap.net will produce estimates of soil property values, their uncertainty and their date of prediction at each of six specified depth increments for the following soil properties.

Table 5. Specifications for primary soil properties

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Unit</th>
<th>Precision^4</th>
<th>Reference</th>
<th>Description of Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Organic Carbon</td>
<td>g/kg</td>
<td>N4.0</td>
<td>ISO 10694</td>
<td>mass fraction of carbon by weight in the &lt; 2 mm soil material as determined by dry combustion at 900° C</td>
</tr>
<tr>
<td>4</td>
<td>pHx10</td>
<td></td>
<td>N3.0</td>
<td>ISO 10390</td>
<td>1:5 soil/water (divide by 10 to get correct pH)</td>
</tr>
<tr>
<td>5</td>
<td>Clay</td>
<td>g/kg</td>
<td>N3.0</td>
<td>Burt, 2004 Page 347</td>
<td>&lt; 2 um mass fraction of the &lt; 2 mm soil material determined using the pipette method</td>
</tr>
</tbody>
</table>

^7 The notation used to describe precision (e.g. N3.0) is interpreted as N = number, 3 = length of number, 0 = number of decimal digits. Wherever possible values are reported in integer format to avoid the extra overhead associated with storing and transmitting real numbers.
Definitions and methods of analysis for most of the soil properties are according to ISO standards as defined in FAO (2006) Annex 1: Methods for Soil Analysis (see Appendix D). Particle size distribution is defined according to the USDA Soil Survey Laboratory Methods Manual (Burt, 2004). The USDA definition of particle size classes has been recommended by FAO for use in the Soil Map of the World. Units for properties are reported in g/Kg or cm (instead of % or m) to reduce data storage and transmission costs by storing integer numbers.

### 2.3 Derived Soil Properties

From the above attributes, where no insufficient measured data exist the following two properties will be predicted using pedo-transfer functions that will be developed and specified by the data provider:

#### Table 6. Specifications for derived soil properties

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Units</th>
<th>Precision</th>
<th>Reference</th>
<th>Description of Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Bulk Density</td>
<td>Mg/m³</td>
<td>N3.1</td>
<td>ISO 11272</td>
<td>Bulk Density of the whole soil (including coarse fragments) in mass per unit volume by a method equivalent to the core method using a pedotransfer function</td>
</tr>
<tr>
<td>11</td>
<td>Bulk Density</td>
<td>Mg/m³</td>
<td>N3.1</td>
<td>ISO 11272</td>
<td>Bulk Density of the fine earth fraction of the soil (&lt; 2 mm) in mass per unit volume by a method equivalent to the core method using a pedotransfer function</td>
</tr>
<tr>
<td>12</td>
<td>Available Water Capacity</td>
<td>mm (total over the depth range)</td>
<td>N4.0</td>
<td>Burt, 2004 Page 137</td>
<td>Available water capacity computed for each of the specified depth increments using a specified pedotransfer function that references the values estimated above for organic carbon, sand, silt, clay and bulk density.</td>
</tr>
</tbody>
</table>

**NOTE:** AWC = \( f \) (total carbon, sand, silt, clay, % coarse fragments, bulk density) for the 6 depths. Profile-AWC is AWC summed over the effective depth.
2.4 Additional Soil Properties – see also Appendix G

The soil properties identified above represent the minimum data set agreed upon by the GlobalSoilMap.net consortium. This list in no way restricts individual countries or nodes from producing a longer list of predicted soil properties for their area of interest. For example, the following secondary variables (Table 7) are considered by some nodes to be important, desirable and feasible to predict. These nodes have indicated an intention to predict these additional soil properties but they are considered optional, from the point of view of these specifications.

Table 7. Specifications for secondary soil properties

<table>
<thead>
<tr>
<th>No.</th>
<th>Property</th>
<th>Units</th>
<th>Precision$^4$</th>
<th>Reference</th>
<th>Description of Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Electrical Conductivity</td>
<td>mS/m</td>
<td>N4.1</td>
<td></td>
<td>Electrical conductivity in 1:1 saturated paste</td>
</tr>
</tbody>
</table>

2.5 Time (Year)

The date of the actual or estimated time of sampling of the legacy soil data will be attached to each of the estimated soil properties at each grid cell. The date reported will reflect the year of publication for a map or the year of analysis for a sampled soil profile.

The maps of soil properties created in the GlobalSoilMap.net project will initially be based on making maximum use of legacy soils data collected and reported over many decades of field work. Data for any point or any map reflect the state of the soil at the time the point was sampled and analysed, or the map was produced. A gridted date map will be made to indicate the date (in years) that the soil property value most closely reflects.\(^8\)

---

\(^8\) It may be possible, in future Tiers (beyond Tier 2) of the GlobalSoilMap.net products to attempt to reconcile differences in soil property values reported for different times and under different land uses to one or more standardized reference dates (e.g. harmonized decadal values at 1970, 1980, 1990, 2000, 2010, etc.) and under the land use conditions current at each date. This will first require that regional legacy soil data sets be analyzed to detect and quantify directions and rates of change in soil property values under known land use and land management regimes. These regional values for rates of change under different land uses could be applied to the original predictions of soil property values, in combination with information on land use history at each grid cell, to harmonize soil property values to common reference years for each major regional land use type. This is a potential future product and is not part of the current specifications.
3. Uncertainty, Date & Accuracy

An important aspect of the GlobalSoilmap.net project is its estimate and reporting of the uncertainty associated with all soil property predictions.

3.1 Uncertainty Definition

For the purposes of these specifications, uncertainty is defined for each location and depth increment at increasing levels of sophistication as in Table 0. For tier 1, it is the 90% Prediction Interval (PI) which reports the range of values within which the true value is expected to occur 9 times out of 10 (or 90% of the time). There is no assumption that this prediction interval is necessarily symmetric around the predicted value.

3.2 Date Stamping

It would be valuable to provide a raster surface to indicate the age of observational data used in the spatial predictions. The practical reason is to show how old or out of date the data is or can be. This is of clear interest to users, and can prioritise future investments in sampling.

With each prediction location the (integer) year of the data related to the prediction can be given. The method of producing this is not specified but will be described as metadata. Methods might include the year of the nearest data point (or weighted average year of the nearest 8 data points for point-based predictions, or the year of publication of the map for map-based predictions.)
3.3 Validation of GlobalSoilMap.net predictions of soil property values

For Tier 1 and 2 products for any contiguous region no statement of accuracy is required (Table 0). For tier 1 as an addendum, but not a requirement, an appropriate measure for each property at each depth increment may be the root mean square error of the point predictions. This can be achieved by cross validation. [It is suggested, on average, one observation point per 10 000 square kilometres may be required and probably at least 50 points are required to obtain an estimate.]

In subsequent tiers beyond tier 2 it is anticipated a richer set of validation criteria will be used, including mean error, and the percentage of the map area that fall within the realised uncertainty limits (Tier 3), and ultimately should meet pre-specified accuracy measures (Tier 4). Ultimately, sampling will be required to produce such quality estimates – see Table 0.

3.4 Reproducibility of GlobalSoilMap.net predictions of soil property values

It is a goal of the GlobalSoilMap.net project that it be possible to replicate or reproduce all output products given access to the inputs used to produce them. In the first instance, for tier 1 products, reproducibility will be enabled by requiring that for each output value reported for any soil property, associated date and associated uncertainty there be full documentation of the methods used to produce those values. Documentation of methods used to produce each output will be reported using the standard template described below.

David Rossiter agreed to insert the metadata template here at a later date.

Future versions of the specifications will be expanded to require contributors to guarantee complete reproducibility by specifying all inputs, processes and outputs and by making all inputs and processes available for access and use by others.

3.5 Policy for release of interim GlobalSoilMap.net predictions of soil property values

It is desirable to upload and release for general use only data that fully meets all specifications by providing predictions of both soil property values and associated uncertainty for all soil properties at all depths. However, in the short term, preliminary or incomplete GlobalSoilMap.net products may be released or made available under the following circumstances.

- Where the data are made available for testing delivery processes or client use (i.e. not a full / official release).
Where an agreed minimum area (i.e. the node and GSM agree on the size and utility of the area covered) has at least some properties available to full specification for those properties including estimates of uncertainty.

References


Appendix A: Minimum data set for each 100 m grid cell

Table 8 below illustrates lists the information that must be provided for each grid cell and suggests a possible structure for providing that information in a data base format. This structure is not part of the current *GlobalSoilMap.net* specifications. A formal process is underway to develop a UML model for describing the delivery products of the *GlobalSoilMap.net* project. This UML model will be implemented in XML once it is defined, in order to facilitate data exchange and inter-operability. It is assumed that all nodes will submit data that has been harmonized to the methods of analysis specified for each of the 8 soil properties in these specifications (e.g. pH, organic carbon, sand, silt, clay, gravel, etc.)

Table 8. Illustration of the minimum data required for each grid cell

<table>
<thead>
<tr>
<th>Long (X)</th>
<th>Lat (Y)</th>
<th>Tile_ID</th>
<th>1 Deg Cell ID</th>
<th>AttributeAnalyzed</th>
<th>yearOfAnalysis</th>
<th>Top_Depth</th>
<th>Bottom_Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>For Example</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>pH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>organic carbon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sand</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>silt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>clay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>gravel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AWC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ECEC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Profile depth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Effective Depth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spline Smoothing Value</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AttributeValue</th>
<th>LowerUncertainty (5%)</th>
<th>Upper Uncertainty (95%)</th>
<th>Prediction Model</th>
<th>Pedo transfer Function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Some possible approaches for computing weighted mean value by grid cell.

