Chapter 6
WorldGrids.org: a repository of global soil covariates

Hengl T., Reuter H.I. & Mendes de Jesus J.

Rationale for WorldGrids.org • WorldGrids.org as a public geoportal for global environmental layers (that stack) • common data sources for WorldGrids.org • standard processing steps — principal component analysis, resampling, aggregation • technical specifications (which type of data is it hosted on WorldGrids.org) • available layers (1 km) • Web Processing Functionality with code examples (overlay, subsetting and aggregating grids) • how to submit a layer to WorldGrids.org • file naming convention • metadata specifications • referent land, water and soil masks • missing functionality and future developments of WorldGrids.org

6.1 Introduction

6.1.1 Rationale for WorldGrids.org

WorldGrids.org is a repository of gridded predictors with global coverage or at least close to global coverage (covering >90% land mass). This is one of the major components of the GSIF system intended to serve global soil mapping applications, but also ISRIC's general repository of referent 1 km resolution environmental layers.

Building repositories of global environmental covariates i.e. geoportals is not a novel thing. In 1985, for example, United Nation started with the UNEP Global

Resource Information Database (GRID) initiative[1]. The objective of GRID was to provide data and information for decision and policy making. Likewise, the Global Change Master directory[2] (GCMD) from NASA contains over 25,000 metadata records which are accessible via the internet. Other similar academic projects focused on building repositories of global GIS layers are e.g. FAO's GeoNetwork[3], Environmental Information Portal of the World Resources Institute[4], Natural Earth Data website[5], GeoPortal.org, and Atlas of the Biosphere portal[6]. Large amount of global GIS data can also be browsed via data portals owned by big GIS / internet companies e.g. the ESRI's ArcGIS Online Explorer[7] and GeoCommons[8] and Google's Earth Engine[9].

Key agenda setters in the terms of production and dissemination of remote sensing and thematic layers include Google — in terms of spatial detail (public access to view on global mosaic of satellite images at a resolution of 2 m), NASA's MODIS mission — in terms of thematic content and usability; Germany's TanDEM-X radar mission[10] aiming at a 12 m resolution DEM (±2 m vertical accuracy) of the world by 2013, and the OneGeology project — in terms of promoting cross-national collaboration (Jackson and Wyborn, 2007).

At the level of continent, various similar initiatives exist that serve even finer resolution data. In Europe, an initiative called INSPIRE[11] was set to serve European environmental data. Large amount of environmental data is now also publicly available for the African continent, e.g. via the AfSIS — Africa soil information system. At country level there are even more such initiatives.

Unfortunately, continent level map repositories often can not be merged to produce global coverage layers. As Maguire and Longley (2005) conclude:

> *"Coordination of parallel initiatives must reconcile different technology standards, administrative schema and funding regimes. Not surprisingly… bottom-up approaches to building inter-organizational enterprise GIS have met with limited success."*

While looking at the existing geodata portals we can observe some disadvantages:

1. numerous data sets with different access and user rights exist,
2. different data types, coordinate systems, resolutions are used in a nonstandard way,

[1] `http://www.grid.unep.ch/`

[2] `http://gcmd.nasa.gov/`

[3] `http://www.fao.org/geonetwork/srv/en/main.home`

[4] `http://earthtrends.wri.org/`

[5] `http://www.naturalearthdata.com/downloads/` — compiled by Nathaniel Vaughn (KELSO) and volunteers.

[6] `http://www.sage.wisc.edu/atlas/maps.php`

[7] `http://www.arcgis.com/explorer/`

[8] `http://geocommons.com/`

[9] `http://earthengine.google.org/`

[10] `http://www.astrium-geo.com/en/2933-first-tandem-x-coverage-completed`

[11] `http://inspire-geoportal.ec.europa.eu/`

3. different methods of soil map generation are used,
4. many nominally global layers are fragmented and do not have a complete global coverage.

In many cases production processes used to create global GIS layers have been closed or are not documented, which makes it rather difficult to re-produce them or objectively assess quality. All this inspired us to build a truly Open and a truly global repository of environmental layers that can be used for global soil mapping and for sharing inputs and outputs of global soil mapping.

In the following sections we describe the content and functionality of the WorldGrids.org geoportal and explain how to contribute to the geoportal with your own data.

6.1.2 Data sources

Recall from page 48 that there are five main groups of soil covariates of interest for global soil mapping: (1) *Spectral and multispectral RS images*, (2) *DEM-derived covariates*, (3) *Climatic and vegetation based covariates*, (4) *Land survey and land use information*, and (5) *Expert-based covariates e.g. soil delineations*. WorldGrids.org hosts already majority of layers from this list, although one can find on WorldGrids.org also layers of general interest for any environmental analysis.

ISRIC facilitates collation and use of multi-thematic gridded repositories of global soil covariates (at standard resolutions of 1 km, 500 m, 250 m, and 5 km). The focus is put on five main groups of covariates: (1) spectral and multispectral RS images, (2) DEM-derivatives, (3) climatic and vegetation maps, (4) land survey and land use information, and (5) soil delineations.

WorldGrids.org focuses only on serving publicly available data and hence is also primarily based on public environmental data. Examples of publicly available global products that are freely available for use (as of 2011) are: NASA's SRTM DEM (RABUS *et al.*, 2003), the European Space Agency's GlobCover[12] v2 land cover map of the world, various MODIS land, atmosphere and ocean products (SAVTCHENKO *et al.*, 2004), the Defence Meteorological Satellite Program Nighttime Lights Time Series v4 (SMALL *et al.*, 2005), Worldclim.org a 1–km resolution repository of some 20 key climatic variables (HIJMANS *et al.*, 2005), the Gridded Population of the World, version 3 (GPWv3)[13], Global Forest Resources Assessment (FRA) and similar (Fig. 6.1).

