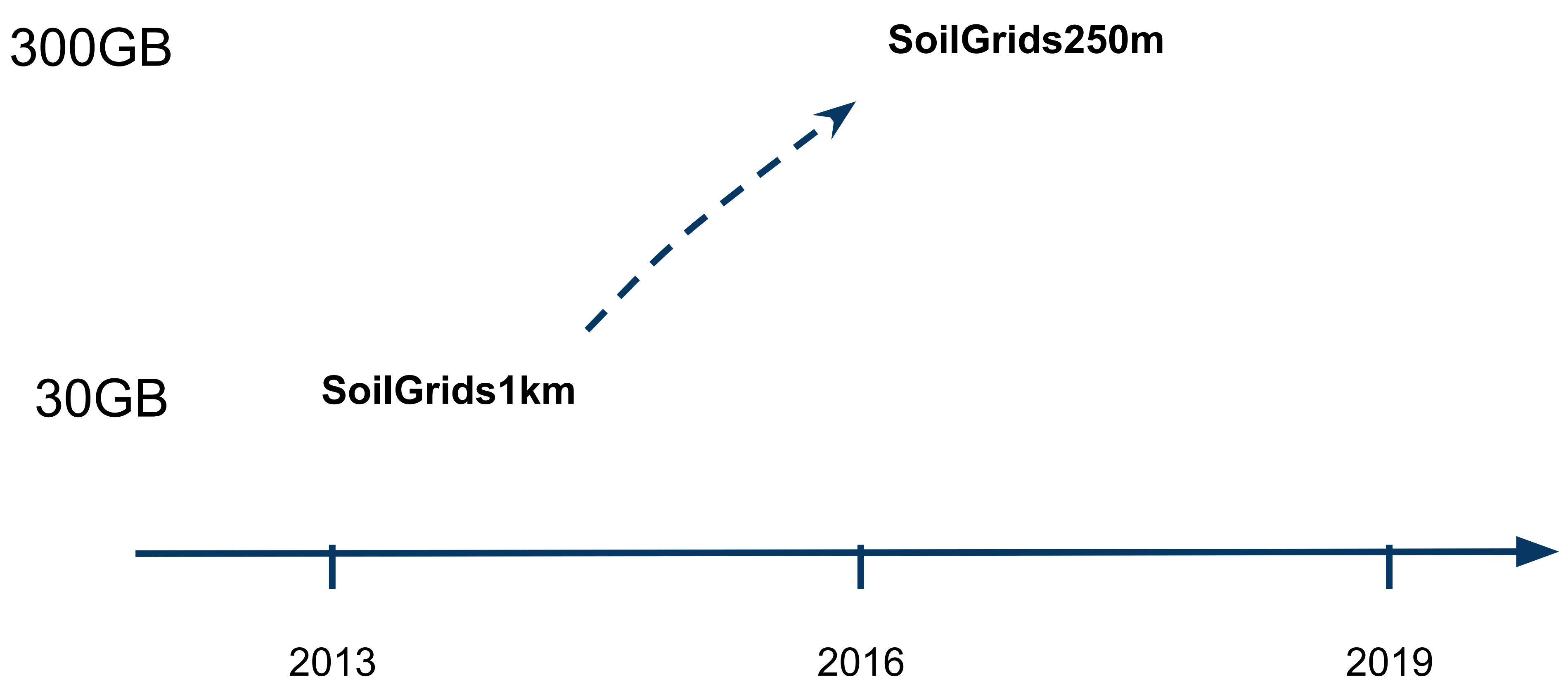
Automated soil mapping based on Machine Learning: towards a soil data revolution

Presented at the DSM 2016 conference

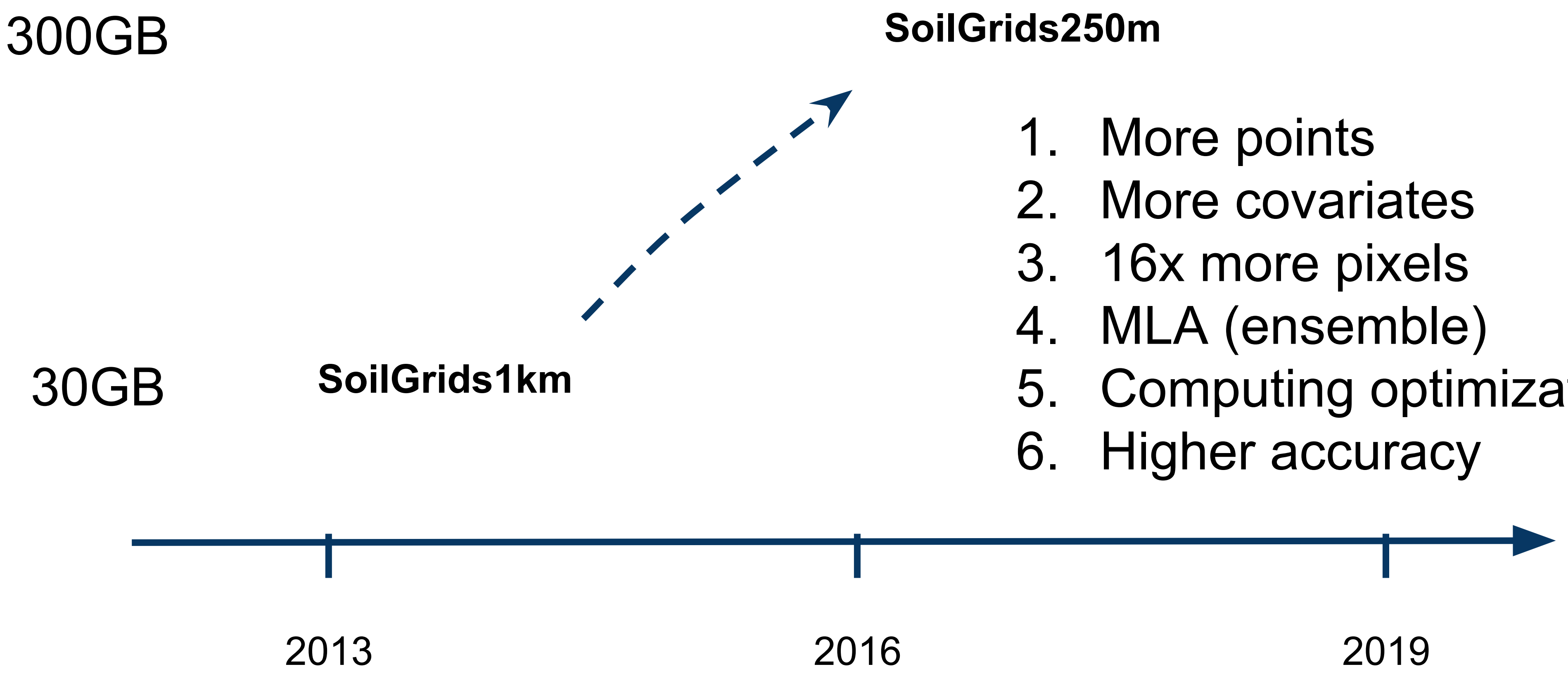
ISRIC World Soil Information

T. (Tom) Hengl <tom.hengl@isric.org>

Since 2013, we have been revising SoilGrids



The new system:



1. More points
2. More covariates
3. 16x more pixels
4. MLA (ensemble)
5. Computing optimization
6. Higher accuracy

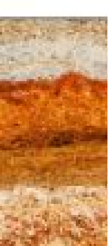
SoilGrids250m: global gridded soil information based on Machine Learning

Tomislav Hengl[1], Jorge Mendes de Jesus[1], Gerard B.M. Heuvelink[1], Maria Ruiperez Gonzalez[1], Milan Kilibarda[2], Aleksandar Blagotić[3], Wei Shangguan[4], Marvin N. Wright[5], Xiaoyuan Geng[6], Bernhard Bauer-Marschallinger[7], Mario Antonio Guevara[8], Rodrigo Vargas[8], Robert A. MacMillan[9], Niels H. Batjes[1], Johan G.B. Leenaars[1], Eloi Ribeiro[1], Ichsani Wheeler[10], Stephan Mantel[1], and Bas Kempen[1]