This Appendix presents and discusses some alternative approaches for computing or assigning a bulked mean value to each grid cell for each soil property of interest. It is necessary to be more specific about how to calculate a bulked mean value for each soil property at each of the 6 depths within each square grid cell.

Below are some ideas or options to consider.

1. If using map based estimates, the bulked mean value for the soil property at a given depth for all soils listed for the polygon can be assumed to represent an areal average already and so should satisfy the requirement that the value for the cell represent a bulked area average.
2. Work at a grid resolution that is finer than the final reporting resolution (25, 30 or 50 m grid cells) and produce estimates of the soil property value at each depth for each of the finer resolution grid cells. Then compute an average value for the 100 m x 100 m grid cell as the mean of all values for the finer resolution cells.
3. Produce point-centred estimates for the centre of each grid cell at the working resolution (100 m or 90 m) and then compute a bulked mean value for each grid cell as the average value within a 3x3 or 5x5 window centred on each grid cell. This way the bulked area average reflects the average value within a larger window centred at each grid cell.
4. Don’t worry about it and just assume that any point centred prediction represents the bulked mean value for the entire grid cell.

Reporting a bulked mean value has the advantage of removing the short range variability in the value of a soil property within the extent of a grid cell. The uncertainty associated with estimation of a mean value for each reference depth within the full extent of a grid cell will also be lower than the uncertainty associated with estimation of a single value for each depth at a single point at the centre of a grid cell. The values reported for each grid cell should therefore be an estimate of the mean value of that property at each of the six specified depths within the extent of the cell occupied by soil materials (excluding non-soils).

It is necessary to be aware of, and to clearly acknowledge, that in some instances of strongly contrasting soils the reported bulked average value may not exist at any single physical location within the grid cell. Consider the case of a cell that is 50% organic peat soils and 50% sandy upland soils with no or very low organic matter content. The bulked mean value for organic carbon for the cell would represent a mean value between the high value for the peat soil and the low value for the sandy soil. This value is not likely to occur anywhere within the grid cell.
but it is representative of the mean value within the grid cell. It will be necessary to live with this dichotomy and acknowledge it.
Appendix C: Correlations of soil properties derived from different soil analytical methods

This Appendix identifies and discusses the need for pedotransfer functions to convert soil property values from their original method of analysis to the standard GlobalSoilMap.net reference method of analysis. For discussion purposes, examples are provided to illustrate conversion of data from several widely used non-reference methods into the specified reference methods.

C1.0 Rationale

A well-known issue with using legacy soils data is the inconsistency that arises from use of many different methods for analysing soils in the laboratory or describing them in the field. These different methods yield different values that are not exactly equivalent or comparable. This creates a need to harmonize values produced using different methods in order to make them roughly equivalent and comparable. Harmonization can be challenging.

In order to make use of legacy soils data in the GlobalSoilMap.net project, it will be necessary to convert measurements made using different laboratory methods into an equivalent value in the specified standard reference method. For example, values reported for organic carbon determined by non-reference methods will need to be converted into equivalent values in the reference method of dry combustion. Similarly, values for pH in 1:1 or 1:2 water will need to be converted the equivalent value in the standard reference method of pH in 1:5 water. Harmonization of values reported for sand, silt and clay computed using methods of textural analysis that use definitions for particle size fractions different from the reference method will also have to be converted to the standard particle size definitions adopted for these specifications.

Default pedotransfer functions could potentially be identified for each of the methods of analysis for each of the soil properties selected for inclusion in the project. However, locally specific pedotransfer functions have consistently proven to be more effective than global ones and there is widespread agreement that there is generally no universal equation for converting from one method to another in all instances (Konen et al., 2002; Meersmans et al., 2009; Jankauskas et al., 2006; Jolivet et al., 1998; de Vos et al., 2007).

Consequently, there will be a need to develop locally relevant pedotransfer functions at the node level that apply to restricted soil-landscape domains. Examples of conversion of values from non reference to reference methods are presented below for the primary soil properties of organic carbon, pH, sand, silt and clay.
C1.1 Organic Carbon

The standard reference method for reporting soil organic carbon for the *GlobalSoilMap.net* project is by dry combustion to at least 900° C (ISO 10694). Values of organic carbon will be reported in g/Kg with integer precision (N4.0). Because of its accuracy and completeness, the dry combustion method (Leco at 1000° C) has been used in many studies as a reference method against which to calibrate other methods (Grewal et al., 1991; Meersmans et al., 2009).

The dry combustion method is based on thermal oxidation of the OC and thermal decomposition of IC minerals by means of a furnace. It is a rapid, reliable method for the determination of the OC when IC is removed prior to combustion. In fact, dry combustion is considered to ensure oxidation of all OC so it is considered the most accurate method. It can be used as a reference to calibrate other methods against it (Biscutti et al., 2004).

In the dry combustion method, the carbon present in the soil is oxidised to carbon dioxide (CO$_2$) by heating the soil to at least 900 °C in a flow of oxygen-containing gas that is free from carbon dioxide. The amount of carbon dioxide released is then measured by titrimetry, gravimetry, conductometry, gas chromatography or using an infrared detection method, depending on the apparatus used. When the soil is heated to a temperature of at least 900 °C, in addition to organic carbon any inorganic carbon present as carbonate is also completely decomposed.

Total organic carbon can be determined directly or indirectly. Direct determination consists of previous removal of any carbonates present by treating the soil with hydrochloric acid. Indirect determination consists of a correction of the total carbon content for the carbonates present.

Examples of studies that have used dry combustion for calibrating other methods of analyzing organic carbon include Biscutti et al., 2004; Byre and Slaton, 2003; de Vos et al., 2007; Grewal et al., 1991; Kalembasa and Jenkinson, 1973; Jankauskas et al., 2006; Jolivet et al., 1998; Konen et al., 2002; Meersmans et al., 2009; Mikhailova et al., 2003; Sleutel et al., 2007; Soon and Abboud, 1991 and Wang et al., 1996.

A review of several studies (Table 9) illustrates that it is possible to produce regression equations to permit conversion of results produced by one method into equivalent values in a specified reference method (usually dry combustion). However, the studies also highlight the fact that local calibration equations that reflect differences in soils on a regional basis are usually needed.

It has not proven possible to provide a single universal equation to convert organic carbon values produced using other methods of analysis to equivalent values in the reference method of dry combustion. Each node will need to develop and apply node-specific conversions.
Table 9. Regression equations for harmonizing values of organic carbon to a reference standard