Target resolution of WorlGrids in 2013 is 1 km, which is primarily for technical reasons. A problem of working with finer resolution data (100 m, 250 m) is

[12] `http://ionia1.esrin.esa.int`

[13] `http://sedac.ciesin.columbia.edu/gpw/`

that only a limited number of publicly available GIS layers with a global coverage are available at present: (1) SRTM DEM and derivatives, (2) 1:200k scale maps of water bodies, forest areas etc, and (3) Landsat ETM and ASTER images (Fig. 6.2). Because Landsat and ASTER images are of very fine resolution (30+ m), a significant effort is required to collate a global mosaic. For example, Landsat's **Global Land Survey**[14] 1990, 2000 and 2005 mosaics are publicly available via the Global Land Cover Facility[15], but the scenes need to be merged and harmonized before they can be used for soil mapping applications. Other problems of using Landsat ETM or MSS imagery are clouds and atmospheric artifacts that still need to be filtered before a >90% cloud free global image can be made available for global soil mapping.

Among the main data sources considered for WorldGrids, MODIS continues to be one of the most widely used global sources of remote sensing-based maps. SRTM DEM also remains an impressive source of scientific information with over 430 publications with the word "*SRTM*" in the article title based on Google Scholar. Both projects can be recognized as best-practice examples of how to design, process and distribute global data.

The problem of MODIS products, however, is that NASA and other collaborating organizations typically distribute only time-series compilations for coarser resolution products and all finer resolution products are available only as raw tiles. Our intention with the WorldGrids.org project is to invest significant resources to extract long-term spatial data sets that are of interest for global soil mapping. For example, daily Enhanced Vegetation Index image (EVI) derived by MODIS is of only minor utility for global soil mapping, but the long term values of EVI and their seasonal variability are definitely of interest for soil mapping because they help parameterize long-term influence of vegetation on soil formation.

6.1.3 Standard processing steps

WorldGrids.org soil covariates are derived out of some input data, which are often unprocessed, in-complete or not fit to be used for global soil mapping. Fig. 6.2 provides an overview of the main data sources and steps followed to produce the global soil covariates. Note that, most of the covariates available at WorldGrids.org, are original products derived by the WorldGrids.org developers. Nevertheless, some 20–30% of maps listed on the repository are simply reformatted/re-projected versions of the original grids, for example the Land cover (GlobCover v2) map of the world.

Each of the thematic groups on WorldGrids.org is dealt with as a separate sub-product and require a specific collection and processing strategy. The most typical procedures used to derived soil covariates are:

[14] `http://gls.umd.edu/`

[15] `ftp://ftp.glcf.umd.edu/glcf/Landsat/`

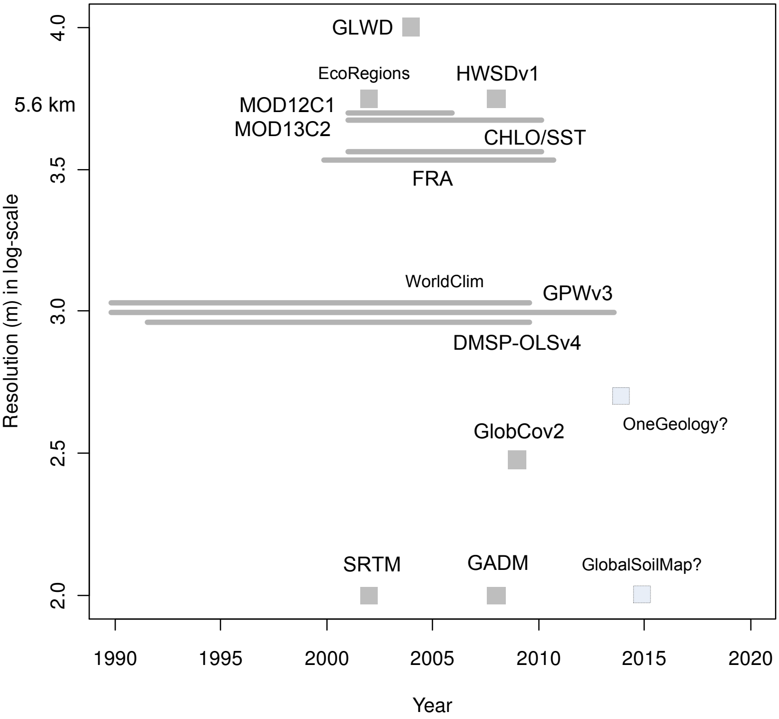


Fig. 6.1 Spatial resolution and temporal coverage of some widely used global data products: GLWD — Global Lakes and Wetlands Database, HWSD — Harmonized World Soil Database, MOD12C1 — MODIS Land Cover Type Yearly L3, MOD13C2 — Vegetation Indices Monthly L3, CHLO/SST — MODIS Aqua Level-3 annual Chlorophyll / mid-IR Sea Surface Temperature, FRA — Forest Resources Assessment, GPW — Gridded Population of the World, DMSP-OLS — Nighttime Lights Time Series, GlobCov — Land Cover classes based on the MERIS FR images, GADM — Global Administrative Areas.

(1) principal component analysis for time-series of images,
(2) resampling indicator values to memberships or percentages, and
(3) aggregating, generalizing maps with too much detail.

The quantity and complexity of the input imagery is often high and the processing can take significant resources. We make frequent use of the techniques such as the principal component analysis or resampling, especially to process the MODIS-derived products. For example the soil under vegetation mask map in Fig. 7.3 on page 192 was derived from a long term time series of Leaf Area Index images (120 of monthly images).

