A Global Soil Data Set for Earth System Modeling

Table of content

Introduction
Data description
Data usage
Citation
Reference
Contact

1. Introduction

We developed a comprehensive, gridded Global Soil Dataset for use in Earth System Models (GSDE) and other applications as well. GSDE provides soil information including soil particle-size distribution, organic carbon, and nutrients, etc. and quality control information in terms of confidence level. GSDE is produced following an improved protocol of the Harmonized World soil Database (HWSD)(FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012). GSDE is based on the Soil Map of the World and various regional and national soil databases, including soil attribute data and soil maps. We used a standardized data structure and data processing procedures to harmonize the data collected from various sources. We then used a soil type linkage method (i.e. taxotransfer rules) and the polygon linkage method to derive the spatial distribution of soil properties. To aggregate the attributes of different compositions of a mapping unit, we used three mapping approaches: area-weighting method, the dominant soil type method and the dominant binned soil attribute method. In the released gridded dataset, we used the area-weighting method as it will meet the demands of most applications. The dataset can be also aggregate to a lower resolution. The resolution is 30 arc-seconds (about 1 km at the equator).The vertical variation of soil property was captured by eight layers to the depth of 2.3 m (i.e. 0- 0.045, 0.045- 0.091, 0.091- 0.166, 0.166- 0.289, 0.289- 0.493, 0.493- 0.829, 0.829- 1.383 and 1.383- 2.296 m).

2. Data description

2.1 NetCDF format

We offered two versions with different resolution, i.e., 30 seconds (~1km) and 5 minutes (~10km).

For the 30 seconds version, we split the data of one soil property into two files to avoid too big files. One file is for the top four layers, the other is for the bottom 4 layers.

Here we take pH value (KCl) file (“PHK1.nc”) as an example to show the data. The dataset takes the NetCDF Climate and Forecast Metadata Convention (CF-1.0). The extent is 180°W -180°E and 84°N- 56°S. The following is the metadata of 30 seconds version :

```
dimensions:
        lon = 43200 ;
        lat = 16800 ;
        depth = 4 ;
variables:
        float lon(lon) ;
                lon:long_name = "longitude" ;
                lon:units = "degrees_east" ;
        float lat(lat) ;
                lat:long_name = "latitude" ;
                lat:units = "degrees_north" ;
        float depth(depth) ;
                depth:long_name = "depth to the bottom of a soil layer" ;
                depth:units = "centimeter" ;
        byte PHK(lon, lat, depth) ;
                PHK:missing_value = -100 ;
                PHK:units = "1/10" ;
                PHK:long_name = "pH(KCl)" ;

// global attributes:
                :Conventions = "CF-1.0" ;
```

The 5 minutes version is produce by aggregating the 30 seconds version. We used the dominant class method for general information, and the area weighting method for soil properties (Shangguan, 2014). We also provide a 5 minutes version of soil organic carbon density calculated using the aggregating after approach (Shangguan, 2014). The 5 minute version is provided in a single file with the metadata as following:

```
dimensions:
        lon = 4320 ;
        lat = 1680 ;
        depth = 8 ;
variables:
        float lon(lon) ;
                lon:long_name = "longitude" ;
                lon:units = "degrees_east" ;
        float lat(lat) ;
                lat:long_name = "latitude" ;
                lat:units = "degrees_north" ;
        float depth(depth) ;
                depth:long_name = "depth to the bottom of a soil layer" ;
                depth:units = "centimeter" ;
        short PHK(lon, lat, depth) ;
                PHK:missing_value = -999 ;
```

```
        PHK:units = "1/10" ;
        PHK:long_name = "pH(KCl)" ;

// global attributes:
        :Conventions = "CF-1.0" ;
```

2.2 Binary format

The spatial coverage of binary files is global, with 43200 columns (longitude) and 21600 rows (latitude). The values are stored by rows, from 180°W to 180°E and from 90°N to 90°S.

There are 11 types of soil general information for soil profiles and 34 soil properties for 8 depths.