<table>
<thead>
<tr>
<th>No.</th>
<th>Target Method Y =</th>
<th>Source Method X</th>
<th>* Slope</th>
<th>+ Intercept</th>
<th>R2</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dry Combustion</td>
<td>Spectro-photonic</td>
<td>0.9800</td>
<td>0.0000</td>
<td>0.98</td>
<td>Soon and Abboud (1991)</td>
</tr>
<tr>
<td>2</td>
<td>Dry Combustion</td>
<td>Walkley-Black</td>
<td>1.0500</td>
<td>0.0000</td>
<td>0.98</td>
<td>Soon and Abboud (1991)</td>
</tr>
<tr>
<td>3</td>
<td>Dry Combustion</td>
<td>modified Tinsley</td>
<td>1.0400</td>
<td>0.0000</td>
<td>0.98</td>
<td>Soon and Abboud (1991)</td>
</tr>
<tr>
<td>4</td>
<td>Dry Combustion</td>
<td>modified Mebius</td>
<td>1.4000</td>
<td>0.0000</td>
<td>0.99</td>
<td>Soon and Abboud (1991)</td>
</tr>
<tr>
<td>5</td>
<td>Dry Combustion</td>
<td>Loss on Ignition (LOI)</td>
<td>0.6330</td>
<td>-9.3600</td>
<td>0.98</td>
<td>Soon and Abboud (1991)</td>
</tr>
<tr>
<td>6</td>
<td>Tinsley (1950)</td>
<td>LOI at 850 C</td>
<td>0.4620</td>
<td>-1.3600</td>
<td>0.99</td>
<td>Ball, 1964</td>
</tr>
<tr>
<td>7</td>
<td>Tinsley (1950)</td>
<td>LOI at 850 C</td>
<td>0.4600</td>
<td>-1.8700</td>
<td>0.99</td>
<td>Ball, 1964</td>
</tr>
<tr>
<td>8</td>
<td>Tinsley (1950)</td>
<td>LOI at 375 C</td>
<td>0.4580</td>
<td>-0.4000</td>
<td>0.99</td>
<td>Ball, 1964</td>
</tr>
<tr>
<td>9</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 360 C MLRA 65NE</td>
<td>1.1414</td>
<td>-0.6791</td>
<td>0.94</td>
<td>Konen et al., 2002</td>
</tr>
<tr>
<td>10</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 360 C MLRA 75NE</td>
<td>0.0672</td>
<td>-0.5359</td>
<td>0.94</td>
<td>Konen et al., 2002</td>
</tr>
<tr>
<td>11</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 360 C MLRA 95B</td>
<td>0.5743</td>
<td>0.1025</td>
<td>0.98</td>
<td>Konen et al., 2002</td>
</tr>
<tr>
<td>12</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 360 C MLRA 103 IA</td>
<td>0.6824</td>
<td>-2.8696</td>
<td>0.97</td>
<td>Konen et al., 2002</td>
</tr>
<tr>
<td>13</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 360 C MLRA 108 IL</td>
<td>0.6094</td>
<td>0.1949</td>
<td>0.98</td>
<td>Konen et al., 2002</td>
</tr>
<tr>
<td>14</td>
<td>DC (Dumas at 1000)</td>
<td>Walkley-Black</td>
<td>1.2500</td>
<td>0.1260</td>
<td>0.99</td>
<td>Grewal et al., 1991</td>
</tr>
<tr>
<td>15</td>
<td>LOI at 550</td>
<td>DC (Dumas at 1000)</td>
<td>1.6700</td>
<td>2.5100</td>
<td>0.76</td>
<td>Grewal et al., 1991</td>
</tr>
<tr>
<td>16</td>
<td>LOI at 550</td>
<td>LOI at 450</td>
<td>0.9970</td>
<td>0.5000</td>
<td>0.98</td>
<td>Grewal et al., 1991</td>
</tr>
<tr>
<td>17</td>
<td>DC (at 680 C)</td>
<td>Wet combustion</td>
<td>0.9920</td>
<td>0.0000</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>DC (at 680 C)</td>
<td>Tinsley I</td>
<td>0.9500</td>
<td>0.0000</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>DC (at 680 C)</td>
<td>Tinsley II</td>
<td>0.9530</td>
<td>0.0000</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>DC (at 680 C)</td>
<td>Tinsley III</td>
<td>0.9680</td>
<td>0.0000</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>DC (at 680 C)</td>
<td>Anne</td>
<td>0.9330</td>
<td>0.0000</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>DC (at 680 C)</td>
<td>Mebius</td>
<td>0.9530</td>
<td>0.4300</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>DC (at 680 C)</td>
<td>Walkley-Black</td>
<td>0.7600</td>
<td>-0.0800</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>DC (at 680 C)</td>
<td>Tyurin</td>
<td>0.9330</td>
<td>0.0000</td>
<td>Kalembasa &amp; Jenkinson, 1973</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>DC (Leico CNS 2000)</td>
<td>Walkley-Black</td>
<td>1.3350</td>
<td>0.5730</td>
<td>0.88</td>
<td>Mikhailova et al., 2003</td>
</tr>
<tr>
<td>26</td>
<td>DC Robo-prep</td>
<td>Walkley-Black</td>
<td>1.4490</td>
<td>0.4110</td>
<td>0.90</td>
<td>Mikhailova et al., 2003</td>
</tr>
<tr>
<td>27</td>
<td>DC (Leico at 1000 C)</td>
<td>Walkley-Black (classic)</td>
<td>1.4700</td>
<td>0.0000</td>
<td>0.84</td>
<td>Meersmans et al., 2009</td>
</tr>
<tr>
<td>28</td>
<td>DC (Leico at 1000 C)</td>
<td>Walkley-Black (modified)</td>
<td>1.2000</td>
<td>0.0000</td>
<td>0.87</td>
<td>Meersmans et al., 2009</td>
</tr>
<tr>
<td>29</td>
<td>Walkley-Black (mod)</td>
<td>Walkley-Black (classic)</td>
<td>0.8200</td>
<td>0.6800</td>
<td>0.53</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>30</td>
<td>DC (Leico at 1000 C)</td>
<td>DC (Carlo–Erba at 1020 C)</td>
<td>1.1300</td>
<td>-0.0600</td>
<td>0.99</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>31</td>
<td>Walkley-Black (modified)</td>
<td>DC (Leico at 1000 C)</td>
<td>0.7200</td>
<td>0.6300</td>
<td>0.73</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>32</td>
<td>Walkley-Black (modified)</td>
<td>DC (Carlo–Erba at 1020 C)</td>
<td>0.8100</td>
<td>0.5800</td>
<td>0.73</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>33</td>
<td>Walkley-Black (classic)</td>
<td>DC (Leico at 1000 C)</td>
<td>0.8900</td>
<td>-0.0900</td>
<td>0.99</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>34</td>
<td>Walkley-Black (classic)</td>
<td>DC (Carlo–Erba at 1020 C)</td>
<td>1.0200</td>
<td>0.1500</td>
<td>0.99</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>35</td>
<td>Walkley-Black (classic)</td>
<td>LOI at 360</td>
<td>0.4300</td>
<td>-0.0900</td>
<td>0.88</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>36</td>
<td>Walkley-Black (modified)</td>
<td>LOI at 360</td>
<td>0.3400</td>
<td>0.6300</td>
<td>0.44</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>37</td>
<td>Walkley-Black (modified)</td>
<td>LOI at 360</td>
<td>0.4300</td>
<td>0.6500</td>
<td>0.98</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>38</td>
<td>DC (Carlo–Erba at 1020 C)</td>
<td>LOI at 360</td>
<td>0.4800</td>
<td>-0.0030</td>
<td>0.89</td>
<td>Brye and Slaton, 2003</td>
</tr>
<tr>
<td>39</td>
<td>DC (Leico at 1000 C)</td>
<td>LOI at 360</td>
<td>0.4800</td>
<td>-0.0030</td>
<td>0.89</td>
<td>Brye and Slaton, 2003</td>
</tr>
</tbody>
</table>
Table 10. Regression equations for harmonizing values of organic carbon to a reference standard

<table>
<thead>
<tr>
<th>No.</th>
<th>Target Method Y =</th>
<th>Source Method X</th>
<th>* Slope</th>
<th>+ Intercept</th>
<th>R²</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>A-I colorimetric</td>
<td>Walkley-Black (classic)</td>
<td>0.5410</td>
<td>-0.0330</td>
<td>0.96</td>
<td>Chacón et al., 2002</td>
</tr>
<tr>
<td>41</td>
<td>A-I colorimetric</td>
<td>Walkley-Black (classic)</td>
<td>0.4590</td>
<td>-0.0580</td>
<td>0.94</td>
<td>Chacón et al., 2002</td>
</tr>
<tr>
<td>42</td>
<td>A-I colorimetric</td>
<td>Walkley-Black (classic)</td>
<td>0.4920</td>
<td>0.0000</td>
<td>0.99</td>
<td>Chacón et al., 2002</td>
</tr>
<tr>
<td>43</td>
<td>DC (Shimadzu at 900 C)</td>
<td>Walkley-Black not corrected</td>
<td>1.5800</td>
<td>0.0000</td>
<td>0.96</td>
<td>De Vos et al., 2007</td>
</tr>
<tr>
<td>44</td>
<td>DC (Shimadzu at 900 C)</td>
<td>Walkley-Black corrected</td>
<td>1.2000</td>
<td>0.0000</td>
<td>0.96</td>
<td>De Vos et al., 2007</td>
</tr>
<tr>
<td>45</td>
<td>LOI at 375 (Lab K)</td>
<td>DC (Vario EL at 1150 C)</td>
<td>1.2530</td>
<td>0.5030</td>
<td>0.87</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>46</td>
<td>LOI at 375 (Lab W)</td>
<td>DC (Vario EL at 1150 C)</td>
<td>1.2790</td>
<td>0.2380</td>
<td>0.89</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>47</td>
<td>Walkley-Black NRCS 1995</td>
<td>DC (Vario EL at 1150 C)</td>
<td>1.0200</td>
<td>0.1680</td>
<td>0.97</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>48</td>
<td>Tyurin photometric</td>
<td>DC (Vario EL at 1150 C)</td>
<td>0.8700</td>
<td>0.3690</td>
<td>0.98</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>49</td>
<td>Tyurin titrametric classic</td>
<td>DC (Vario EL at 1150 C)</td>
<td>0.8690</td>
<td>0.1620</td>
<td>0.91</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>50</td>
<td>LOI at 375 (Lab W)</td>
<td>LOI at 375 (Lab K)</td>
<td>0.8750</td>
<td>0.1500</td>
<td>0.88</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>51</td>
<td>Walkley-Black NRCS 1995</td>
<td>LOI at 375 (Lab K)</td>
<td>0.6100</td>
<td>0.3570</td>
<td>0.83</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>52</td>
<td>Tyurin photometric</td>
<td>LOI at 375 (Lab K)</td>
<td>0.5220</td>
<td>0.5250</td>
<td>0.84</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>53</td>
<td>Tyurin titrametric classic</td>
<td>LOI at 375 (Lab K)</td>
<td>0.5280</td>
<td>0.1390</td>
<td>0.87</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>54</td>
<td>Walkley-Black NRCS 1995</td>
<td>LOI at 375 (Lab W)</td>
<td>0.6350</td>
<td>0.4200</td>
<td>0.86</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>55</td>
<td>Tyurin photometric</td>
<td>LOI at 375 (Lab W)</td>
<td>0.5510</td>
<td>0.5570</td>
<td>0.89</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>56</td>
<td>Tyurin titrametric classic</td>
<td>LOI at 375 (Lab W)</td>
<td>0.5670</td>
<td>0.3060</td>
<td>0.85</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>57</td>
<td>Tyurin photometric</td>
<td>Walkley-Black NRCS 1995</td>
<td>0.8130</td>
<td>0.3110</td>
<td>0.97</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>58</td>
<td>Tyurin titrametric classic</td>
<td>Walkley-Black NRCS 1995</td>
<td>0.8240</td>
<td>0.0810</td>
<td>0.91</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>59</td>
<td>Tyurin titrametric classic</td>
<td>Tyurin photometric</td>
<td>0.9540</td>
<td>-0.1120</td>
<td>0.89</td>
<td>Jankauskas et al., 2006</td>
</tr>
<tr>
<td>60</td>
<td>Walkley-Black NRCS 1995</td>
<td>DC (Leico at 875 C)</td>
<td>0.9180</td>
<td>1.0000</td>
<td>0.99</td>
<td>Jolivet et al., 1998</td>
</tr>
<tr>
<td>61</td>
<td>Walkley-Black NRCS 1995</td>
<td>DC (Leico at 875 C)</td>
<td>0.9470</td>
<td>0.0000</td>
<td>0.99</td>
<td>Jolivet et al., 1998</td>
</tr>
<tr>
<td>62</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 550 C</td>
<td>0.6130</td>
<td>0.6000</td>
<td>0.99</td>
<td>Jolivet et al., 1998</td>
</tr>
<tr>
<td>63</td>
<td>DC (Leico at 875 C)</td>
<td>LOI at 550 C</td>
<td>0.6240</td>
<td>0.0000</td>
<td>0.99</td>
<td>Jolivet et al., 1998</td>
</tr>
<tr>
<td>64</td>
<td>DC (Shimadzu at 900 C)</td>
<td>Walkley-Black NRCS 1995</td>
<td>1.5060</td>
<td>0.0000</td>
<td>0.99</td>
<td>Lettens et al., 2007</td>
</tr>
<tr>
<td>65</td>
<td>DC (Shimadzu at 900 C)</td>
<td>Walkley-Black NRCS 1995</td>
<td>1.5940</td>
<td>0.0000</td>
<td>0.99</td>
<td>Lettens et al., 2007</td>
</tr>
<tr>
<td>66</td>
<td>DC (Shimadzu at 900 C)</td>
<td>Walkley-Black NRCS 1995</td>
<td>1.7740</td>
<td>0.0000</td>
<td>0.98</td>
<td>Lettens et al., 2007</td>
</tr>
<tr>
<td>67</td>
<td>Walkley-Black 6A1</td>
<td>DC (Leico at 1000 C)</td>
<td>0.9700</td>
<td>0.0000</td>
<td>0.99</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>68</td>
<td>DC (Leico at 1000 C)</td>
<td>LOI at 375 C siltstone</td>
<td>0.7320</td>
<td>-1.6100</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>69</td>
<td>DC (Leico at 1000 C)</td>
<td>LOI at 375 C sandstone</td>
<td>0.5620</td>
<td>-0.9950</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>70</td>
<td>DC (Leico at 1000 C)</td>
<td>LOI at 375 C basalt</td>
<td>0.4690</td>
<td>-0.9410</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>71</td>
<td>DC (Leico at 1000 C)</td>
<td>LOI at 375 C combined</td>
<td>0.7260</td>
<td>-1.5980</td>
<td>0.96</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>72</td>
<td>DC (Leico at 1000 C)</td>
<td>LOI at 375 C basalt</td>
<td>0.4690</td>
<td>-0.9410</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>73</td>
<td>Walkley-Black 6A1</td>
<td>DC (Leico at 1000 C) other</td>
<td>0.7390</td>
<td>-1.7590</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>74</td>
<td>Walkley-Black 6A1</td>
<td>DC (Leico at 1000 C) basalt</td>
<td>0.4520</td>
<td>-0.8910</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>75</td>
<td>LOI at 375 C basalt</td>
<td>DC (Leico at 1000 C)</td>
<td>0.4692</td>
<td>-0.9410</td>
<td>0.95</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>76</td>
<td>Walkley-Black 6A1</td>
<td>LOI at 375 C combined</td>
<td>0.4880</td>
<td>-2.3360</td>
<td>0.91</td>
<td>Wang et al., 1996</td>
</tr>
<tr>
<td>77</td>
<td>Walkley-Black 1934</td>
<td>DC (Variomax CNS)</td>
<td>1.0340</td>
<td>0.0160</td>
<td>0.99</td>
<td>Sleutel et al., 2007</td>
</tr>
<tr>
<td>78</td>
<td>Walkley-Black 1934</td>
<td>DC (Variomax CNS)</td>
<td>1.0130</td>
<td>0.0000</td>
<td>0.99</td>
<td>Sleutel et al., 2007</td>
</tr>
<tr>
<td>79</td>
<td>Springer-Klee, 1954</td>
<td>DC (Variomax CNS)</td>
<td>1.0020</td>
<td>0.0000</td>
<td>0.98</td>
<td>Sleutel et al., 2007</td>
</tr>
<tr>
<td>80</td>
<td>DC (Shimadzu at 900 C)</td>
<td>DC (Variomax CNS)</td>
<td>0.9430</td>
<td>0.0000</td>
<td>0.99</td>
<td>Sleutel et al., 2007</td>
</tr>
</tbody>
</table>
As a single measurement, pH describes more than relative acidity or alkalinity. It also provides information on nutrient availability, metal dissolution chemistry, and the activity of microorganisms (Miller and Kissel, 2010).