Eastman and Fulk (1993) have observed that principal component analysis is an attractive technique to analyze time-series of images and reduce their dimensionality. The first Principal Component (PC) of a long term monthly time-

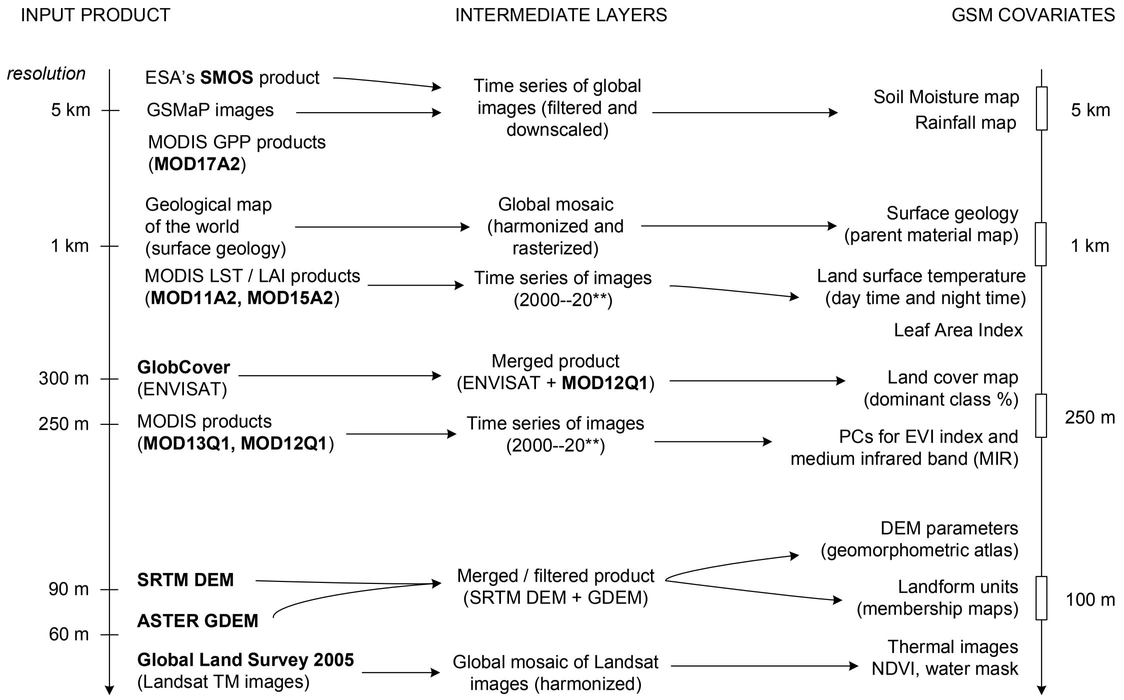

INPUT PRODUCT
INTERMEDIATE LAYERS
GSM COVARIATES
resolution
5 km
1 km
300 m
250 m
90 m
60 m
ESA's **SMOS** product
GSMaP images
MODIS GPP products (**MOD17A2**)
Geological map of the world (surface geology)
MODIS LST / LAI products (**MOD11A2, MOD15A2**)
GlobCover (ENVISAT)
MODIS products (**MOD13Q1, MOD12Q1**)
SRTM DEM
ASTER GDEM
Global Land Survey 2005 (Landsat TM images)
Time series of global images (filtered and downscaled)
Global mosaic (harmonized and rasterized)
Time series of images (2000--20**)
Merged product (ENVISAT + **MOD12Q1**)
Time series of images (2000--20**)
Merged / filtered product (SRTM DEM + GDEM)
Global mosaic of Landsat images (harmonized)
Soil Moisture map
Rainfall map
Surface geology (parent material map)
Land surface temperature (day time and night time)
Leaf Area Index
Land cover map (dominant class %)
PCs for EVI index and medium infrared band (MIR)
DEM parameters (geomorphometric atlas)
Landform units (membership maps)
Thermal images
NDVI, water mask
5 km
1 km
250 m
100 m

Fig. 6.2 Key input global GIS layers and intermediate and final products (DSM covariates).

series of MODIS EVI images, for example, represent clearly the mean biomass, while the following PCs show vegetation seasonality, but also different land use practice, vegetation succession and degradation. Example of a MODIS time-series derived products is shown in Fig. 6.3.

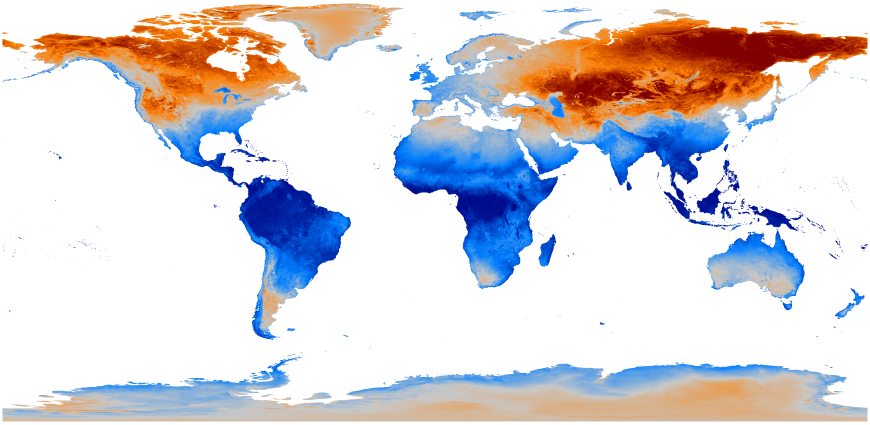


Fig. 6.3 Long-time deviation pattern of the Day-time MODIS Land Surface Temperature based on the monthly MODIS images. A covariate that can potentially explain soil-water content and soil mineralogy. This types of global GIS layers can not be found on the original NASA's MODIS distribution websites.