The soil general information is as follows:

Filename	description	units	Data type	Missing value[a]
ADD_PROP	additional property		Signed byte	-100
AWC_CLASS	available water capacity		Signed byte	-100
DRAINAGE	drainage class		Signed byte	-100
IL	impermeable layer		Signed byte	-100
NONSOIL	nonsoil class		Signed byte	-100
PHASE1	phase1		Signed byte	-100
PHASE2	phase2		Signed byte	-100
REF_DEPTH	reference soil depth	cm	Signed byte	-100
ROOTS	obstacle to roots		Signed byte	-100
SWR	soil water regime		Signed byte	-100
T_TEXTURE	topsoil texture		Signed byte	-100

[a]The data type of Signed byte may not be legal in some programming language. However, It can be seen as character type, and the missing value is 156 (equivalent to -100 in short integer).

The 34 soil properties had 8 files for each of them. Each file contains the information of 1 layer. The depths to the bottom of the 8 layers are 4.5, 9.1, 16.6, 28.9, 49.3, 82.9, 138.3 and 229.6 cm. The name of each file end with a number to indicate the layer order from the top to the bottom. To get the correct value, you should multiply the values from the file with the scale factor.

Filename	description	units	Scale factor[a]	Data type[b]	Missing value
TC_1L~TC_8L	total carbon	% of weight	0.01	Short Integer	-999
OC_1L~OC_8L	organic carbon	of weight	0.01	Short Integer	-999
TN_1L~TN_8L	total N	% of weight	0.01	Short Integer	-999
TS_1L~TS_8L	total S	% of weight	0.01	Short Integer	-999
CACO3_1L~CACO3_8L	CaCO3	% of weight	0.01	Short Integer	-999
GYP_1L~GYP_8L	gypsum	% of weight	0.01	Short Integer	-999
PHH2O_1L~PHH2O_8L	pH(H2O)		0.1	Signed byte	-100
PHK_1L~PHK_8L	pH(KCl)		0.1	Signed byte	-100
PHCA_1L~PHCA_8L	pH(CaCl2)		0.1	Signed byte	-100
ECE_1L~ECE_8L	Electrical conductivity	ds/m	0.01	Short Integer	-999
EXCA_1L~EXCA_8L	Exchangeable calcium	cmol/kg	0.01	Short Integer	-999
EXMG_1L~EXMG_8L	Exchangeable magnesium	cmol/kg	0.01	Short Integer	-999
EXNA_1L~EXNA_8L	Exchangeable sodium	cmol/kg	0.01	Short Integer	-999
EXK_1L~EXK_8L	Exchangeable potassium	cmol/kg	0.01	Short Integer	-999
EXAL_1L~EXAL_8L	Exchangeable aluminum	cmol/kg	0.01	Short Integer	-999
EXH_1L~EXH_8L	Exchangeable acidity	cmol/kg	0.01	Short Integer	-999
CEC_1L~CEC_8L	Cation exchange capacity	cmol/kg	0.01	Short Integer	-999
BS_1L~BS_8L	Base saturation	%		Signed byte	-100
SAND_1L~SAND_8L	Sand content[c]	% of weight		Signed byte	-100

SILT_1L~SILT_8L	Silt content	% of weight		Signed byte	-100
CLAY_1L~CLAY_8L	Clay content	% of weight		Signed byte	-100
GRAV_1L~GRAV_8L	Gravel content	% of volume		Signed byte	-100
BD_1L~BD_8L	Bulk density	g/cm3	0.01	Short Integer	-999
VMC1_1L~VMC1_8L	Volumetric water content at -10 kPa	% of volume		Signed byte	-100
VMC2_1L~VMC2_8L	Volumetric water content at -33 kPa	% of volume		Signed byte	-100
VMC3_1L~VMC3_8L	Volumetric water content at -1500 kPa	% of volume		Signed byte	-100
PBR_1L~PBR_8L	The amount of phosphorous using the Bray1 method	ppm of weight	0.01	Short Integer	-999
POL_1L~POL_8L	The amount of phosphorous by Olsen method	ppm of weight	0.01	Short Integer	-999
PNZ_1L~PNZ_8L	Phosphorous retention by New Zealand method	% of weight	0.01	Short Integer	-999
PHO_1L~PHO_8L	The amount of water soluble phosphorous	ppm of weight	0.0001	Short Integer	-999
PMEH_1L~PMEH_8L	The amount of phosphorous by Mehlich method	ppm of weight	0.01	Short Integer	-999
ESP_1L~ESP_8L	exchangeable sodium percentage	% of weight	0.01	Short Integer	-999
TP_1L~TP_8L	Total phosphorus	% of weight	0.0001	Short Integer	-999
TK_1L~TK_8L	Total potassium	% of weight	0.01	Short Integer	-999

[a]The valuesshould be multiplied by the scale factor to get the target values.