This standard specifies an instrumental method for the routine determination of pH using a glass electrode in a 1:5 (volume fraction) suspension of soil in water (pH in H2O), in 1 mol/l potassium chloride solution (pH in KCl) or in 0.01 mol L\(^{-1}\) calcium chloride solution (pH in CaCl\(_2\)).

Values for pH for the GlobalSoilmap.net project will be reported for a 1:5 suspension of soil in water. Values will be reported in byte format as pH x 10 with a precision of (N3.0) (value range of 0-149). These values will need to be divided by 10 to produce a correct pH value with a precision of 1 decimal place.

ISO 10390:2005 is applicable to all types of air-dried soil samples, for example pre-treated in accordance with ISO 11464. The most common method for analyzing pH in North America is a 1:1 soil/water suspension (Miller and Kissel, 2010). Adopting ISO 10390:2005 as a standard with its specification of pH measured in a 1:5 suspension of soil in water will require many values to be converted from 1:1 soil/water to 1:5 soil/water equivalent values.

The ratio of soil to water in a suspension has a net effect of increasing the pH with a decrease in the soil/water ratio. Keaton (1938) and Davis (1943) have shown that decreasing the soil/water ratio from 10:1 to 1:10 resulted in an increase of 0.40 pH units. Values for pH computed using methods with a lower ratio of soil to water (e.g. 1:1 or 1:2.5) will generally be lower than equivalent values for pH in 1:5 CaCl\(_2\) solution and will need to be adjusted higher. Several authors have demonstrated that fitting quadratic or curvilinear functions to soil pH data produces regression equations with higher coefficients of determination that those obtained from a linear fit (Aitken and Moody, 1991; Miller and Kissel, 2010).

Soil pH varies with season and soil moisture content with higher pH values associated with wetter soils and winter conditions and lower pH values with drier soils and summer conditions (Miller and Kissel, 2010). The effects of both temporal variation in pH and variation due to different methods means that small differences in pH may not be meaningful in the context of predictions made for the GlobalSoilmap.net project using legacy soils data.
### Table 11 Example regression equations for converting values of pH between different methods

<table>
<thead>
<tr>
<th>No.</th>
<th>Target Method (Y)</th>
<th>Source Method (X)</th>
<th>Equation</th>
<th>R²</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pH (1:1 0.01 m CaCl₂)</td>
<td>pH (1:1 water)</td>
<td>$y = 1.08(x) - 0.973$</td>
<td>0.98</td>
<td>Miller and Kissel (2010)</td>
</tr>
<tr>
<td>2</td>
<td>pH (1:1 0.01 m CaCl₂)</td>
<td>pH (saturated paste)</td>
<td>$y = 1.10 (x) - 0.923$</td>
<td>0.98</td>
<td>Miller and Kissel (2010)</td>
</tr>
<tr>
<td>3</td>
<td>pH (1:1 0.01 m CaCl₂)</td>
<td>pH (1:2 water)</td>
<td>$y = 1.05 (x) - 0.950$</td>
<td>0.97</td>
<td>Miller and Kissel (2010)</td>
</tr>
<tr>
<td>4</td>
<td>pH (1:1 water)</td>
<td>pH (1:1 0.01 m CaCl₂)</td>
<td>$y = x + 0.267 \times \text{EC} \times 1:1 \text{water}^{0.446}$</td>
<td>0.99</td>
<td>Miller and Kissel (2010)</td>
</tr>
<tr>
<td>5</td>
<td>pH (1:2 water)</td>
<td>pH (1:1 0.01 m CaCl₂)</td>
<td>$y = x + 0.239 \times \text{EC} \times 1:1 \text{water}^{0.506}$</td>
<td>0.98</td>
<td>Miller and Kissel (2010)</td>
</tr>
<tr>
<td>6</td>
<td>pH (1:5 0.01 m CaCl₂)</td>
<td>pH (1:5 water)</td>
<td>$y = 1.012 (x) - 0.76$</td>
<td>0.99</td>
<td>Conyers and Davey (1988)</td>
</tr>
<tr>
<td>7</td>
<td>pH (1:5 0.01 m CaCl₂)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.979 (x) - 0.71$</td>
<td>0.68</td>
<td>Bruce et al., (1989)</td>
</tr>
<tr>
<td>8</td>
<td>pH (1:5 0.01 m CaCl₂)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.887 (x) - 0.199$</td>
<td>0.88</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>9</td>
<td>pH (1:5 0.01 m CaCl₂)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.197 (x^2) - 1.21 (x) + 5.78$</td>
<td>0.92</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>10</td>
<td>pH (1:5 0.002 m CaCl₂)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.948 (x) - 0.308$</td>
<td>0.90</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>11</td>
<td>pH (1:5 0.002 m CaCl₂)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.178 (x^2) - 1.043 (x) + 5.10$</td>
<td>0.94</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>12</td>
<td>pH (1:5 1 m KCl)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.803 (x) + 0.077$</td>
<td>0.81</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>13</td>
<td>pH (1:5 1 m KCl)</td>
<td>pH (1:5 water)</td>
<td>$y = 0.233 (x^2) - 1.797 (x) + 7.143$</td>
<td>0.98</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>14</td>
<td>pH (soil solution)</td>
<td>pH (1:5 water)</td>
<td>$y = 1.28 (x) - 0.613$</td>
<td>0.78</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>15</td>
<td>pH (soil solution)</td>
<td>pH (1:5 0.01 m CaCl₂)</td>
<td>$y = 1.105 (x) - 0.140$</td>
<td>0.79</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>16</td>
<td>pH (soil solution)</td>
<td>pH (1:5 0.002 m CaCl₂)</td>
<td>$y = 1.050 (x) - 0.112$</td>
<td>0.80</td>
<td>Aitken and Moody (1991)</td>
</tr>
<tr>
<td>18</td>
<td>pH (soil solution)</td>
<td>pH (1:5 1 m KCl)</td>
<td>$y = 1.175 (x) - 0.262$</td>
<td>0.80</td>
<td>Aitken and Moody (1991)</td>
</tr>
</tbody>
</table>
C1.3 Particle Size Distribution (sand, silt and clay)