[1]ISRIC — World Soil Information, Wageningen, the Netherlands
[2]Faculty of Civil Engineering, University of Belgrade, Serbia
[3]GILab Ltd, Belgrade, Serbia
[4]College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
[5]Institut für Medizinische Biometrie und Statistik, Lübeck, Germany
[6]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada
[7]Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria
[8]University of Delaware, Newark DE, USA
[9]LandMapper Environmental Solutions Inc., Edmonton, Canada
[10]Envirometrix Inc., Wageningen, the Netherlands

Correspondence to: T. Hengl (tom.hengl@isric.org)

Abstract. This paper describes the technical development and accuracy assessment of the most recent and improved version of the SoilGrids system at 250 m resolution (June 2016 update). SoilGrids provides global predictions for standard numeric soil properties (organic carbon, bulk density, Cation Exchange Capacity (CEC), pH, soil texture fractions and coarse fragments) at seven standard depths (0, 5, 15, 30, 60, 100 and 200 cm), in addition to predictions of depth to bedrock and distribution of soil classes based on the World Reference Base (WRB) and USDA classification systems (ca. 280 raster layers in total). Predictions were based on ca. 150,000 soil profiles used for training and a stack of 158 remote sensing-based soil covariates (primarily derived from MODIS land products, SRTM DEM derivatives, climatic images and global landform and lithology maps), which were used to fit an ensemble of machine learning methods — random forest and gradient boosting and/or multinomial logistic regression — as implemented in the R packages ranger, xgboost, nnet and caret. The results of 10–fold



SOILGRIDS

A system for automated global soil mapping

www.soilgrids.org

ISRIC

World Soil Information

SoilGrids are heavily based on:

- → OpenStreetMap
- → OpenWeatherMap
- → Wikipedia
- → Global Biodiversity Information Facilities

Important info about SoilGrids

1. Open Data license
2. Updatable maps
3. Code on Github (reproducibility)
4. Diversity of access (FTP, WCS, REST API)
5. ... moving towards crowdsourcing

1. Open Data license

http://opendatacommons.org/licenses/odbl/summary/

ODC Open Database License (ODbL) Summary

This is a human-readable summary of the ODbL 1.0 license. Please see the disclaimer below.

You are free:

To Share: To copy, distribute and use the database.

To Create: To produce works from the database.

To Adapt: To modify, transform and build upon the database.

As long as you:

Attribute: You must attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.

Share-Alike: If you publicly use any adapted version of this database, or works produced from an adapted database, you must also offer that adapted database under the ODbL.

Keep open: If you redistribute the database, or an adapted version of it, then you may use technological measures that restrict the work (such as DRM) as long as you also redistribute a version without such measures.

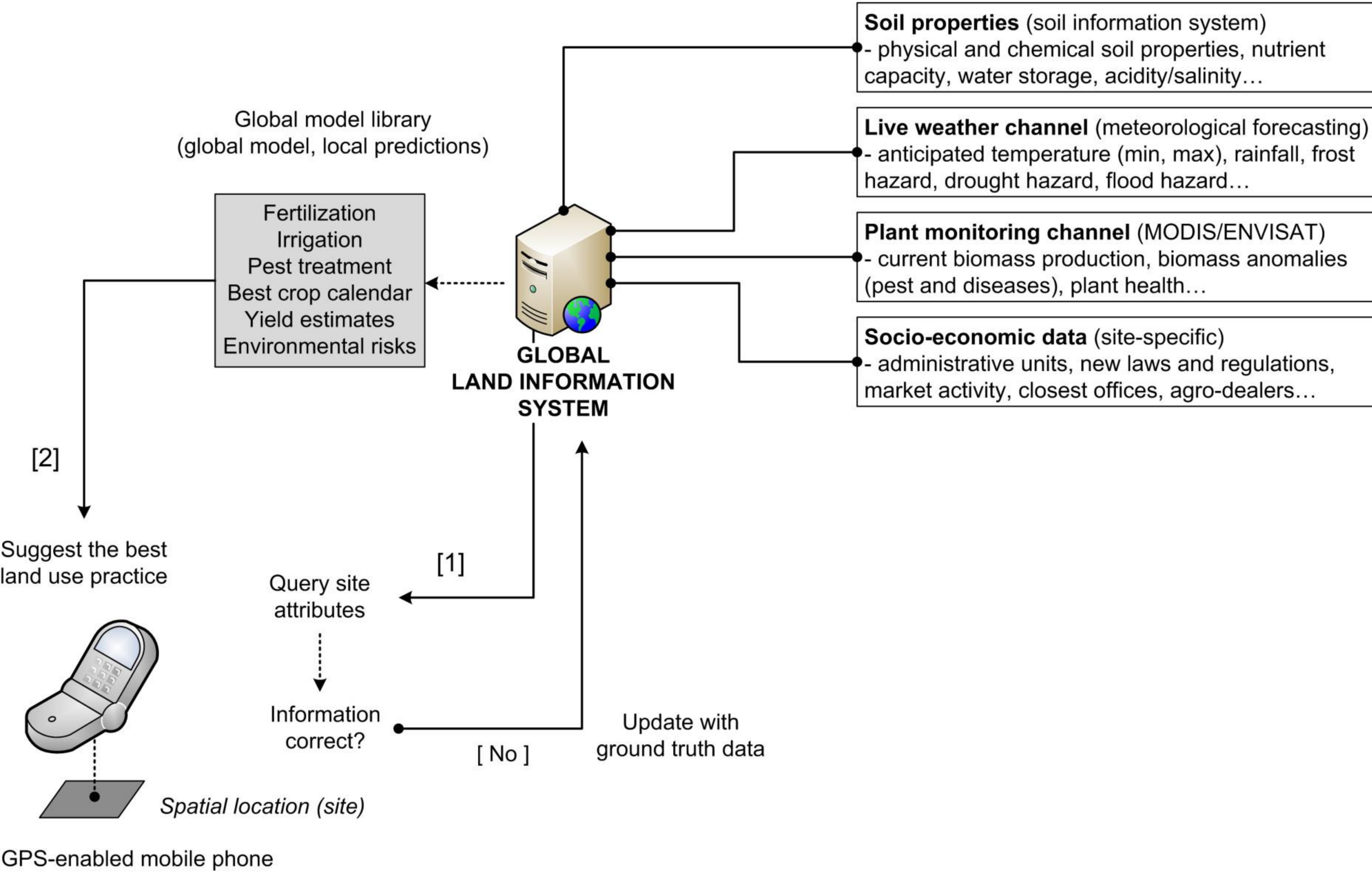
Disclaimer

This is not a license. It is simply a handy reference for understanding the ODbL 1.0 — it