6.2 Available layers and functionality

6.2.1 Technical specifications

Three most important technical specifications of the GIS data available on the WorldGrids.org geoportal are:

1. The target resolution of the WorldGrids.org in 2013 is 1/120 decimal degrees in WGS84 coordinate system (about 1 km);
2. Each GIS layers is available as compressed GeoTIFF in 4–5 target resolutions (see Tbl. 6.1);
3. All processing steps carried to produce some GIS layer are fully documented and allow for reproducible research;

World map at resolution of 0.008333333 arcdegrees (1/120 decimal degrees or ca. 1 km) resolution is an image of size 43,200 × 21,600 pixels.

Each data set available at WorldGrids.org comes with accompanying metadata file, visualization specification (legend) and a processing script. The processing scripts are written for the R-project (extension *.R) or in Python (extension *.py) and might make use of several other packages. All processing scripts can be obtained via the Google Code project[16].

Table 6.1 Standard resolutions and corresponding scale numbers used by the WorldGrids.org.

Code	Decimal degrees	Resolution (km)	Scale number	Google scale level (1–19)	Columns	Rows
0	1/5	20	40M	5	900	1800
1	1/20	5.6	10M	6	3600	7200
2	1/40	2.5	5M	7	7200	14400
3	1/120	1	2M	8	21600	43200
4	1/240	.500	1M	9	–	–
5	1/600	.250	500K	10	–	–
6	1/1200	.100	200K	11	–	–
7	1/3600	.030	70K	12	–	–

Before publishing new data, each submitted layer is checked via the GDALinfo command for completeness and consistency. The submitted layers MUST closely match the listed resolutions and be projected in the WGS84 coordinate system to allow for stacking of multiple grids. This is an example of a correct header for a GeoTIFF[17] available on WorldGrids.org:

```
> GDALinfo("X:/WORLDGRIDS/maps/DEMSRE3a.tif")

rows        21600
columns     43200
bands       1
origin.x        -180
origin.y        -90
res.x       0.008333333
res.y       0.008333333
ysign       -1
oblique.x   0
oblique.y   0
driver      GTiff
projection  +proj=longlat +datum=WGS84 +no_defs
file        X:/WORLDGRIDS/maps/DEMSRE3a.tif
apparent band summary:
  GDType   Bmin  Bmax Bmean Bsd hasNoDataValue
1  Int16 -32768 32767     0   0          FALSE
  NoDataValue
1           0
Metadata:
AREA_OR_POINT=Area
```

[16] http://code.google.com/p/worldgrids/

[17] http://trac.osgeo.org/geotiff/

```
Warning message:
statistics not supported by this driver
```

The proj4 string is embedded in the file header and that the resolution is exactly 1/120 rounded to the 9th decimal place. World map at resolution of 0.008333333 arcdegrees (ca. 1 km) resolution is an image of size 43,200 × 21,600 pixels. Loading such large raster requires >8GB of RAM; their processing can take days or even weeks.

6.2.2 Available layers

Currently, WorldGrids.org contains some 50 data sets at 1 km resolution and over 100 layers at 5 km resolution. The 1 km resolution layers available at the moment of writing this book were:

I Climatic and meteorological images:

  i TDMMOD3: Mean value the 8-day MODIS day-time LST time series data;
  ii TDSMOD3: Standard deviation of the 8-day MODIS day-time LST time series data;
  iii TDLMOD3: Minimum value of the 8-day MODIS day-time LST time series data;
  iv TDHMOD3: Maximum value of the 8-day MODIS day-time LST time series data;
  v TNMMOD3: Mean value the 8-day MODIS night-time LST time series data;

II DEM-derived parameters:

  i DEMSRE3: Global Relief Model based on SRTM 30+ and ETOPO DEM;
  ii SLPSRT3: Slope map in percent derived using the DEMSRE3;
  iii L3POBI3: Physiographic landform units (SCALA project);
  iv TWISRE3: SAGA Topographic Wetness Index derived using the DEMSRE3;
  v OPISRE3: SAGA Topopgraphic Openess Index derived using the DEMSRE3;
  vi INMSRE3: Mean potential incoming solar radiation derived in SAGA GIS;
  vii INSSRE3: Standard deviation of the potential incoming solar radiation derived in SAGA GIS;

III MODIS products:

  i EVMMOD3: Mean value of the monthly MODIS EVI time series data;
  ii EVSMOD3: Standard deviation of the monthly MODIS EVI time series data;
  iii LAMMOD3: Mean value of the 8-day MODIS LAI time series data;
  iv LASMOD3: Standard deviation of the 8-day MODIS LAI time series data;

IV Land Cover and Land Use maps:

- i GLCESA3: Land Cover classes based on the MERIS FR images;
- ii G**ESA3: coverage percentages for classes from GLCESA3;
- iii GLCJRC3: Global Land Cover map for the year 2000 (GLC2000);

V Urbanization and Lights at night images:

- i LN1DMS: First principal component of the long-term lights at night images;
- ii LN2DMS: Second principal component of the long-term lights at night images;
- iii LNMDMS: Mean value of the long-term lights at night images;
- iv GACGEM: Global accessibility map;

VI Biodiversity and human impact maps:

- i IFLGRE: Intact forest landscapes based on Greenpeace;
- ii ECOWWF: Global Eco-floristic regions;

An up-to-date list of available layers and connected input data can be browsed at any time via:

`http://worldgrids.org/doku.php?id=wiki:layers`

6.2.3 Web Processing Functionality

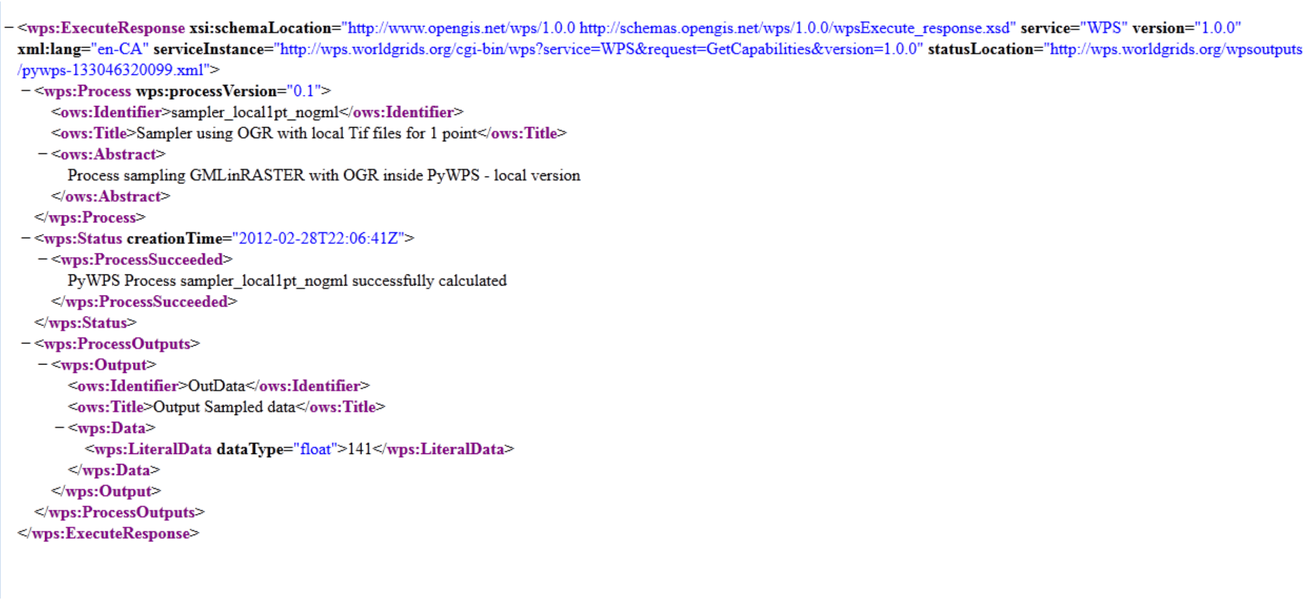
In addition to serving GeoTIffs, WorldGrids.org geoportal also provides some limited Web Processing Services: for point overlay, subsetting and aggregation of values. These methods have been implemented via the pyWPS[18] (OGC compliant) Open Source Web Processing Service (WPS) framework. We basically extend the methods available in the following packages: GDAL, Numpy and Scipy. To hide the complexity of the process as executed in the browser, wrapper functions have been developed for the different WPS functions (see GSIF package for R). The success of data distribution strongly depends on the simplicity of its access. Therefore, we first introduce the simplest methods available via a browser. In practice, we expect that most of the users will automate querying and operations through scripting or through some Graphical User Interface.

The simplest way to perform a given WPS method is the execution in a browser and the display of the resulting XML file (Fig. 6.4). A list of generic WPS functions and input parameters is available also via the GSIF package for R.

To overlay and subset some layer directly from R, we can start by loading all the required packages:

```
> library(XML)
> library(sp)
> library(GSIF)
```

[18] `http://pywps.wald.intevation.org/`



```
– <wps:ExecuteResponse xsi:schemaLocation="http://www.opengis.net/wps/1.0.0 http://schemas.opengis.net/wps/1.0.0/wpsExecute_response.xsd" service="WPS" version="1.0.0"
  xml:lang="en-CA" serviceInstance="http://wps.worldgrids.org/cgi-bin/wps?service=WPS&request=GetCapabilities&version=1.0.0" statusLocation="http://wps.worldgrids.org/wpsoutputs
  /pywps-133046320099.xml">
  – <wps:Process wps:processVersion="0.1">
      <ows:Identifier>sampler_local1pt_nogml</ows:Identifier>
      <ows:Title>Sampler using OGR with local Tif files for 1 point</ows:Title>
    – <ows:Abstract>
        Process sampling GMLinRASTER with OGR inside PyWPS - local version
      </ows:Abstract>
    </wps:Process>
  – <wps:Status creationTime="2012-02-28T22:06:41Z">
    – <wps:ProcessSucceeded>
        PyWPS Process sampler_local1pt_nogml successfully calculated
      </wps:ProcessSucceeded>
    </wps:Status>
  – <wps:ProcessOutputs>
    – <wps:Output>
        <ows:Identifier>OutData</ows:Identifier>
        <ows:Title>Output Sampled data</ows:Title>
      – <wps:Data>
          <wps:LiteralData dataType="float">141</wps:LiteralData>
        </wps:Data>
      </wps:Output>
    </wps:ProcessOutputs>
  </wps:ExecuteResponse>
```

Fig. 6.4 Example of the XML file displayed in a web browser and produced as a result of running the overlay method on the WPS server (value of the layer at 11.3E and 12.1N).

```
GSIF version 0.3-4 (2013-07-24)
URL: http://gsif.r-forge.r-project.org/
```

Table 6.2 Command Settings for the different Methods available via the WorldGrids.org WPS.

Method	Execute	Parameters
I (list server functionality)	`list`	–
II (1 point overlay)	`sampler_local1pt_nogml`	`x`, `y`, `inRastername`
III (multiple point overlay)	`sampler_local`	`inGML`, `inRastername`
IV (zonal statistics)	`overlay`	`inZone`, `inRastername`, `stype`
V (subset)	`subset`	`bbox`, `inRastername`

Next, we need to define the location of the repository and server settings:

```
> URI = "http://wps.worldgrids.org/pywps.cgi"
> server <- list(URI=URI, request="execute",
+     version="version=1.0.0", service.name="service=wps",
+     identifier="identifier=sampler_local1pt_nogml")
```

Where `inRastername` corresponds to the chosen Worldgrids.org data set, and the identifier:

```
identifier=sampler_local1pt_nogml
```

indicates WPS functionality of interest (point overlay without a GML input). Total list of methods available via `http://wps.worldgrids.org` is shown in Tbl. 6.2. Next, we can create an object of class `WPS`, that will emulate a virtual `Spatial` object from the sp package:

```
> glcesa3.wps <- new("WPS", server=server, inRastername="glcesa3a")
```

```
Loading required package: RCurl
Loading required package: bitops
```