Soil texture represents the relative composition of sand, silt, and clay in soil. The particle-size distribution is usually represented in a texture diagram, relating the percentages of sand, silt, and clay to a texture class (Minasny and McBratney, 2001). The standard reference method adopted by the GlobalSoilMap.net project for reporting particle size classes of sand, silt and clay (g/Kg), is as per the USDA Soil Survey Laboratory Methods Manual (3A1a) (Burt, 2004 page 34). The Kilmer and Alexander (1949) pipet method was chosen by the USDA Soil Conservation Service because it is reproducible in a wide range of soils.

The current standard for particle size classes adopted by FAO for use in the Harmonized World Soil Database (FAO/IIASA/ISRIC/ISSCAS/JRC, 2009) is ISO 10390:2005. This standard differs from the USDA definition in defining the size range for silt as 2-63 µm instead of 2-50 µm and sand as 63-2000 µm instead of 50-2000 µm. This is a relatively new standard for FAO which previously adopted the USDA definitions for the digital soil map of the world (FAO, 1990).

Differences in values reported for soil particle size fractions can arise because of differences in method of analysis (e.g. hydrometer, pipette, laser diffraction) or differences classification of particle size fractions. Most literature on harmonization of soil texture data deals with harmonizing differences in reported particle size fractions (Figure 4).

![Figure 4. Particle size limits used in European countries, Australia and America (Adapted from Nemes et al., 1999a and Minasny and McBratney, 2001)](image-url)
Minasny and McBratney (2001) identified two major textural classifications in the world as the International and USDA/FAO systems (Table 12). The significant difference between these two was the choice of a threshold value for differentiating silt from clay of 20 µm for the International and 50 µm for the USDA. The new ISO/FAO standard adds an additional difference by changing the threshold value between silt and sand from 50 µm to 63 µm. This is a relatively minor difference but it still needs to be addressed.

Table 12. Differences between the International, USDA and ISO/FAO particle size classifications

<table>
<thead>
<tr>
<th>Size Fraction</th>
<th>International</th>
<th>USDA</th>
<th>ISO/FAO</th>
</tr>
</thead>
<tbody>
<tr>
<td>clay</td>
<td>&lt; 2 µm</td>
<td>&lt; 2 µm</td>
<td>&lt; 2 µm</td>
</tr>
<tr>
<td>silt</td>
<td>2 - 20 µm</td>
<td>2 - 50 µm</td>
<td>2 - 63 µm</td>
</tr>
<tr>
<td>sand</td>
<td>20-2000 µm</td>
<td>50-2000 µm</td>
<td>63-2000 µm</td>
</tr>
</tbody>
</table>

Both Minasny and McBratney (2001) and Nemes et al., (1999a) investigated options for harmonizing values for sand, silt and clay reported using different systems for classifying particle size fractions.

Using a compilation of four large databases consisting of a total of 1620 samples, Minasny and McBratney (2001) developed a single multiple linear regression model for converting between silt fraction based on the international standard of 2-20 µm (P2-20) to the 2-50 µm range of the USDA standard (P2-50) and vice versa. The equations are as follows:

\[
P_{2-50} = -18.3914 + 2.0971 (P_{2-20}) + 0.6726 (P_{20-2000}) - 0.0142 (P_{2-20})^2 - 0.0049 (P_{20-2000})^2 \\
(R^2 = 0.823)
\]

If \( P_{2-50} < 0 \) then \( P_{2-50} = 0.8289 (P_{2-20}) + 0.0198 (P_{20-2000}) \)

and

\[
P_{2-20} = -0.4070 - 0.1271 (P_{<2}) + 0.5527 (P_{2-50}) + 0.0017 (P_{<2})^2 - 0.0019 (P_{2-50})^2 + 0.0059 (P_{<2}) (P_{2-50}) \\
(R^2 = 0.818)
\]

If \( P_{2-20} < 0 \) then \( P_{2-20} = 0.1147 (P_{<2}) + 0.2212 (P_{2-50}) \)

Minasny and McBratney (2001) argued that most countries should consider adopting the particle size limits and texture classes of the USDA system. They noted that the 2- 50 µm particle size range is usually more useful than the 2- 20 µm range for estimating water retention in pedo transfer functions and observed that translations from one system into another were relatively easy, given improved computing power and algorithms.
There is already a package in R that supports conversion of particle size data reported in one system of classification to values in any specified other system. This package, provided by Julien Moeys with contributions by Wei Shangguan, applies a log-linear transformation of soil texture data from one particle size system into another (Moeys, 2010). Two modules exist, one that only accepts three data values as input (TT.text.transf) and the other that can translate any number of values for any number of size fractions (TT.text.transf.X). Log linear transformations have been shown to be the least reliable method for converting between different particle size classifications (Minasny and McBratney, 2001; Nemes, 1999a) but the simple fact that routines already exist in R to support rapid and efficient conversion from different systems into the USDA reference standard is encouraging. The GlobalSoilMap.net project will look at extending the functionality of this R package provided by Moeys (2010) to include additional options for converting between particle size classification systems.

The GlobalSoilMap.net project will develop an extended library of R functions for converting from systems of particle size classification different from the USDA to the standard particle size classes of the USDA system (clay = < 2 µm, silt = 2-50 µm and sand = 50-2000 µm). We will investigate and implement three main options of a) the spline and similarity methods of Nemes et al., (1999a,b) b) the regression equations of Minasny and McBratney (2001), and c) the graphical PSD conversion nomograms of Shirazi et al., 2001.
C1.4 Bulk Density

The standard reference method for reporting bulk density for the GlobalSoilMap.net project is the core method (ISO 11272).

The dry bulk density (BD) is the ratio between the mass of oven dry soil material and the volume of the undisturbed fresh sample. The ISO standard defines dry bulk density as the ratio of the oven-dry mass of the solids to the volume (the bulk volume includes the volume of the solids and of the pore space) of the soil.

The recommended ISO method (core method) uses steel cylinders of known volume (100 mL, 400 mL) that are driven in the soil vertically or horizontally by percussion. Sampling large volumes results in smaller relative errors but requires heavy equipment. The method cannot be used if stones or large roots are present or when the soil is too dry or too hard.

For soils with a high stone or root content or when the soil is too dry or too hard, methods based on the excavation technique are used as an alternative to the core method. In the excavation method a hole on a horizontal surface is dug and then filled with a material with a known density (e.g. sand which packs to a calibrated volume or water separated from the soil material by an elastic membrane). The soil obtained from the hole, is dried to remove the water and the dry mass is weighed.

The volumetric percentage of the coarse fragments needs to be determined in order to calculate the bulk density of the fine earth.

Experience has shown that organic carbon (OC) and texture predominately determine soil bulk density. Organic carbon and texture information is often available in soil survey campaigns. Therefore many attempts have been made to estimate soil bulk densities through some pedo-transfer functions (PTFs) based on soil OC and texture data (Curtis and Post 1964; Adams 1973; Alexander 1980; Federer 1983; Rawls 1983; Huntington et al. 1989; Manrique and Jones 1991; Bernoux et al. 1998; Tomasella and Hodnett 1998).

Heuscher et al., (2007) applied a stepwise multiple regression procedure to predict oven-dried bulk density from soil properties using the NRCS National Soil Survey Characterization Data. The database included both subsoil and topsoil samples. An overall regression equation for predicting oven-dried bulk density from soil properties ($R^2 = 0.45$, $P < 0.001$) was developed using almost 47,000 soil samples. Partitioning the database by soil suborders improved regression relationships ($R^2 = 0.62$, $P < 0.001$). Of the soil properties considered, the stepwise multiple regression indicated that organic C content was the strongest contributor to bulk density prediction. Other significant variables included clay content, water content and to a lesser extent, silt content, and depth.
Tranter et al., 2007 proposed a conceptual model that incorporated *a priori* knowledge for predicting soil bulk density from other more regularly measured properties. The model considers soil bulk density to be a function of soil mineral packing structures ($p_m$) and soil structure ($\Delta p$). Bulk-density maxima were found for soils with approximately 80% sand. Bulk densities were also observed to increase with depth, suggesting the influence of over-burden pressure. Residuals from the $p_m$ model, referred to as $\Delta p$, correlated with organic carbon.

Torri et al., (2007) developed a nomogram for transforming rock fragment content from a by-mass to a by-volume basis and vice versa. This nomogram facilitates comparison of data on rock fragment content expressed in different units.

Most PTFs for predicting bulk density, except those developed by Rawls (1983), Tomasella and Hodnett (1998), and Bernoux et al. (1998), are a function only of organic matter (OM)/OC content. Although studies conducted by Saini (1966) and Jeffrey (1970) have shown that OM has a dominating effect on soil bulk density and that it can be used alone as a good predictor of soil bulk density, it has been observed (e.g. Alexander 1980; Huntington et al. 1989; Manrique and Jones 1991) that soil texture plays a major role in controlling bulk density where OM is a minor component.