MORE INFORMATION

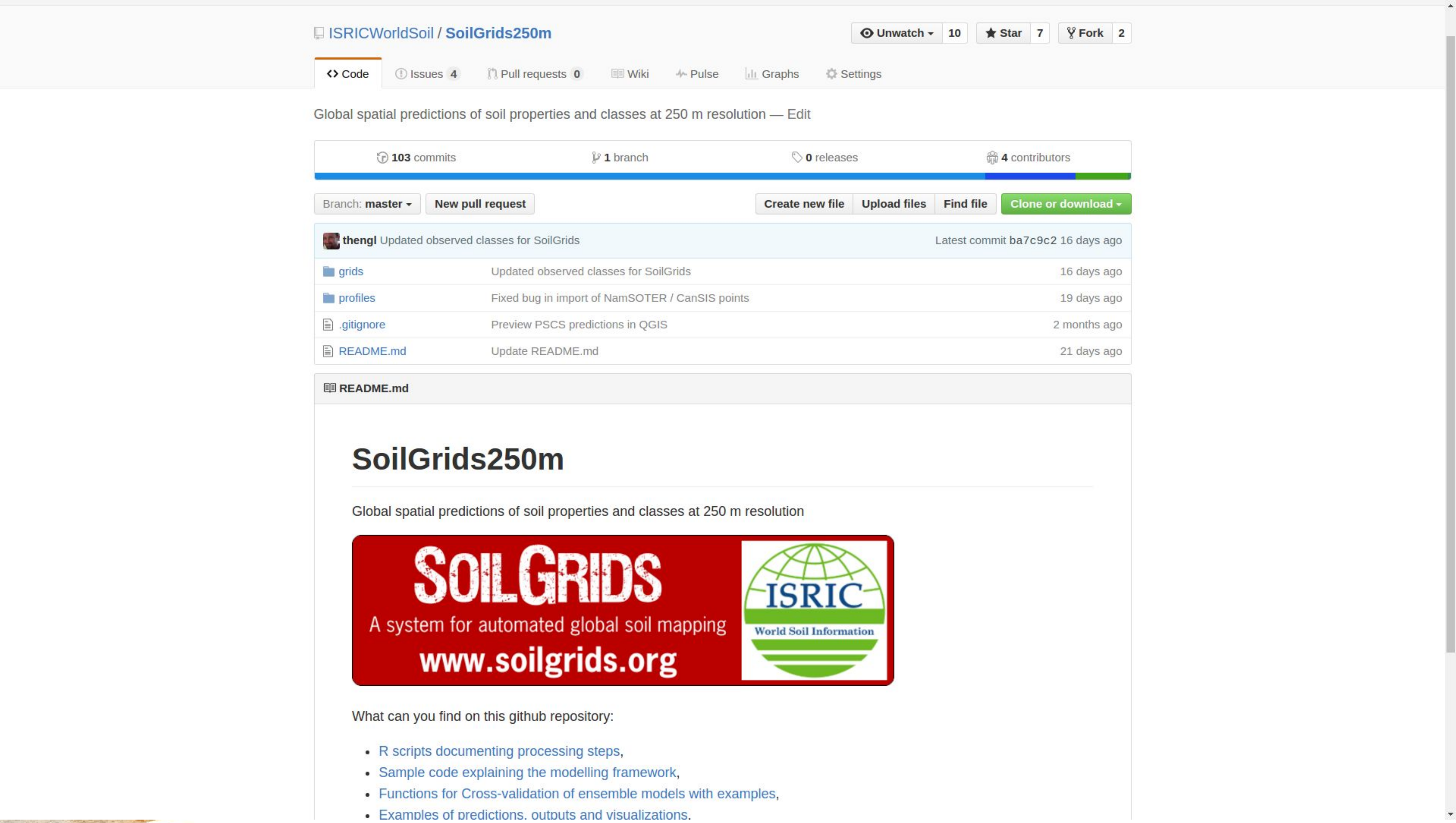
- Introduction to Open Data
- Open Definition for Data
- Quick guide to making data open
- Open Data Handbook

2. Versioning (automated mapping system)


Soil properties (soil information system)
- physical and chemical soil properties, nutrient capacity, water storage, acidity/salinity…
Live weather channel (meteorological forecasting)
- anticipated temperature (min, max), rainfall, frost hazard, drought hazard, flood hazard…
Plant monitoring channel (MODIS/ENVISAT)
- current biomass production, biomass anomalies (pest and diseases), plant health…
Socio-economic data (site-specific)
- administrative units, new laws and regulations, market activity, closest offices, agro-dealers…
Global model library
(global model, local predictions)
Fertilization
Irrigation
Pest treatment
Best crop calendar
Yield estimates
Environmental risks
GLOBAL
LAND INFORMATION
SYSTEM
[2]
[1]
Suggest the best land use practice
Query site attributes
Information correct?
[No]
Update with ground truth data
Spatial location (site)
GPS-enabled mobile phone

3. Reproducibility / open code

https://github.com/ISRICWorldSoil/



ISRICWorldSoil / **SoilGrids250m**

Unwatch 10 | Star 7 | Fork 2

Code | Issues 4 | Pull requests 0 | Wiki | Pulse | Graphs | Settings

Global spatial predictions of soil properties and classes at 250 m resolution — Edit

103 commits | **1** branch | **0** releases | **4** contributors

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

thengl Updated observed classes for SoilGrids — Latest commit ba7c9c2 16 days ago

grids	Updated observed classes for SoilGrids	16 days ago
profiles	Fixed bug in import of NamSOTER / CanSIS points	19 days ago
.gitignore	Preview PSCS predictions in QGIS	2 months ago
README.md	Update README.md	21 days ago

README.md

SoilGrids250m

Global spatial predictions of soil properties and classes at 250 m resolution

SOILGRIDS
A system for automated global soil mapping
www.soilgrids.org
ISRIC
World Soil Information

What can you find on this github repository:

- R scripts documenting processing steps,
- Sample code explaining the modelling framework,
- Functions for Cross-validation of ensemble models with examples,
- Examples of predictions, outputs and visualizations.

SoilGrids inputs:

- ➔ **ca 150,000 points ("World's largest" compilation of soil profile / soil sample data sets)** based on national and international datasets from over 45 countries.
- ➔ **40TB repository of MODIS land products, climatic images, DEM derivatives, geological and geomorphological data** (all at 250 m resolution)
- ➔ ISRIC's international network that can cross-check and validate spatial prediction patterns / values.

Data holdings in WoSIS 2

(December 2015)

- About 98,000 unique profiles
- Some 76,000 profiles are georeferenced within defined limits
- Number of measured data for each property varies between profiles with depth, generally depending on the purpose of the initial studies
- Source data based on diverse (inter)national standards
- Generally, limited quality information provided with the source (analytical) data

Lineage:
- Datasets, reports & maps

Soil observations and measurements:
- Feature (georeferenced profiles & layers)
- Attribute (x-y-z-t, map, class, site, layer-field, layer-lab)
- Method
- Value, including units of expression

SoilGrids are possible mainly thanks to:

NASA

NRCS
Natural Resources Conservation Service

USGS
science for a changing world

And thanks to: AfSIS project

Bill & Melinda Gates Foundation

Business Operation

Bill & Melinda Gates Foundation is one of the largest private foundations in the world, founded by Bill and Melinda Gates. It was launched in 2000 and is said to be the largest transparently operated private foundation in the world.
Wikipedia

Nonprofit category: Private Grantmaking Foundations

Founded: 2000

Assets: 36.79 billion USD (2010)

Income: 53 billion USD (2010)

Founders: Melinda Gates, Bill Gates

Also thanks to:

Center for International Forestry Research

CIFOR

Woods Hole Research Center

UN-REDD PROGRAMME

Food and Agriculture Organization of the United Nations

UNDP

Empowered lives. Resilient nations.