This is basically a pointer to the WPS geoportal layer GLCESA3 (Land Cover classes based on the MERIS FR images), i.e. a virtual raster layer in our R session. We can now try to make connection with the server and test if the WPS is available:

```
> prl <- getProcess(glcesa3.wps)
> prl[7]
```

```
overlay TIFF and report statistics
                         "overlay"
```

This confirms that a function overlay is available. We can test overlaying some point (e.g. at lon=15, lat=15):

```
> p1 <- data.frame(lon=15, lat=15)
> coordinates(p1) <- ~lon+lat
> proj4string(p1) <- CRS("+proj=longlat +datum=WGS84")
> p1
```

```
SpatialPoints:
     lon lat
[1,]  15  15
Coordinate Reference System (CRS) arguments: +proj=longlat
+datum=WGS84
```

which we overlay by using the sp package default over operation as with any other Spatial object:

```
> over(glcesa3.wps, p1)

[1] "200"
```

in this case the number 200 indicates the class "Bare areas", which is also available via the layer description file available on the WorldGrids.org server (Fig. 6.5):

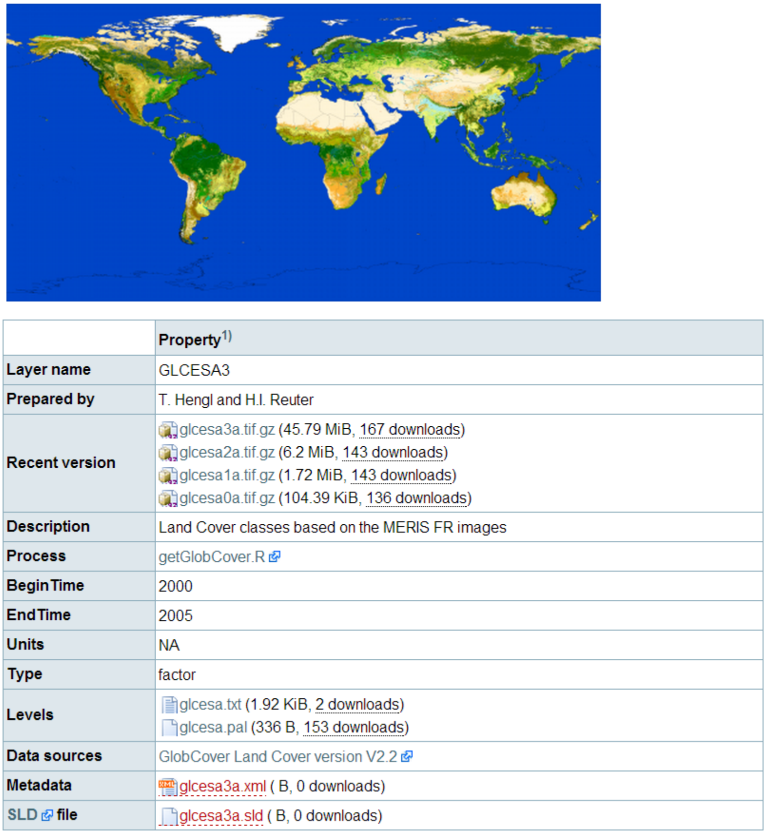
```
> cls <- read.table("http://worldgrids.org/lib/exe/fetch.php?media=glcesa.txt",
+    sep="\t", header = TRUE)
> cls[cls$DESCRIPTION=="CL200",]

      COLOR       NAME DESCRIPTION MINIMUM MAXIMUM
20 14153215 Bare areas       CL200   190.1   200.1
```

Similar calls would be executed for any of the other methods where the identifier for the EXECUTE parameter and the data inputs would change accordingly (Tbl. 6.2).

We can also fetch some smaller block of data by specifying the bounding box of interest:

```
> glcesa3 <- subset(glcesa3.wps, bbox=matrix(c(20,40,22,42), nrow=2))

trying URL 'http://wps.worldgrids.org/wpsoutputs/OutData-18037IVw2Ob.tif'
Content type 'image/tiff' length 65912 bytes (64 Kb)
opened URL
downloaded 64 Kb

glcesa3a_20_40_22_42.tif has GDAL driver GTiff
and has 240 rows and 240 columns

> str(glcesa3@data)

'data.frame':   57600 obs. of  1 variable:
 $ glcesa3a: int  30 14 20 50 30 30 30 20 14 14 ...
```

which is a very fast and efficient way to load the data of interest to an existing R session for some limited extent.



	Property[1]
Layer name	GLCESA3
Prepared by	T. Hengl and H.I. Reuter
Recent version	glcesa3a.tif.gz (45.79 MiB, 167 downloads) glcesa2a.tif.gz (6.2 MiB, 143 downloads) glcesa1a.tif.gz (1.72 MiB, 143 downloads) glcesa0a.tif.gz (104.39 KiB, 136 downloads)
Description	Land Cover classes based on the MERIS FR images
Process	getGlobCover.R
BeginTime	2000
EndTime	2005
Units	NA
Type	factor
Levels	glcesa.txt (1.92 KiB, 2 downloads) glcesa.pal (336 B, 153 downloads)
Data sources	GlobCover Land Cover version V2.2
Metadata	glcesa3a.xml (B, 0 downloads)
SLD file	glcesa3a.sld (B, 0 downloads)

Fig. 6.5 Example of a GIS layer on WorldGrids. See Tbl. 6.1 for explanation of scale codes.