McBratney et al., (2002) proposed the concept of a soil inference system (SINFERS) that incorporated both expert soil knowledge and statistical prediction equations. The proposed system was intended to implement two major functions, namely:

1. Predict all soil properties using all possible (known) combinations of inputs and pedotransfer functions (PTFs).
2. Select the combination that leads to a prediction with the minimum variance.

The SINFER approach proposed by McBratney et al., (2002) will be the basis for efforts to create and apply PTFs for predicting soil bulk density for the GlobalSoilMap.net project.
C1.5 Available Water Capacity

The standard reference method adopted by the GlobalSoilMap.net project for reporting available water capacity is as per the USDA Soil Survey Laboratory Methods Manual (3D5a) (Burt, 2004 page 137).

Calculation of the water retention difference (WRD) is considered the initial step in the approximation of the available water capacity (AWC). WRD is a calculated value that denotes the volume fraction for water in the whole soil that is retained between 1500-kPa suction and an upper limit of usually 33 or 10-kPa suction (Burt, 2004 page 137). The upper limit (lower suction) is selected so that the volume of water retained approximates the volume of water held at field capacity. The 10-, 33- and 1500-kPa gravimetric water contents are then converted to a whole soil volume basis by multiplying by the bulk density (Db33) and adjusting downward for the volume fraction of rock fragments, if present in the soil. The lower suctions, e.g., 10 or 5-kPa, are used for coarse materials.

Results of research to develop hydraulic PTFs have been reported widely, including in the USA (Rawls et al., 1982), the UK (Mayr and Jarvis, 1999), the Netherlands (Wösten et al., 1995), and Germany (Scheinost et al., 1997b).” This research has attempted to correlate particle size distribution, bulk density and organic matter content with water content at field capacity (FC, θ at -33 kPa), permanent wilting point (PWP, θ at -1500 kPa), and available water content (AWC = FC - PWP) (Minasny, 2007). Other examples include studies by Nielsen and Shaw (1958), Burrows and Kirkham (1958), Slater and Williams (1965a, 1965b, 1966, 1967, 1969), Hall et al., (1977)Gupta and Larson (1979) Clapp and Hornberger (1978) and Bloemen (1980).

Gijsman et al., (2007) reported that many PTFs for estimating soil hydraulic properties have been published (see overviews by Rawls et al. (1991), Timlin et al. (1996) and Wösten et al. (2001). Timlin et al. (1996) reported 49 methods and estimated that this covers only about 30% of the total. Gijsman et al. (2002) compared eight methods for all the soil classes that make up the texture triangle. They went through the triangle in steps of 1% sand, 1% silt and 1% clay and determined the estimated values of wilting point or lower limit of plant extractable water (LL), field capacity, also referred to as the drained upper limit (DUL) and soil saturation (SAT). Gijsman et al. (2002) concluded that none of the methods were universally good. The best method in the comparison of Gijsman et al. (2002) was Saxton et al. (1986), closely followed by Rawls et al. (1982).

Jagtap et al. (2004) developed an approach that does not fit a mathematical equation through the data, but rather compares the soil layer for which the key soil water contents of LL, DUL and SAT have to be estimated with all layers in a database of field-measured soil–water-retention data. The layer that is most similar in texture and organic carbon concentration is considered to be the ‘nearest neighbor’ among all the layers in the database and its soil–water-retention values are assumed to be similar to those that need to be estimated. To avoid making estimated soil–water-retention values dependent on only one soil in the database, the six
'nearest neighbors' are used and weighted according to their degree of similarity (Jagtap et al., 2004). This is a non-parametric procedure, in the sense that it does not assume a fixed mathematical relationship between the physical properties and the water holding properties of soils. The similarity method to convert soil particle size fraction data proposed by Nemes et al. (1999a,b) is a direct analogue of this similarity method of Jagtap et al., (2004).

Zacharias and Wessolek (2007) identified three different approaches for deriving the WRC from more easily available parameters as:

1. Point-based estimation methods: estimating the water content of selected matric potentials from predictors such as the percentage of sand, silt, or clay, the amount of organic matter, or the bulk density (e.g., Gupta and Larson, 1979; Rawls and Brakensiek, 1982).
2. Semiphysical approach: deriving the WRC from information on the cumulative particle size distribution (Arya and Paris, 1981); theoretically, this approach is based on the similarity between cumulative particle size distribution and water retention curves. The water contents are derived from the soil’s predicted pore volume and the hydraulic potentials are derived from capillarity relationships.
3. Parameter estimation methods: using multiple regression to derive the parameters of an analytical closed-form equation for describing the WRC, using predictors such as the percentage of sand, silt, and clay, the amount of organic matter, or the bulk density (e.g., Vereecken et al., 1989; Wösten et al., 1999).

Zacharias and Wessolek (2007) concluded that approach 1 has the disadvantage that it uses a large number of regression parameters depending on the number of WRC sampling points, which makes its use in the mathematical modelling more difficult while for approach 2 very detailed information about the particle size distribution is required. They therefore preferred use of the parameter estimation methods.

Zacharias and Wessolek (2007) observed that pedotransfer functions that do not use the OM are rare and gave the following examples. Hall et al. (1977) developed point-based regression equations using soil texture and bulk density (only for subsoils) for British soils. Oosterveld and Chang (1980) developed an exponential regression equation for Canadian soils for fitting the relationship between clay and sand content, depth of soil, and moisture content. Equations to estimate the WRC from mean particle diameter and bulk density have been proposed by Campbell and Shiozawa (1989). Williams et al. (1992) analyzed Australian data sets and developed regression equations for the relationship between soil moisture and soil texture, structure information, and bulk density including variants for both the case where there is available information on OM and where the OM is unknown. Rawls and Brakensiek (1989) reported regression equations to estimate soil water retention as a function of soil texture and bulk density. Canarache (1993) developed point based regression equations using clay content and bulk density for Romanian soils. More recently, Nemes et al. (2003) developed different PTFs derived from different scales of soil data (Hungary, Europe, and international data) using artificial neural network modeling including a PTF that uses soil texture and bulk density only.
Zacharias and Wessolek (2007) developed two different regression equations depending upon the percentage of sand in a soil as follows:

Sand content < 66.5%

\[ \Theta_r = 0 \]
\[ \Theta_s = 0.788 + 0.001 \text{clay} - 0.263 D_b \]
\[ \ln(\alpha) = -0.648 + 0.023 \text{sand} + 0.044 \text{clay} - 3.168 D_b \]
\[ n = 1.392 - 0.418 \text{sand}^{-0.024} + 1.212 \text{clay}^{-0.704} \]

Sand content > 66.5%

\[ \Theta_r = 0 \]
\[ \Theta_s = 0.890 - 0.001 \text{clay} - 0.332 D_b \]
\[ \ln(\alpha) = -4.197 + 0.013 \text{sand} + 0.076 \text{clay} - 0.276 D_b \]
\[ n = 2.562 + 7 \times 10^{-9} \text{sand} + 3.750 \text{clay}^{-0.016} \]

The regression coefficients from these models were almost identical to those reported by Vereecken et al., (1989) (Vereecken \( \Theta_s = 0.81 + 0.001 \text{clay} - 0.283 D_b \)) for a different data set, adding further credibility to their general applicability. Zacharias and Wessolek (2007) recommended using the PTFs of Vereecken et al., (1989) if data on OM were available but felt that their proposed equations were suitable for use where OM data were not available.

Empirical equations developed by Williams et al. (1992) for the prediction of the constants \( A \) and \( B \) in the Campbell function have been widely used in Australia and elsewhere. These regression equations require particle size distribution, field texture and bulk density inputs as follows:

\[ A = 1.996 + 0.136(\ln C) - 0.00007(\text{FS215}) + 0.145(\ln SI) + 0.382(\ln TEX) \]
\[ B = -0.192 + 0.0946(\ln TEX) - 0.00151(\text{FS}) \]

\( C \) is % clay (< 0.002 mm); \( SI \) is % silt (0.002-0.02 mm); \( FS \) is % fine sand (0.02-0.20 mm), and \( TEX \) is texture group from 1-6 as defined by Northcote (1971).

Cresswell et al., (2006) demonstrated applicability of the Williams et al. (1992) method for French soils and confirmed that the approach of assuming a Campbell SWC model and
empirically predicting the slope and air entry potential has merit. They concluded that the empirical regression equations of Campbell appeared transferable to different data sets from very different geographical locations. They provided regression equations for all samples and stratified by horizon type that had $r^2$ values ranging from 0.81 to 0.91.

Cresswell et al., (2006) suggested a strategy for achieving adequate coverage of soil hydraulic property data for France that included an efficient sampling strategy based on the use of functional horizons (Bouma 1989) and a series of reference sites where soil hydraulic properties could be measured comprehensively. They argued that functional horizon method recognizes the soil horizon rather than the profile as the individual or building block for prediction (Wösten et al. 1985; Wösten and Bouma 1992). A significant feature of this approach is the capacity to create a complex range of different hydrologic soil classes from simple combinations of horizon type, sequence, and thickness.

It is anticipated that the SINFER approach proposed by McBratney et al., (2002) will be the basis for efforts to create and apply PTFs for predicting available water capacity for the GlobalSoilMap.net project. These PTFs have yet to be developed.
References for Appendix C


Appendix D: Equal Area projections for use by each of the GlobalSoilmap.net nodes

This Appendix proposes to suggest a preferred equal area projection for each node to use for collating and processing projected data sets used to predict soil properties (Table D1).