UNEP

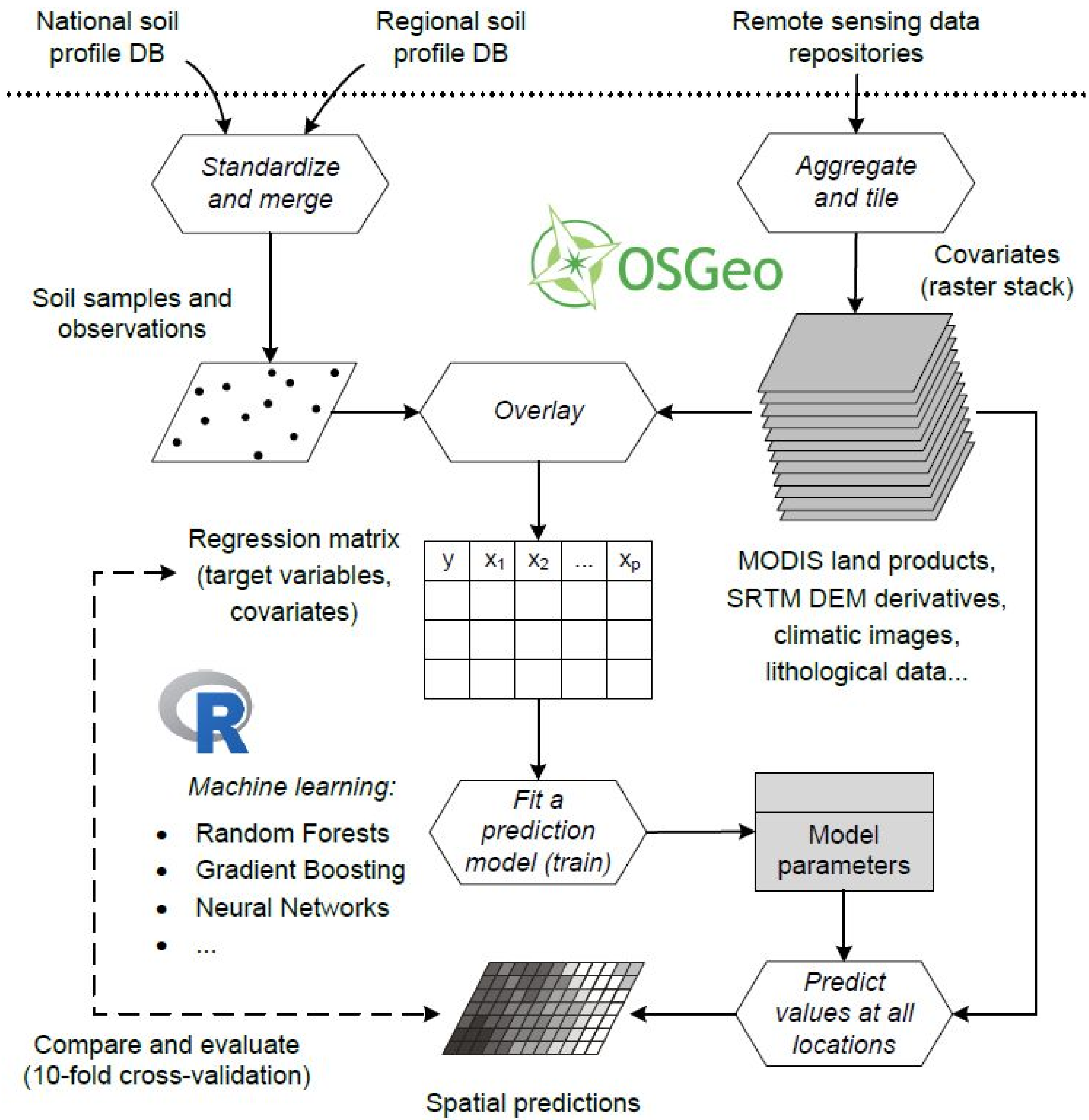
The United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries

ISRIC World Soil Information

Machine learning as a framework for automated soil mapping

Methods

- ➔ 2D and 3D soil properties: **ensemble random forest and gradient boosting** (ranger, xgboost)
- ➔ soil types: **ensemble random forest and nnet::multinom**
- ➔ Cross-validation, post-processing, pseudo-observations


National soil profile DB
Regional soil profile DB
Remote sensing data repositories
Standardize and merge
Aggregate and tile
OSGeo
Covariates (raster stack)
Soil samples and observations
Overlay
Regression matrix (target variables, covariates)
y
x1
x2
...
xp
MODIS land products, SRTM DEM derivatives, climatic images, lithological data...
R
Machine learning:
Random Forests
Gradient Boosting
Neural Networks
...
Fit a prediction model (train)
Model parameters
Predict values at all locations
Compare and evaluate (10-fold cross-validation)
Spatial predictions

SoilGrids are possible espoecially thanks to authors of: ranger, xgboost, caret, raster, nnet, SAGA GIS, GDAL, Geoserver...

ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R

Marvin N. Wright
Universität zu Lübeck

Andreas Ziegler
Universität zu Lübeck,
University of KwaZulu-Natal

Abstract

We introduce the C++ application and R package **ranger**. The software is a fast implementation of random forests for high dimensional data. Ensembles of classification, regression and survival trees are supported. We describe the implementation, provide examples, validate the package with a reference implementation, and compare runtime and memory usage with other implementations. The new software proves to scale best with the number of features, samples, trees, and features tried for splitting. Finally, we show that **ranger** is the fastest and most memory efficient implementation of random forests to analyze data on the scale of a genome-wide association study.

Keywords: C++, classification, machine learning, R, random forests, **Rcpp**, recursive partitioning, survival analysis.

XGBoost: A Scalable Tree Boosting System

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

ABSTRACT

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

CCS Concepts

•Methodologies → Machine learning; •Information systems → Data mining;

Keywords

many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks [14]. LambdaMART [4], a variant of tree boosting for ranking, achieves state-of-the-art result for ranking problems. Besides being used as a stand-alone predictor, it is also incorporated into real-world production pipelines for ad click through rate prediction [13]. Finally, it is the de-facto choice of ensemble method and is used in challenges such as the Netflix prize [2].

In this paper, we describe XGBoost, a scalable machine learning system for tree boosting. The system is available as an open source package[2]. The impact of the system has been widely recognized in a number of machine learning and data mining challenges. Take the challenges hosted by the machine learning competition site Kaggle for example. Among the 29 challenge winning solutions [3] published at Kaggle's blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with neural nets in en-

Journal of Statistical Software

November 2008, Volume 28, Issue 5. http://www.jstatsoft.org/

Building Predictive Models in R Using the caret Package

Max Kuhn
Pfizer Global R&D

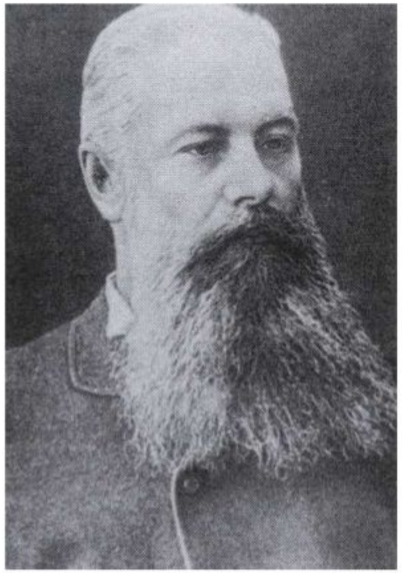
Abstract

The **caret** package, short for classification and regression training, contains numerous tools for developing predictive models using the rich set of models available in R. The package focuses on simplifying model training and tuning across a wide variety of modeling techniques. It also includes methods for pre-processing training data, calculating variable importance, and model visualizations. An example from computational chemistry is used to illustrate the functionality on a real data set and to benchmark the benefits of parallel processing with several types of models.

Keywords: model building, tuning parameters, parallel processing, R, **NetWorkSpaces**.