6.3 Submission requirements

6.3.1 Minimum requirements

ISRIC will not limit the list of covariates in the repository but will instead keep this repository open. The minimum specifications for each submitted layer area as follows:

1. Full coverage with <5% of missing pixels for the domain of interest (land or water mask);
2. Saved in the GeoTIFF format with proj4 string embedded in the file header;
3. Projected in the WGS84 coordinate system with the bounding box covering the whole extent (longitudes: -180 to 180, latitudes: -90 to 90)
4. GeoTIFF file accompanied with:

  - a metadata file in the XML format and following the ISO19139 standard i.e. it must pass the metadata validity check;
  - a styled Layer Description (SLD) file i.e. the color legend and/or class names;
  - a process description file (R, Python or similar script);

5. The GeoTIFF file, metadata file and the SLD files with the same name;
6. A unique assigned name following the WorldGrids.org naming convention outlined below (all WorldGrids.org layers require a unique name);

Detailed instructions how to prepare a layer for WorldGrids.org and attach documentation is available on the geoportal DokuWiki.

The GeoTIFF files can be further compressed by using e.g. the GZIP compression. To compress rasters in R, you can consider using the 7z programme. First download the 7za program file in the working directory, then run e.g.:

```
> system("7za a -tgzip -mx9 DEMSRE3a_P24.tif.gz DEMSRE3a_P24.tif")

7-Zip (A) 9.20  Copyright (c) 1999-2010 Igor Pavlov  2010-11-18
Scanning
Creating archive DEMSRE3a_P24.tif.gz
Compressing  DEMSRE3a_P24.tif
Everything is Ok
```

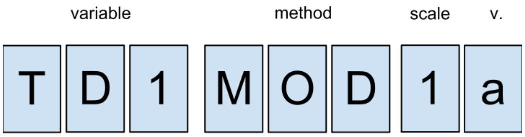
where -mx9 arguments sets the compression level to maximum.

6.3.2 File naming convention

All layers listed on the WorldGrids.org follow a standardized naming convention and data format. Most of GSIF columns follow a standard 8.3 filename convention[19] with at most eight characters:

- First three letter for the variable type: TD1 (daily surface temperature PC1);
- The next three letters represent the data source or collection method: MOD (MODIS day time imagery);
- The 7th character is the effective scale: 1 to 7;
- The 8th letter is the product version number (a for v1, b for v2, c for v3 etc.);

This is an example of a file name for principal component (PC1) of the land surface temperature derived from the MODIS time series data at 5 km resolution:



[19] http://en.wikipedia.org/wiki/8.3_filename

6.3.3 Metadata

Each dataset is described in ISRICs Metadata system under the catalog WorldGrids.org with a full ISO 19139 record. ISO19139 is the XML implementation schema for ISO19115 specifying for example details like data type, abstract, point of contact, descriptive keywords, temporal and spatial extent, links to processing description, data download, data quality.

To prepare your metadata, consider using some of the freely available Metada tools[20] provided kindly by the the Federal Geographic Data Committee. Before submitting the metadata, it is advisable to run a validity check[21] to pick up any missing information which will greatly speed up submission acceptance times.

6.3.4 Land and water referent maps and administrative boundaries

One key initial activity for global soil mapping is to determine the spatial domain of interest — the mask map representing all soil areas in the world. For example, it has been agreed by the *GlobalSoilMap* consortium that no predictions of soil properties will be made for grid cells that are considered to be occupied wholly or dominantly (>50%) by non-soil materials. Non-soil materials include permanent water and ice, bare rock and permanently sealed surfaces (urban areas and pavements).

Although this seems to be trivial, delineation of global soil and non-soil areas might require significant resources. World soil mask is not exactly equivalent to a land mask because not all potential soil areas represent areas of productive soils capable of supporting vegetative growth. In addition, rules for distinguishing between actual and potential soil areas differ from country to country. Hence the first step in defining the domain of interest is to specify rules to classify a pixel as soil or non-soil (see further section 7.3.2 on page 189). It is also important to consider that different soil prediction models will likely have different domains of applicability as soils develop differently under different environmental settings. Consequently, it is useful to distinguish between land surface areas that show significantly different soil forming processes and that have been sampled using different concepts (or omitted from sampling completely).

For practical purposes WorldGrids.org hosts a number of official mask maps representing a particular spatial domain of interest for soil mapping. Such referent maps would become important as they prevent us from extrapolating in areas where either not enough field samples are available or the models show distinct jumps. For example to use the same model developed over moderate climates to

[20] `http://www.fgdc.gov/metadata/geospatial-metadata-tools`

[21] `http://geo-nsdi.er.usgs.gov/validation/`

predict organic carbon content over Sahara would probably not work because under extreme conditions soil-forming processes are completely different.

ISRIC distributes and maintains a number of global land and water masks (land mask, soil mask, soil with vegetation cover, peatlands and arable land). These serve as global reference for any global calculus and analysis.

The following global land and water masks are of especial interest:

- **Land mask**: The total land area, defined as the total area not covered by permanent water. This mask map can be determined, for example, by using the Global Administrative boundaries database (GADM[22]), European Space Agency's GlobCover v2[23] land cover map of the world and/or Global Self-consistent, Hierarchical, High-resolution Shoreline Database GSHHA[24].
- **Water mask**: Water mask shows areas covered by water. Oceans and seas are relatively easy to map, but shallow lakes and rivers are more tricky. Water masks also change from year to year, so they need to be regularly updated. The suggested input for derivation of water mask is the MODIS MOD44W product[25].
- **Soil mask**: This is the total land area with soil materials deemed to be potentially or actually productive. This mask map includes areas identified as glaciers, deserts, swamps and similar areas that are not currently productive soils, but could theoretically be converted to productive areas (see further section 7.3.2 on page 189).
- **Soils with vegetation cover**: These are areas which can be demonstrated to be capable of supporting the growth of plants. An estimate of the extent of productive soils can be produced by using, for example, the MODIS long-term series of Leaf Area Index images. The soil productive areas mask map illustrated in Fig. 7.3 was derived by identifying all land areas with a Leaf Area Index >0 at any time since the beginning of the MODIS monitoring project (2000–2010).
- **Peatlands mask**: Peatlands are defined as areas covered with peat ($>20\%$ organic carbon and at least 50 cm deep). Peat forms in wetland bogs, moors, muskegs, mires, and peat swamp forests. According to the International Peat Society[26], approximately 60% of the world's wetlands are peat, and about 2% of the land surface is covered with peatlands, hence it is important to map the peatland areas accurately.
- **Arable land mask**: These are defined as areas that are used for agricultural crop production i.e. soil areas that are cultivated for production of food or fibre either manually or using mechanized systems (tractors). This is basically