Table D1. Suggested continental scale projections and their parameters for each node

<table>
<thead>
<tr>
<th>Abbr</th>
<th>Description</th>
<th>Australia</th>
<th>Africa</th>
<th>Asia</th>
<th>Europe</th>
<th>N. America</th>
<th>S. America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proj</td>
<td>Projection</td>
<td>Albers EA</td>
<td>Lambert EA</td>
<td>Mercator</td>
<td>Lambert EA</td>
<td>Albers EA</td>
<td>Albers EA</td>
</tr>
<tr>
<td>lat_1</td>
<td>Latitude of 1st standard parallel</td>
<td>-18</td>
<td></td>
<td>0</td>
<td></td>
<td>29.5</td>
<td>23</td>
</tr>
<tr>
<td>lat_2</td>
<td>Latitude of 2nd standard parallel</td>
<td>-36</td>
<td></td>
<td>0</td>
<td></td>
<td>45.5</td>
<td>23</td>
</tr>
<tr>
<td>lat_0</td>
<td>Latitude of Origin</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
<td>52</td>
<td>23</td>
</tr>
<tr>
<td>lat_ts</td>
<td>Latitude of true scale</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lon_0</td>
<td>Central Meridian</td>
<td>132</td>
<td>20</td>
<td>0</td>
<td>10</td>
<td>-96</td>
<td>0</td>
</tr>
<tr>
<td>x_0</td>
<td>False Easting - X</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4321000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y_0</td>
<td>False Northing - Y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3210000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ellps</td>
<td>Ellpsoide</td>
<td>GRS80</td>
<td>WGS84</td>
<td>GRS80</td>
<td>GRS80</td>
<td>aust_SA</td>
<td></td>
</tr>
<tr>
<td>datum</td>
<td>Datum</td>
<td>toWGS84</td>
<td>WGS84</td>
<td>NAD83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>units</td>
<td>Units of Distance</td>
<td>metres</td>
<td>metres</td>
<td>metres</td>
<td>metres</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>Semimajor radius of the ellipsoid axis</td>
<td>6378137</td>
<td>6378137</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>Semiminar radius of the ellipsoid axis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>Scaling factor</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nadgrids</td>
<td>Grid based datum adjustment</td>
<td>&quot;@null&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wktext</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_defs</td>
<td>Don't use the defaults file</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is expected that each node will define a single node-wide projection in which to work. It is further expected that this projection will be some type of equal area projection in which all grid cells have the same fixed resolution. Equal area projections which organize data into grid cells of fixed horizontal dimensions are required by some of the key programs used to compute terrain attributes from DEM data or to implement geostatistical procedures such as kriging.

Nodes that work at a grid resolution finer than 100 m will be able to use the finer resolution data to compute bulked mean values for a 3 arc-second by 3 arc-second grid cell by averaging the values for all grid cells that occupy a target 3 arc-second by 3 arc-second reporting grid cell.

Nodes that elect to work at a grid resolution of 100 m or greater will need to use the property values of surrounding grid cells to compute a weighted average value for each property at each depth for each target 3 arc-second by 3 arc-second reporting grid cell.
Suggested projections expressed in terms of R-code

- **1st level - Whole world compilation projections**
  - **world** proj4: `+proj=lonlat +ellps=WGS84`
  - **Googlemaps** proj4: `+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0 +k=1.0 +units=m +nadgrids=@null +wktext +no_defs`

- **2nd level – Continental scale compilation projections at the node level**
  - **au** (Australia and New Zealand) proj4: `+proj=aea +lat_1=-18 +lat_2=-36 +lat_0=0 +lon_0=132 +x_0=0 +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs`
  - **af** (Africa) proj4: `+proj=laea +lat_0=0 +lon_0=20 +x_0=0 +y_0=0 +units=m +ellps=WGS84 +datum=WGS84`
  - **as** (Asia) proj4: `+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0 +k=1.0 +units=m +nadgrids=@null +wktext +no_defs`
  - **eu** (Europe) proj4: `+proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80 +units=m +no_defs`
  - **na** (North America) proj4: `+proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=23 +lon_0=-96 +x_0=0 +y_0=0 +ellps=GRS80 +datum=NAD83 +units=m +no_defs`
  - **sa** (South/Central America) proj4: `+proj=aea +lat_1=-5 +lat_2=-42 +lat_0=-32 +lon_0=-60 +x_0=0 +y_0=0 +ellps=aust_SA +units=m +no_defs`
Appendix E: Background on uncertainty and guidelines for uncertainty methods

A general framework for assessing and representing uncertainties in environmental data is provided by Brown (2004).

Heuvelink and Brown (DSM 2007) observed that “soil data are rarely certain or ‘error free’, and that these errors may be difficult to quantify in practice”. Indeed, the quantification of error (defined here as a ‘departure from reality’) implies that the ‘true’ state of the environment is known. They reported that “in recent years, a distinct spectrum of methods, not altogether statistical, has emerged for dealing with situations of ‘imperfect knowledge’ in scientific research (see Ayyub, 2001 also)”. A spectrum of methods for uncertainty analysis is indeed important for DSM (and GSM) with due consideration of the potential sources of uncertainty—namely from inputs (observed data and covariate information), model parameters and model structure. Similarly, methods of uncertainty analysis will vary on the basis of whether soil point data or existing soil maps are used for producing GSM outputs. Moreover, methods will also vary dependent on the density of point data as well.

As such, guidance on uncertainty analysis for GSM product will take the following form:

When there are sufficiently many point observations, there are two general approaches:

1. Statistical modeling (principally geostatistical models) of the soil properties directly. The uncertainty of predictions is generated from the model as a byproduct.
2. Statistical modeling (principally geostatistical models) of residuals from independent data set or resampling techniques

If there are insufficient point observations then the use of expert knowledge may be a viable option. Examples of this include:

1. Uncertainty parameters e.g. the lower and upper limits for 90% prediction intervals for soil attributes. In practice this could mean the upper and lower limits for each property and soil class.
2. Expert elicitation for (parametric form of) distributions of soil attributes in specific soil classes
3. For soil maps where soil class proportions and attribute ranges are quantified, these values can be combined in a conservative manner (via lowest of the low etc.) or via Monte Carlo simulation.
4. Expert elicitation for variogram specification
As stated in the body of the specification document, for tier 1 GSM, our uncertainty of the reality is to be expressed as the 90% Prediction Interval (PI) which reports the range of values within which the true value is expected to occur 9 times out of 10 (or 90% of the time). There is no assumption that this prediction interval is necessarily symmetric around the predicted value. For all uncertainty methods, it is the probabilistic ones that may be most practicable for GSM. In regards to probability density functions (pdfs), Heuvelink and Brown (DSM 2007) argued that they confer a number of advantages over non-probabilistic techniques. For example, pdfs include methods for describing interdependence or correlation between uncertainties, methods for propagating uncertainties through environmental models and methods for tracing the sources of uncertainty in environmental data and models (Heuvelink, 1998). Notwithstanding these advantages, and the current popularity of stochastic methods in environmental research, there are a number of ongoing challenges for the successful application of pdfs to environmental data. In particular, there is a need to support the identification and estimation of pdfs in specific cases, as well as their storage in environmental databases.

Thus, the general pdfs need to be simplified in order to make them estimable in practice and tractable to storage within a soil database. The pdf of a numerical or categorical constant may be simplified by describing the uncertainty with a characteristic shape function, for which a small number of parameters must be estimated. Rather than specifying the entire pdf it is therefore sufficient to define the shape function and to estimate its parameters. For example, measurement error in a continuous numerical attribute is often assumed to follow a normal distribution (Heuvelink, 1998). This implies that the pdf is reduced to only two parameters, namely the mean and standard deviation, which describe the bias and average magnitude of uncertainty in the soil attribute, respectively. Similarly, it may be reasonable to assume that the number of stones in a volume of soil is Poisson distributed, for which the discrete pdf is reduced to only one parameter.

Useful simplifications must satisfy two conditions. First, the simplified pdfs must be estimable in practice, as well as tractable to storage within a soil database. Secondly, they must approximate the uncertainty in a soil variable sufficiently for their intended application. Among others, the elaboration and subsequent storage in the database must include the following aspects:

1. Uncertainty is subjective. The database must allow the opinions of different ‘experts’ to be stored.
2. Uncertainty information is very sensitive to the support size of the data items (Heuvelink and Pebesma, 1999). Here ‘support’ refers to the volume, magnitude and length of the entity described. Support size (in time and space) should always be specified in a database.
3. The uncertainty in a particular variable may well be statistically dependent on the uncertainty in another variable. Statistical dependencies (and cross correlations) between uncertain variables can have a marked influence on how uncertainties propagate in a modelling study. These create a need to address uncertainty in spatially...
distributed or dynamic attributes because these are strongly affected by about dependencies and correlations.

In terms of implementation, where there is sufficient point data to define the underlying probability distribution function (pdf) for conventional statistical analysis, a 90% confidence interval will be calculated to establish the range of the prediction interval (PI) for each predicted soil property. When using statistical modeling (principally geostatistical models) of the soil properties directly. The uncertainty of predictions is generated from the model as a byproduct from which the necessary PIs can be derived. Sufficient data or computing power is often not available to achieve this. (Could flesh this out a bit more???)