Results

They would have been interested in this…



Vasili Dokuchaev

The Russian School

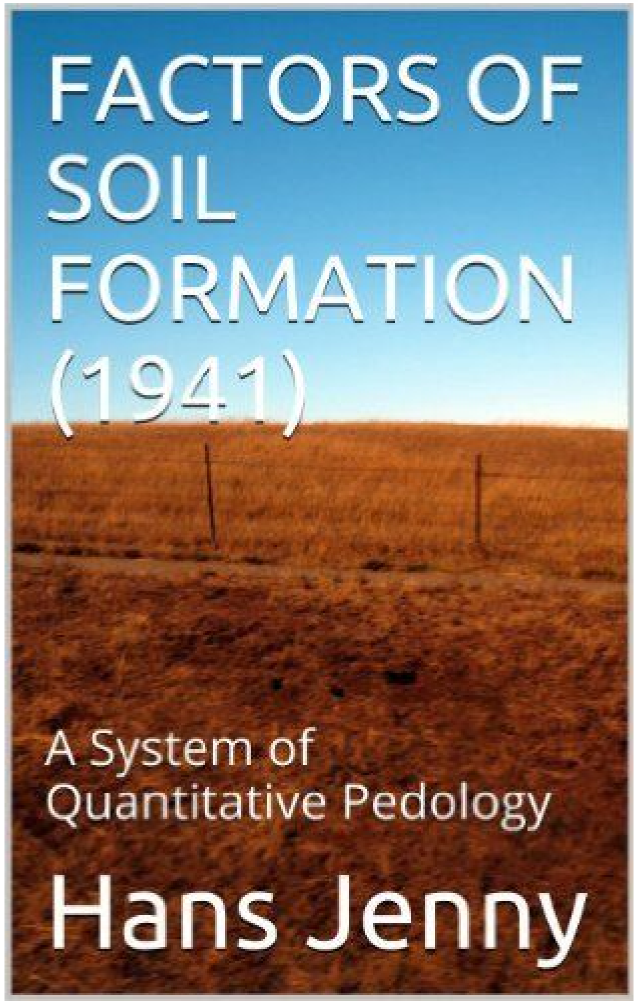
Soil forming factors

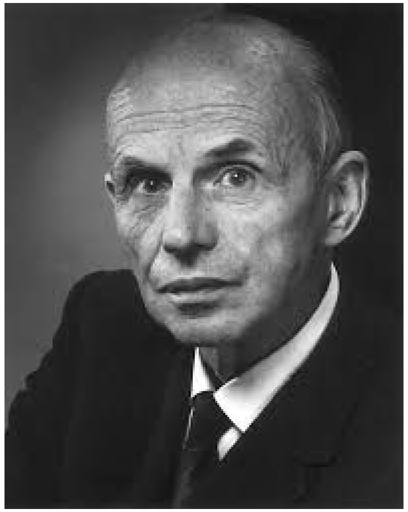
⇩

Soil forming processes

⇩

Different Soils





Bulk density

Soil pH

Soil organic carbon

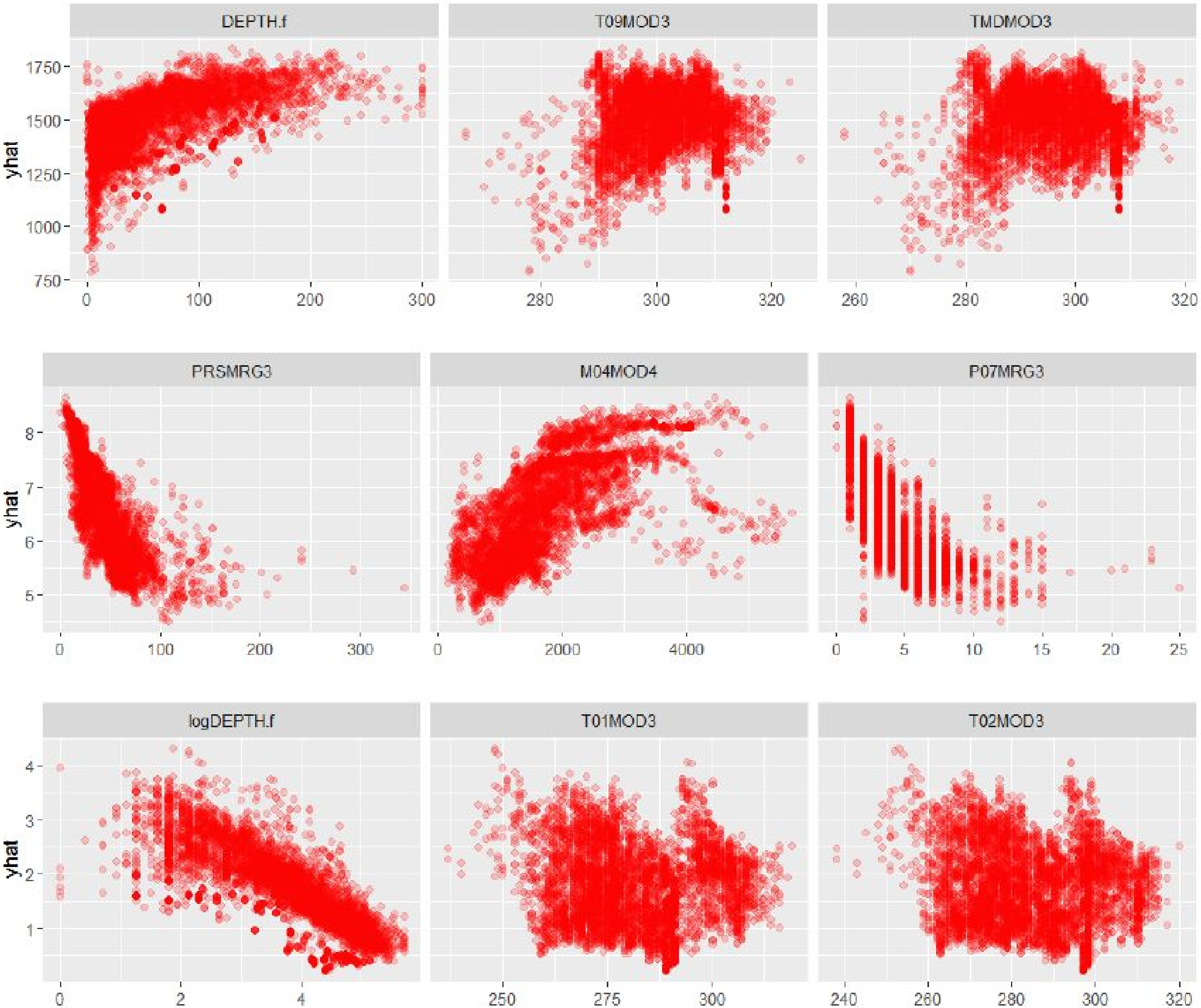


Figure 6. Examples of fitted relationships for bulk density (above), pH (middle) and soil organic carbon (below). Plots show target variables and top three most important covariates as reported by the random forest model. `DEPTH.f` is the depth from soil surface, `T09MOD3` is mean monthly temperature for September, `TMDMOD3` is mean annual temperature, `PRSMRG3` is total annual precipitation, `M04MOD4` is mean monthly MODIS NIR band reflectance, `P07MRG3` is mean monthly precipitation for July, `T01MOD3` is mean monthly temperature for January, and