[22] `http://www.gadm.org`

[23] `http://ionia1.esrin.esa.int`

[24] `http://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html`

[25] `http://glcf.umd.edu/data/watermask/`

[26] `http://www.peatsociety.org`

an inventory of all agricultural croplands. See for example the global map of croplands produced by the SAGE Research Center of the Nelson Institute for Environmental Studies at the University of Wisconsin-Madison[27].

Note that, from the map shown in Fig. 7.3 it is questionable whether there is any point presently in mapping soils at latitudes beyond 65° South or 84° North. Nevertheless, WorldGrids.org focuses on the global extent, then leaves the user to decide which mask to use and how to subset.

For administrative boundaries, we advise using one of the two sources of vector data:

- GADM database of Global Administrative Areas and/or
- vector and raster map data at 1:10M scales from the Natural Earth website.

But are publicly available and allow matching of spatial patterns by referent administrative boundaries. For more detailed visualizations we advise users to download and use the vector data that can be obtained from the OpenStreetMap geoportal.

6.4 Summary points

WorldGrids.org is possibly the most comprehensive public repository of covariates of interest for global soil mapping. Unfortunately, not all soil covariates required to fit significant spatial prediction models are available globally (e.g. airborne radiometric data) and many are available at only coarse resolution. Variety of available global covariates typically decreases with spatial detail i.e. finer and finer resolution. At the moment, most global environmental layers are available at resolutions of 1–5 km, i.e. at scales <1:1M. On the other hand, it appears that there is a tremendous amount of publicly available data waiting to be used for global mapping projects. From WorldGrids.org a variety of publicly available maps can be obtained at no cost at resolutions of 1 km or better, so that global soil mappers can focus more on building prediction models and producing data products. A number of referent land, water and soil mask can also be obtained and used to ensure that predictions produced by various teams match the same referent system. The issue for mapping teams is no longer whether to use this data, but where and how to obtain it, how to load it to an existing GIS and where to find necessary metadata.

A problem of switching to better resolutions than 1 km is that the computing time and storage jumps exponentially with a linear decrease in pixel size. For example, to go from 5.6 km to 1 km resolution requires 25+ times more resources (more storage capacity, more processing power, longer programming). The world at 5.6 km can be represented using ca. 25 million pixels; while the 1 km resolution worldgrids are 43,200 × 21,600 or 933 million pixels. Processing and visualization

[27] `http://www.sage.wisc.edu`

of such images is beyond capacities of a standard PC. This means that it will take a significant amount of time and effort before all predictors listed in Fig. 6.1 can be made available for public use.

WorldGrids.org is a multiscale repository of soil covariates. Son Jin Park[28], for example, has shown using the data in Africa that different covariates actually display spatial structure over different distances, and so it is advisable to use covariates at resolutions that correspond to their main observed spatial structure. For example climatic variables are spatially correlated over the large areas, while topography, parent material and hydrological processes operate at local scales of few hundreds of meters. These principles can be incorporated into global soil mapping so that, at each scale, we deliberately focus on different soil predictors (see previously Fig. 1.6).

Future developments of WorldGrids.org components will likely focus on the following aspects:

1. Versioning system: WorldGrids.org should have a versioning system that allows the user to trace which data set has been used in the production.
2. File traffic tracking system: WorldGrids.org needs to tracks number of downloads so that resources can be dedicated to the most used layers.
3. User-friendly export methods: we need to improve the ease of data exporting. At the moment we do support GeoTIFF export, however further data formats should be also possible. The question also remains open if any additional services like OpenDAP or THREDDS are required.
4. Full data processing automation: the processing steps should ideally be completely automated so that one can extend the structure of the portal to allow a seamless transition between core data sets and remote re-sources.
5. Server stability optimization: appropriate resourcing will need to be set to ensure the stability of the portal, its adequate perfor-mance under the expected access patterns and its scalability.
6. Active and large user and developers community: the role of ISRIC is to build up a community of users and stakeholders. Such public data portals can only survive if the community of active users continues to grow.

In the near future, we aim at collecting and processing the state of the art meteorological, vegetation, geology and ecosystem dynamics indices produced by European, USA and Japanese satellite agencies. For example, an important direct estimate of surface (0–5 cm) soil moisture is the ESA's Soil Moisture product (BARRÉ *et al.*, 2008). This product is available at relatively high temporal resolution, but then at a coarse spatial resolution (35–45 km) and needs to be processed to filter the inaccuracies, and downscaled to 5 km resolution using the water/land masks from supplementary MODIS images.

Also note that, the regression-kriging model explained in Eq.(4.13) is handicapped because the way in which the explanatory variables appear in the trend is highly empirical, i.e. it does not reflect the actual physical processes. In later

[28] Department of Geography, Seoul National University.

phases of collection of covariates the *GlobalSoilMap* team proposes also to consider developing *soil–process based covariates* — simulated estimates of soil processes such as soil erosion, flooding and deposition processes, carbon and nitrogen fixation processes and similar — by implementing various dynamic models on a global scale. Finally, we may eventually also consider building actual dynamic models to explain the distribution of the target soil properties, so that the results of predicting pH or organic carbon using dynamic models will only need to be calibrated versus the real data.