[b]The data type of Signed byte may not be legal in some programming language. However, It can be seen as character type, and the missing value is 156 (equivalent to -100 in short integer).

[c]The sum of sand, silt and clay is not always 100 due to round-off errors. In most cases, it is ok to calculate one of them from the other two. Or, you can scale them to 100.

2.3 Coordinate system of the dataset

The coordinate system is WGS_1984, and the parameters are:
Semimajor Axis: 6378137.000000000000000000
Semiminor Axis: 6356752.314245179300000000
Inverse Flattening: 298.257223563000030000

2.4 Description of the codes for the soil general information

The codes of the soil general information are given in the following tables, which is the same as the HWSD (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012).

ADD_PROP (Additional Property)

Certain soil properties, inherent to the soil unit definition that are relevant for agricultural use of the soil are vertic, gelic and petric; the latter property refers to petric Calcisols and petric Gypsisols (FAO-90). The additional field provides details on Petric, Gelic Vertic properties.

ADD_PROP	
CODE	**VALUE**
0	None
1	Petric
2	Gelic
3	Vertic

Available water storage capacity in mm/m of the soil unit

For the soil units of the Soil Map of the World (FAO-74) and for the revised legend (FAO-90), FAO has developed procedures for the estimation of Available Water Capacity in mm/m (AWC) (FAO, 1995). The AWC classes have been estimated for all soil units of both FAO classifications accounting for topsoil textural class and depth/volume limiting soil phases.

AWC_CLASS	
CODE	**VALUE**
1	150
2	125
3	100
4	75
5	50
6	15
7	0

Soil drainage

Soil drainage is indicated by 7 classes:

- EXCESSIVE: Water is removed from the soil very rapidly. The soils are commonly very coarse textured or rocky, shallow or on steep slopes.

- SOMEWHAT EXCESSIVE: Water is removed from the soil rapidly. The soils are commonly sandy and very pervious.

-WELL: Water is removed from the soil readily but not rapidly. The soils commonly retain optimal amounts of moisture, but wetness does not inhibit root growth for significant periods.

-MODERATELY WELL: Water is removed from the soil somewhat slowly during some periods of the year. The soils are wet for short periods within the rooting depth. They commonly have an almost impervious layer or periodically receive heavy rainfall.

-IMPERFECTLY: Water is removed slowly so that the soil is wet at a shallow depth for significant periods. Soils commonly have an impervious layer, a high water table,additions of water by seepage or very frequent rainfall.

-POOR: Water is removed so slowly that the soils are commonly wet at a shallow depth for considerable periods. The soils commonly have a shallow water table which is usually the result of an almost impervious layer, seepage or very frequent rainfall.

-VERY POOR: Water is removed so slowly that the soils are wet at shallow depths for long periods. The soils have a very shallow water table and are commonly in level or depressed sites or have very high rainfall falling almost every day.

DRAINAGE	
CODE	**VALUE**
1	Very Poor
2	Poor
3	Imperfectly
4	Moderately Well
5	Well
6	Somewhat Excessive

IL (Impermeable Layer): Indicates the presence of an impermeable layer within the soil profile of the STU. The code is only available in ESDB.

IL	
CODE	**VALUE**
0	None
1	> 150cm
2	80-150 cm
3	40-80 cm
4	< 40 cm

NONSOIL

Recoding is the process of harmonizing different coding systems to a unique system. This was required for the coding of non-soil units and phases, which were different in the various source databases. For instance the table below illustrates the harmonized coding systems for *non-soil units* in the different soil classifications (FAO-74, FAO-85 and FAO-90). All non-soil units represented in the four source databases are listed and a new unique coding is applied in the harmonized database.

NONSOIL	
CODE	**VALUE**
-20	No data
-19	Inland water
-18	Urban
-17	Salt flats
-16	Rock debris
-15	No data due to soil map
-14	Island
-13	Humanly disturbed
-12	Glaciers & permanent snow
-11	Fishponds
-10	Dunes & shifting sands
1	soil

Note: -20 is no data because there is no information from the soil maps. -15 is no data due to the soil maps.