An alternative method to estimate PI has been presented and described in Malone et al. 2011). Here uncertainty is treated as the probability distribution of the output model errors, which comprises all sources of uncertainty (model structure, model parameters and input data). And since it is estimated through an empirical distribution, it is not necessary to make any assumption about residuals (Solomatine and Shrestha 2009). This method is particularly useful when we are dealing with soil spatial prediction functions that include data-mining tools or neural networks (as examples) in combination with the regression-kriging approach, where it would be difficult to use other existing methods (of uncertainty analysis) to estimates of uncertainty. The approach can be summarized in the following steps:

1. Apply an unsupervised classification technique (e.g. fuzzy k-means) to the covariate data layers assembled and used to make the predictions of soil properties for a particular area or soil-landscape zone to produce functional classes (4-5).
2. Overlay all available geo-registered soil profile analytical data on the resulting 4-5 functional class map for a particular region or soil-landscape zone of interest.
3. Compute the probability distribution function (pdf) for each soil property of interest, at each depth of interest, within the 4-5 functional classes. This establishes the range and distribution of observed soil property values within each of the 4-5 functional classes.
4. Use the pdf computed for each soil property at each depth for each class to identify the values at the 5% and 95% confidence limits (the 90% prediction interval or PI).
5. Use the values at the 5% and 95% confidence limits of the pdf for each class as inputs in calculating a weighted fuzzy mean value for the upper and lower confidence intervals for each grid cell.
6. The method for computing upper and lower confidence limits for any grid cell is based on computing a weighted average of the confidence limit value for each of the 4-5 classes times the fuzzy likelihood value of that class for all n classes at each location.
7. The estimate of uncertainty at each grid cell is a weighted average of the similarity of the conditions at each cell to the conditions that define each of the N classes.

The approach of Malone et al., (2011) requires that there be a sufficient number and density of point observations within any given prediction area (30 per class) to support a data driven assessment of the pdf of a given soil property by class within the geographic extent of an area of interest.
If sufficient information does not exist to support conventional statistical analysis, the range will have to be assessed by appropriate local or national experts. Fuzzy logic (Cazemier et al., 2001) and Bayesian beliefs (O’Hagan et al., 2006) have been proposed as suitable frameworks for establishing estimates of uncertainty in the absence of sufficient hard field data.

See also Appendix H

As an example, Lilburne et al., (2009) presented a method based on using expert knowledge to estimate the pdf in situations where there is insufficient information to support conventional statistical analysis. This following method is adapted and presented as one example of how expert knowledge can be used to estimate uncertainty for data derived from a polygon soil map, as follows:

1. Best available expert knowledge and observed or measured data are examined for each soil map unit, or for taxonomic classes included within the map unit.
2. The variability of the property is described in the form of a probability distribution function (pdf).
3. If data are available, a normal, lognormal, or beta function can be used. An additional combination pdf termed a duplex function has also been proposed. This combines a triangular or uniform distribution with a single-valued discrete pdf for the minimum or maximum value.
4. Confidence in the base property data is indicated by an expert assigned confidence code.

For the present, uncertainty will be reported as the best feasible estimate of the range of values within which a prediction of a soil property at any depth and any location is expected with 90% confidence.

References for Appendix E (What exactly are the other that are cited here???)

Appendix F: North American Node tiering concepts

The North America Node of GlobalSoilMap.net has developed a tiering concept where each successive tier of map products incorporates improvements over the previous tier. In the United States, initial tiers will rely heavily on the use of existing area-class maps as held in STATSGO2 and SSURGO, with the understanding that later tiers will make better use of spatial disaggregation and point observations and eventually lead to truly continuous raster property maps. For the United States, the proposed tiering system is as follows:

“Tiers 0.1–0.4”. Essentially rasterised versions of the existing STATSGO2 map. Soil components have been fitted with equal-area splines (Bishop et al., 1999; Malone et al., 2009), which allows property estimates to be made at the GlobalSoilMap.net standard depth increments. Weighted means are then calculated within map units where more than one soil component possesses data, otherwise the single soil component’s property estimate is reported. The products may have gaps where soil exists but property values were not recorded. Map units will not have been harmonized so artifacts (data discontinuities) may exist at political boundaries. The products are not truly continuous. Pre-existing scale discrepancies are inherited from STATSGO2. For initial tiers, the “upper” and “lower” values reported in STATSGO2 are assumed to be the 95% confidence limits for uncertainty estimation. Later tiers may estimate probability distribution functions and hence confidence intervals from legacy point observations in the National Soil Characterization Database.

“Tiers 0.5–0.9”. Essentially as per tiers 0.1–0.4 but the higher-detail SSURGO map will be used. Tiers will begin to incorporate harmonization of map units at political boundaries, which may necessitate the use of spatial disaggregation and other predictive techniques. Techniques will need to be developed to fill gaps in SSURGO map where no SSURGO data currently exists. Disaggregation of STATSGO2 may be a potential solution, as may other predictive approaches such as the homosoil concept (Mallavan et al., 2010). Uncertainty as per tiers 0.1–0.4, or other quantitative techniques.

“Tiers 1.0+”. These tiers will be produced using predictive techniques. Products will not contain gaps other than where bedrock or water exists. Products will not contain scale discrepancies. Uncertainty will be represented as 95% confidence interval estimated by quantitative techniques will be refined as more point observations become available. Tiers <1 will be released for comment and feedback, but it is anticipated that collated international products from nodes will meet v1 standards.
Appendix G: National variants of globalsoilmap Products

Besides the concept of tiering there will be a need for various individual countries to produce products that otherwise meet globalsoilmap tiering standards in terms of resolution, depth range, uncertainty etc., but *do not correspond to* the standard set of properties. These are designated a national superset, GLOBALSOILMAPnss ‘name of country’ ‘tier no’ ‘soil property’ e.g., GLOBALSOILMAPnss USA V1 pH in KCl(1:2), and are intended for national use only, Common examples may be soil properties based on local analytical methods, particularly Ph and carbon, cation exchange. It is envisaged that in some or many instances these may be produced first and pedotransfer functions (Appendix C) used to convert these to the globalsoilmap standard.
Appendix H: An approach for uncertainty estimation of soil spatial predictions based on soil class (maps) with limited within-class variability information

Introduction
Often we do not have a sufficient quantity of soil property data per soil class in order to generate soil property distributions and estimate their parameters (e.g., mean and variance), which may then be used to derive confidence intervals. Techniques exist for estimating parameters of distributions based on small samples but they may be computationally expensive or specific for a particular distribution (normal, lognormal, etc.). Although sample data may be limiting, often we have some idea of the range and typical value of soil properties for soil classes. In the presence of such data, a useful distribution to use is the triangular distribution. It may be positively or negatively skewed or symmetrical, but it must be unimodal (Johnson, 1997). It has been used in risk analysis, where it is often a proxy for the more computationally complex beta distribution often used when sample data is limiting (Williams, 1992). The probability density function of the triangular distribution is given in Figure 1.

![Figure H5. Probability density function of the triangular distribution.](image)

We can use the triangular distribution in a sampling simulation approach to estimate uncertainty, as will be explained in the following section.

General approach
For some spatial entity (e.g., a map unit delineation or a grid cell), assume $s$ soil classes $S_i, i = 1...s$ exist or have the potential to exist. Draw $n$ times from the triangular soil property distribution of each soil class. A random draw from the triangular distribution is determined as follows:

$$
\begin{align*}
g &= \begin{cases} 
a + \sqrt{U(b-a)(c-a)} & \text{for } 0 < U < F(c) \\
b - \sqrt{(1-U)(b-a)(b-c)} & \text{for } F(c) \leq U < 1 \end{cases}
\end{align*}
$$

(H1)
where \( g \) is the soil property value drawn from the distribution, \( U \) is a random variate drawn from the uniform distribution in the interval \((0,1)\), \( F(c) = (c - a)/(b - a) \) and \( a, b \) and \( c \) are the parameters of the triangular distribution (Figure ).

For each draw \( N = 1 \ldots n \), calculate the weighted mean property value, \( \mu_N^* \), as:

\[
\mu_N^* = \sum_{i=1}^{s} w_i g_i N
\]

where \( g_i N \) is the soil property value drawn from the triangular distribution and \( w_i \) is the weight associated with \( S_i \). Weights are explained below.

After all draws have taken place, we will have a set of \( n \) weighted mean soil property values from which a distribution can be generated. The lower and upper bounds of the 90% confidence interval are then the 5th percentile and 95th percentile of the distribution of weighted means, respectively.

**Special cases**

Two special cases of this approach exist. The first case involves situations where the most feasible approach is to generate within-map unit spatially weighted mean soil property maps. In this case the spatial entity upon which calculations are carried out is the soil map unit polygon. For each polygon, \( s \) soil classes are assumed to exist in known (or assumed) proportions, but the precise spatial distribution of each soil class is unknown. The soil property map thus reports a single weighted mean soil property value for each map unit, for each depth increment. In this case, the weight \( w_i \) in equation 2 is the areal proportion of soil class \( S_i \) in the map unit (Odgers et al., 2012).

The second case involves situations where the soil polygon map has been spatially disaggregated. The result of the spatial disaggregation is a set of \( s \) raster maps where the map for \( S_i \) indicates the probability of occurrence of \( S_i \) at each grid cell. This probability varies continuously in space. The spatial entity upon which calculations are carried out is the grid cell. In this case the weight \( w_i \) in equation 2 is the probability of occurrence of \( S_i \) at the given grid cell.

Obviously these methods can be applied for any probability density function.

**References**


Appendix I Suggested data density for point-based methods

Experience suggests that a density of between 1 and 10 observations per 1000 km$^2$ is required for point predictions. The larger densities are required for regions with larger intrinsic pedodiversity and where the legacy sampling is clustered (in space or with respect to the covariates).