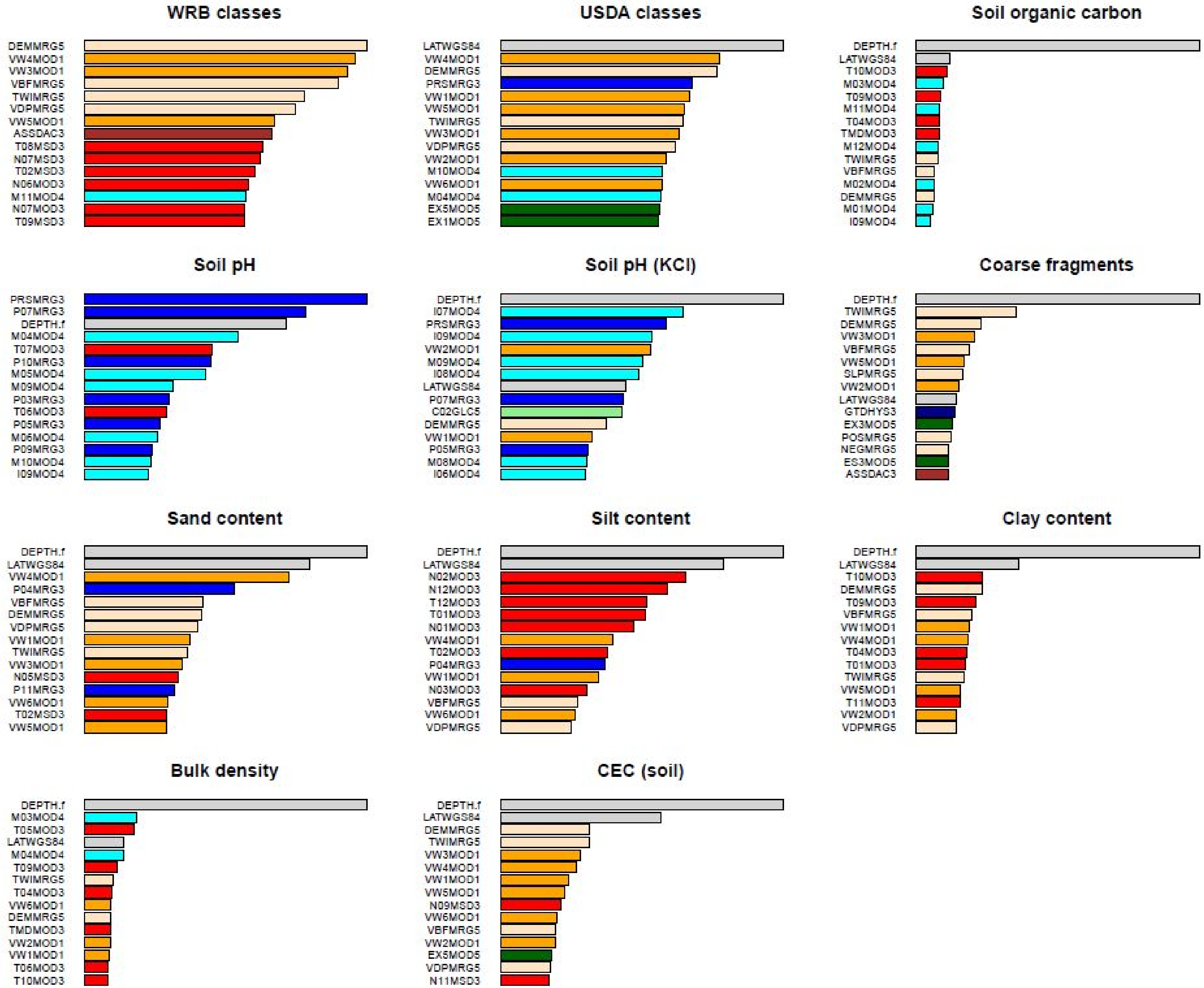


Figure 5. Fitted variable importance plots for target variables. Generated as an average between using the `ranger` and `xgboost` packages, (for soil types results are based on the `ranger` model only). `DEPTH.f` is the depth from soil surface, `T**MOD3` and `N**MOD3` are mean monthly temperatures daytime and nighttime (red color), `TWI`, `DEM`, `VBF` and `VDP` are DEM-parameters (bisque color), `M**MOD4` are mean

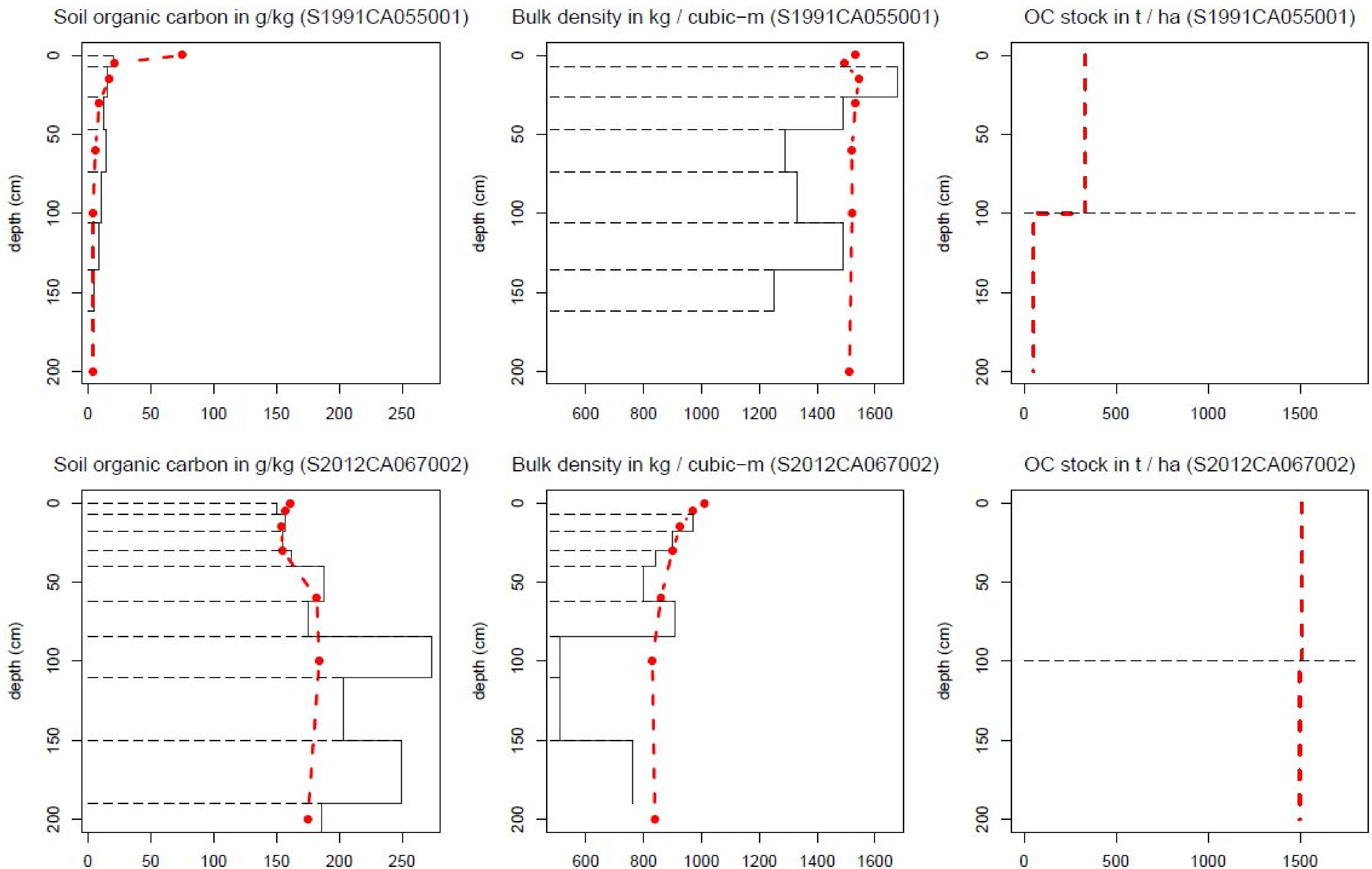
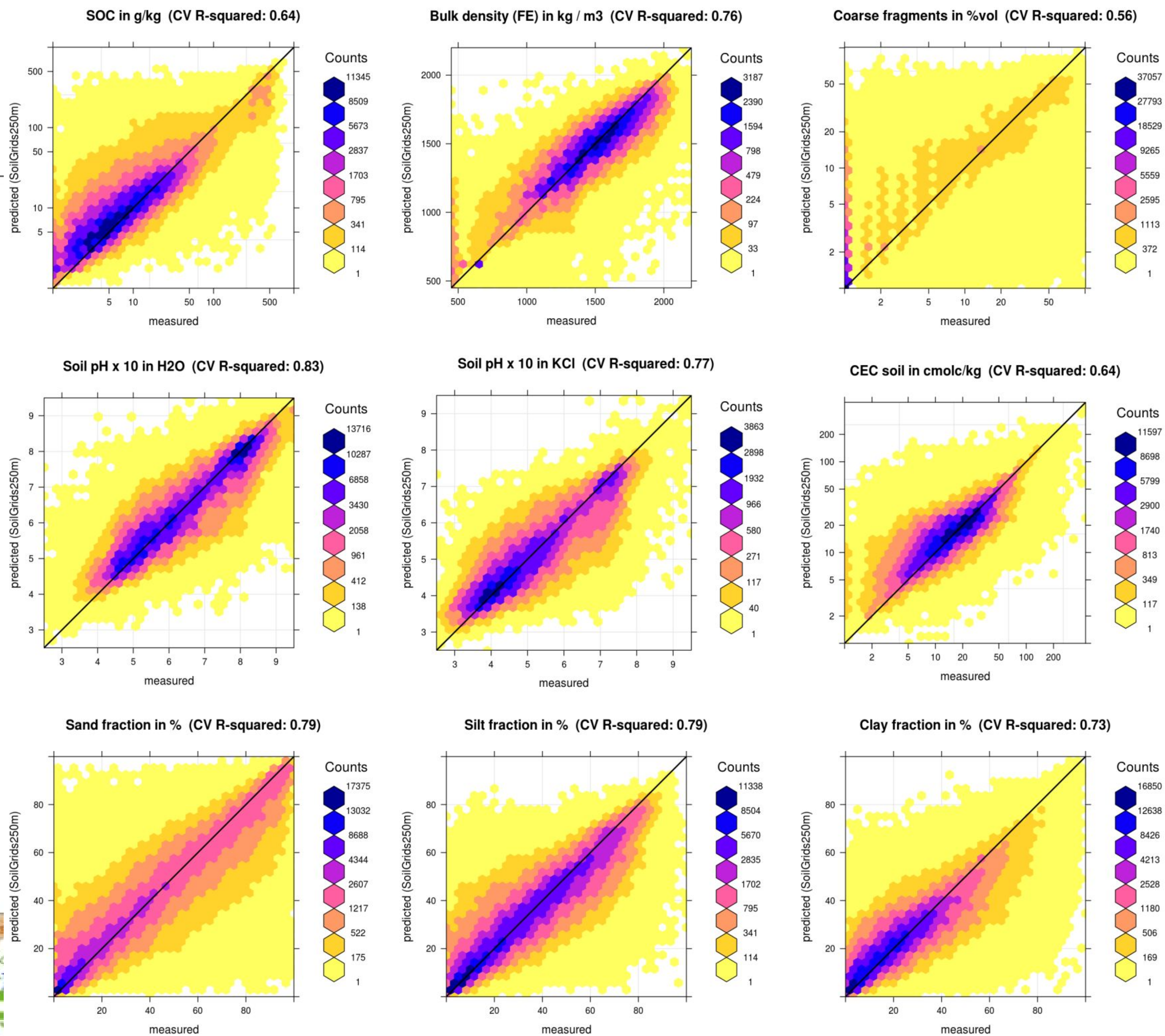
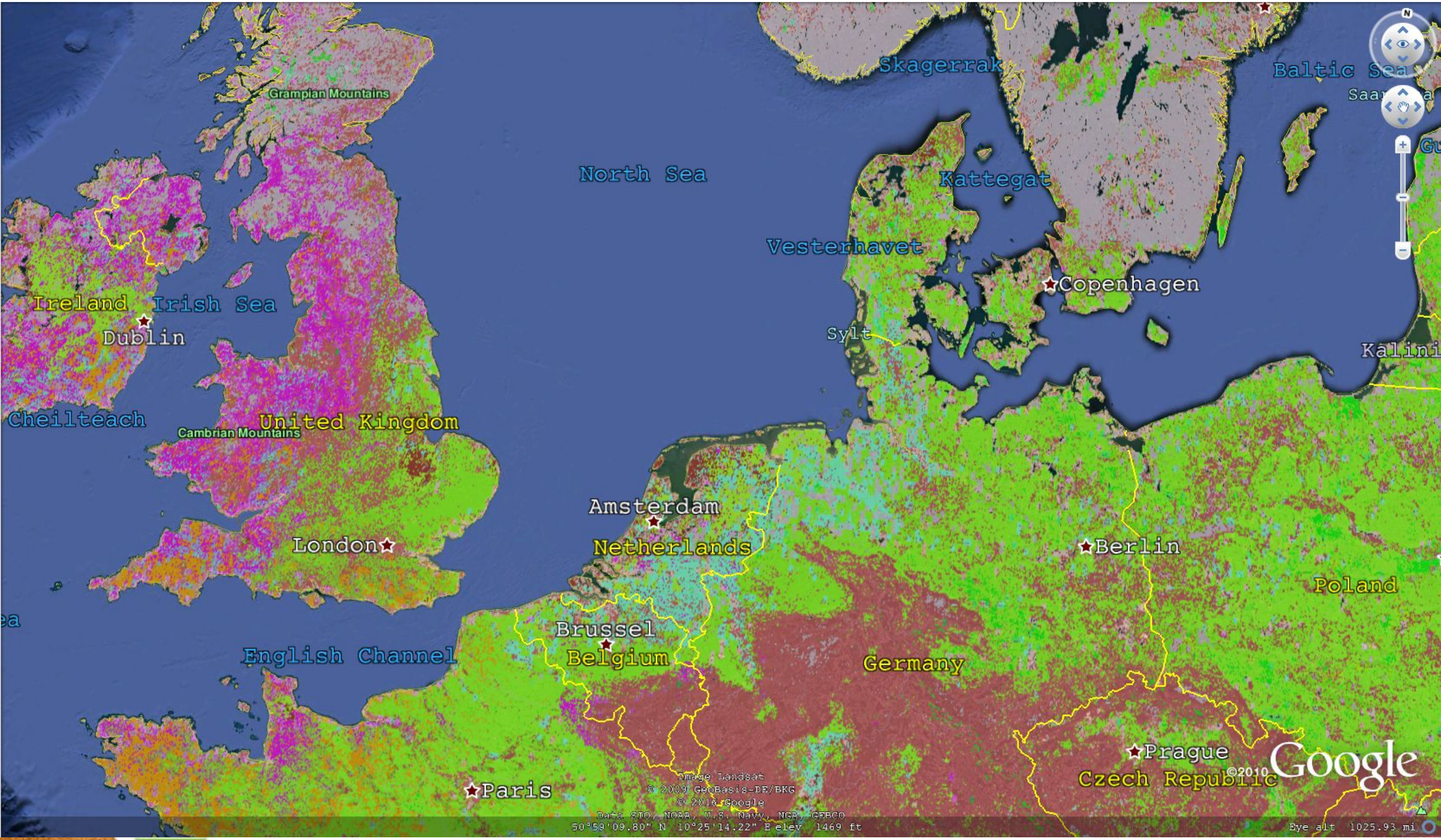