PHASE1 – PHASE2

Phases are subdivisions of soil units based on characteristics which are significant for the use or management of the land but are not diagnostic for the separation of the soil units themselves. Phases numbered 1 to 12 were used in the Soil Map of the World (FAO-74), phases 13 to 22 were used in association with the Revised Legend of the Soil Map of the World (FAO-90), while phases 23 to 30 are specific for the European Soil Database.

Two phases can be listed for each soil unit, in order of importance.

PHASE	
CODE	**VALUE**
1	Stony
2	Lithic
3	Petric

PHASE	
CODE	VALUE
4	Petrocalcic
5	Petrogypsic
6	Petroferric
7	Phreatic
8	Fragipan
9	Duripan
10	Saline
11	Sodic
12	Cerrado
13	Anthraquic
14	Gelundic
15	Gigai
16	Inundic
17	Placic
18	Rudic
19	Salic
20	Skeletic
21	Takyric
22	Yermic
23	Erosion
24	No limitation to agricultural use
25	Gravelly
26	Concretionary
27	Glaciers
28	Soils disturbed by man
29	Excessively drained
30	Flooded

ROOTS (Obstacle to Roots): provides the depth class of an obstacle to roots within the STU.

ROOTS	
CODE	**VALUE**
0	None
1	>80 cm
2	60-80 cm
3	40-60 cm
4	20-40 cm
5	0-80 cm

SWR (Soil Water regime): Indicates the dominant annual average soil water regime class of the soil profile of the STU. The code is only available in ESDB.

SWR	
CODE	**VALUE**
0	None
1	Wet: (0-80 cm) < 3 months; (0-40cm) < 1 month
2	Wet: (0-80 cm) 3-6 months; (0-40cm) < 1 month
3	Wet: (0-80 cm) > 6 months; 0-40 cm > 11 months
4	Wet: 0-40 cm > 11 months

T_TEXTURE (Topsoil texture class)

Topsoil textural class refers to the simplified textural classes for 0–30cm used in the Soil Map of the World (FAO/Unesco, 1970-1980). Because of the scale of the map (1:5 million) only three simplified textural classes were used.

T_TEXTURE	
CODE	**VALUE**
0	None
1	Coarse
2	Medium
3	Fine
4	None

3. Data Usage

The data in binary format can be easily used by many programming language. It is recommended to load the binary format for ArcGIS users (3.4). The data in NetCDF file format can be used by multiply software. Here we give three example softwares, i.e. Panoply, NCL, R and ArcGIS.

3.1 Panoply

This software is recommended to have a fast visual look at the data. It can be downloaded here (www.giss.nasa.gov/tools/panoply).

3.2 NCAR Command Language (NCL)

Here is an example of NCL script to use the data:

```
load "$NCARG_ROOT/lib/ncarg/nclscripts/csm/gsn_code.ncl"
load "$NCARG_ROOT/lib/ncarg/nclscripts/csm/gsn_csm.ncl"

begin

SAdata = addfile("PHK 1.nc","r")
lat = SAdata->lat
lon = SAdata->lon
SA = SAdata-> PHK
;printVarSummary(PHK)

PHK @_FillValue = -100

wks = gsn_open_wks("pdf"," PHK ")
gsn_define_colormap(wks,"rainbow+white+gray")

res   = True
res@gsnAddCyclic = False
res@mpLimitMode = "LatLon"
res@mpMaxLatF =-180
res@mpMinLatF = 180.0
res@mpMaxLonF = 83.0
res@mpMinLonF = -56.0

res@cnFillOn=True
res@cnLinesOn=False

res@lbLabelAutoStride=True
res@lbBoxLinesOn=False

res@gsnSpreadColors=True
res@gsnSpreadColorStart=50
res@gsnSpreadColorEnd=-3

res@cnFillMode = "RasterFill"
```

```
res@cnLevelSelectionMode="ManualLevels"
res@cnMinLevelValF=0.0
res@cnMaxLevelValF=90.0
res@cnLevelSpacingF = 5.0

plot = gsn_csm_contour_map(wks,SA(0,:,:),res)

end
```

Note that workspace reallocation would exceed maximum size 32556688, the easiest way to increase the size is to put a line like the following into your ~/.hluresfile:

```
*wsMaximumSize : 500000000
```

3.3 R language

The NetCDF files can be used by loading "RNetCDF" package, and the corresponding maps can be drawn by loading “raster” package. The following is an example, reading part of the data :

```
rm(list=ls(all=TRUE))
setwd("D:\\NC\\data") # The directory (windows) of NetCDF file
#setwd("/media/shanggv/data/data/soildata/all/datarealse/nc/nc") # The directory (Linux-like) of NetCDF file
library("RNetCDF")
library(raster)
cnfile<-"PHK1.nc"
q3<-open.nc(cnfile, write=FALSE)
print.nc(q3)

#an exmaple for a region(Southeast China):
#you may set the xmn, xmx, ymn,ymx to get the region you want.
xmn=110
xmx=120
ymn=20
ymx=30
coln= round((xmx-xmn)*120)  #column number in the region
rown=round((ymx-ymn)*120)  #row number in the region
lurow= round((84-ymx)*120)  #row number of left upper corner of the region
lucol=round((xmn-(-180))*120)   #column number of left upper corner of the region
r <-  raster(ncol=coln, nrow=rown,xmn=xmn, xmx=xmx,
             ymn=ymn, ymx=ymx)
#read value
tmp<-var.get.nc(q3,"PHK",c(lucol,lurow,1),c(coln,rown,1))
tmp <- tmp*0.1 #the scale foctor of PHK is 0.1 according to the download table
```

```
range(tmp,na.rm=T)
#plot maps
values(r)<-as.vector(tmp)
plot(r, asp=1)

#get a value at a specific location (take lon=110.002, lat=24.51 as an example)
#you need to set xmn, xmx, ymn,ymx to cover only one grid
#you need to set xmn<lon<xmx and ymn<lat<ymx
#note that 0.008333333 is the grid size
xmn=110
xmx=110+0.008333333
(24.51-20)/0.008333333 # you get 61.2 here. Set ymn,ymx according to this value
ymn=24+0.008333333*61
ymx=24+0.008333333*62
coln= round((xmx-xmn)*120)  #column number in the region
rown=round((ymx-ymn)*120)  #row number in the region
lurow= round((84-ymx)*120)  #row number of left upper corner of the region
lucol=round((xmn-(-180))*120)   #column number of left upper corner of the region
tmp<-var.get.nc(q3,"PHK",c(lucol,lurow,1),c(coln,rown,1)) # you may get a null value if there is no data
tmp <- tmp*0.1 #the scale foctor of PHK is 0.1 according to the download table
tmp
close.nc(q3)
```

3.4 ArcGIS

The binary files can be loaded into ArcGIS by changing into .bil file. Here we take the “T_TEXTURE” as an example. The following changes should be done:

a, change the name into “T_TEXTURE.bil”.

b, create an head file “T_TEXTURE.hdr”, edit it with a text editor containing the following:

```
BYTEORDER I
LAYOUT BIL
NROWS 21600
NCOLS 43200
NBANDS 1
NBITS 8
PIXELTYPE SIGNEDINT
ULXMAP -179.99583333333334
ULYMAP 89.99583333333334
XDIM 0.00833333333333
YDIM 0.00833333333333
```

For all signed byte data, the content is the same as above. For short integer data, the “NBITS” should be set to “16”. BYTEORDER may be set to “M” in some computers.

4. Citation

Details about the dataset are in the peer-reviewed paper. Full acknowledgement and referencing of all sources must be included in any documentation using any of the material contained in the Global Soil Dataset for Earth System Modeling, as follows:

Shangguan, W., Y. Dai, Q. Duan, B. Liu and H. Yuan (2014), A Global Soil Data Set for Earth System Modeling, Journal of Advances in Modeling Earth Systems, 6: 249-263.

5. Reference

FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012. Harmonized World Soil Database (version1.2), FAO, Rome, Italy and IIASA, Laxenburg, Austria.

6. Contact

If you have any questions or feedbacks when using the data sets, please email: shgwei@mail.sysu.edu.cn (Dr. Wei Shangguan).