Figure 2. Example of soil variable-depth curves: original sampled soil profiles vs predicted values (SoilGrids) at seven standard depths (broken red line) and estimated soil organic carbon stock for depths 0–100 and 100–200 cm. Locations of points: mineral soil S1991CA055001 (-122.37°W, 38.25°N), and an organic soil profile S2012CA067002 (-121.62°W, 38.13°N).

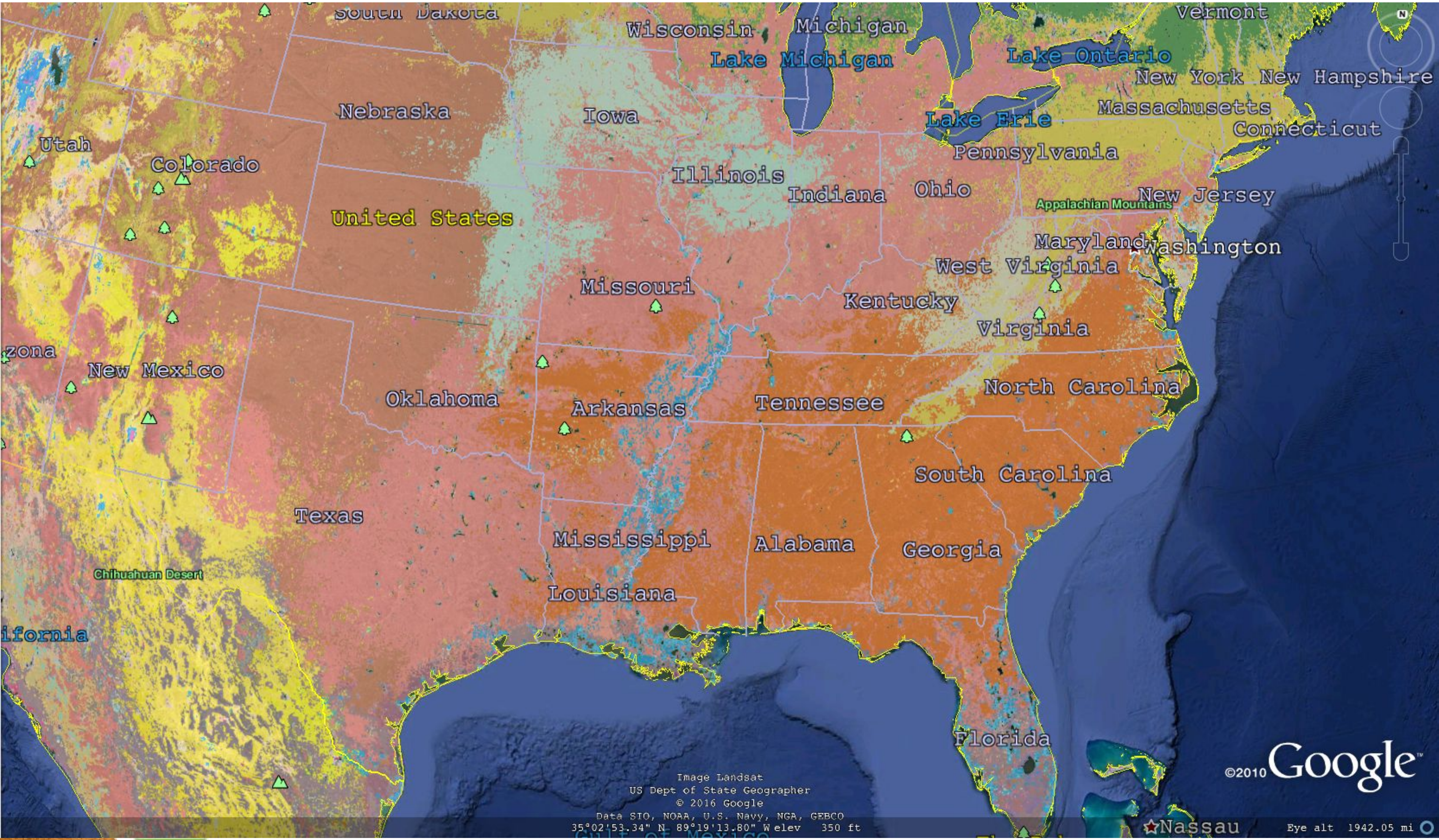

SOC in g/kg (CV R-squared: 0.64)
Counts
11345
8509
5673
2837
1703
795
341
114
1
Bulk density (FE) in kg / m3 (CV R-squared: 0.76)
Counts
3187
2390
1594
798
479
224
97
33
1
Coarse fragments in %vol (CV R-squared: 0.56)
Counts
37057
27793
18529
9265
5559
2595
1113
372
1
Soil pH x 10 in H2O (CV R-squared: 0.83)
Counts
13716
10287
6858
3430
2058
961
412
138
1
Soil pH x 10 in KCl (CV R-squared: 0.77)
Counts
3863
2898
1932
966
580
271
117
40
1
CEC soil in cmolc/kg (CV R-squared: 0.64)
Counts
11597
8698
5799
2900
1740
813
349
117
1
Sand fraction in % (CV R-squared: 0.79)
Counts
17375
13032
8688
4344
2607
1217
522
175
1
Silt fraction in % (CV R-squared: 0.79)
Counts
11338
8504
5670
2835
1702
795
341
114
1
Clay fraction in % (CV R-squared: 0.73)
Counts
16850
12638
8426
4213
2528
1180
506
169
1
predicted (SoilGrids250m)
measured

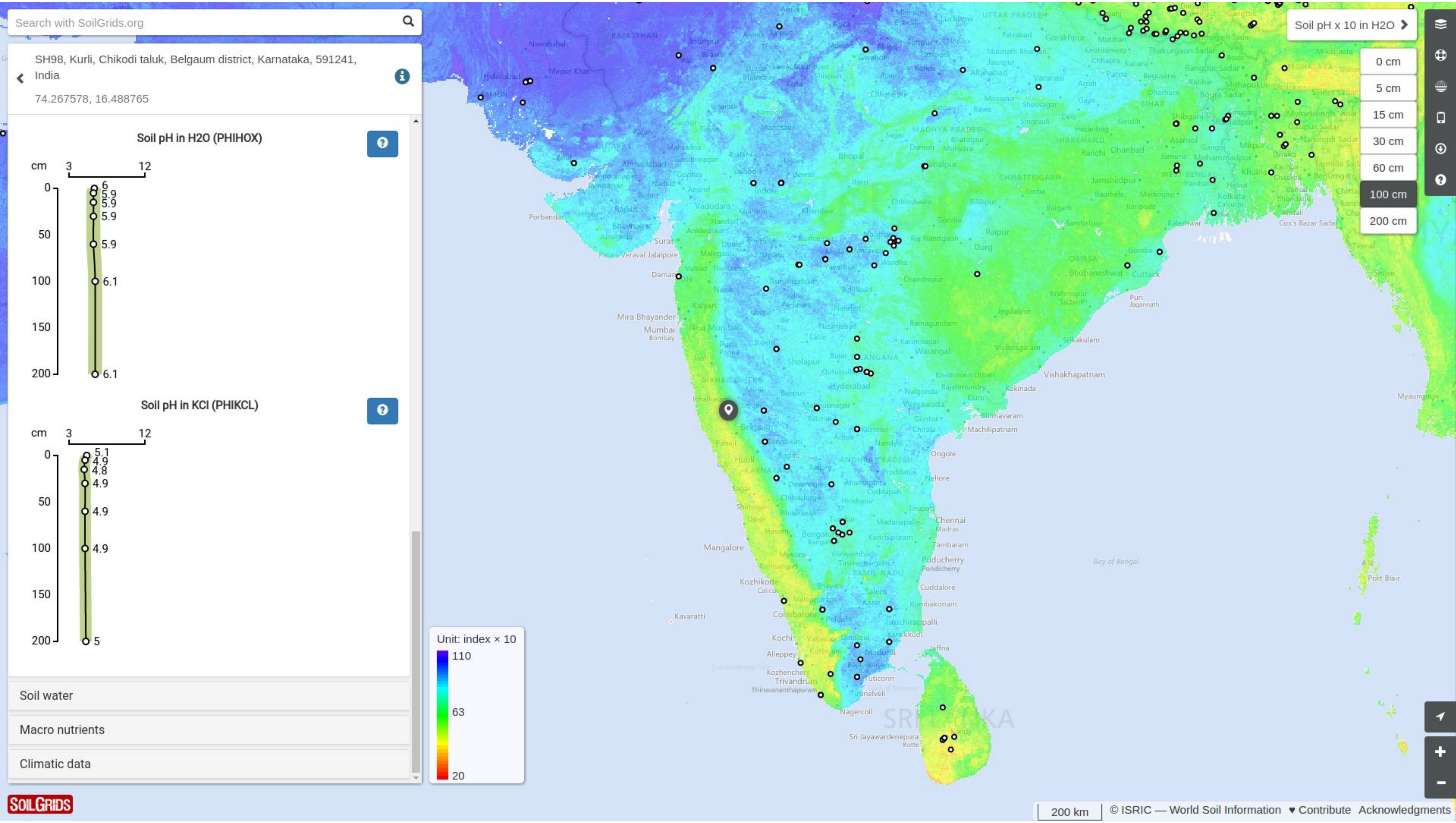
Maps

USDA suborders -> in Europe!



WRB 2nd level -> in USA!




Search with SoilGrids.org
SH98, Kurli, Chikodi taluk, Belgaum district, Karnataka, 591241, India
74.267578, 16.488765
Soil pH in H2O (PHIHOX)
cm
3
12
0
50
100
150
200
6
5.9
5.9
5.9
5.9
6.1
6.1
Soil pH in KCl (PHIKCL)
cm
3
12
0
50
100
150
200
5.1
4.9
4.8
4.9
4.9
4.9
5
Soil water
Macro nutrients
Climatic data
Unit: index × 10
110
63
20
Soil pH x 10 in H2O
0 cm
5 cm
15 cm
30 cm
60 cm
100 cm
200 cm
SOILGRIDS
200 km
© ISRIC — World Soil Information
♥ Contribute
Acknowledgments

Search with SoilGrids.org

Sei Baru Tewu, Central Kalimantan, Indonesia

114.038086, -2.537012

Soil classification

Predicted USDA Soil Taxonomy class (Twelfth Edition; 2014)

Fibrists (20%)

(TAXOUSDA)

Histosols that are primarily made up of only slightly decomposed organic materials, often called peat.

Udults (13%) | Fluvents (10%)

Predicted World Reference Base (2006) soil class

Fibric Histosols (13%)

(TAXNWRB)

Histosols = Soils consisting primarily of organic materials. They are defined as having 40 centimetres or more of organic soil material in the upper 80 centimetres. Having, after rubbing, two-thirds or more (by volume) of the organic material consisting of recognizable plant tissue within 100 cm of the soil surface (in Histosols only).

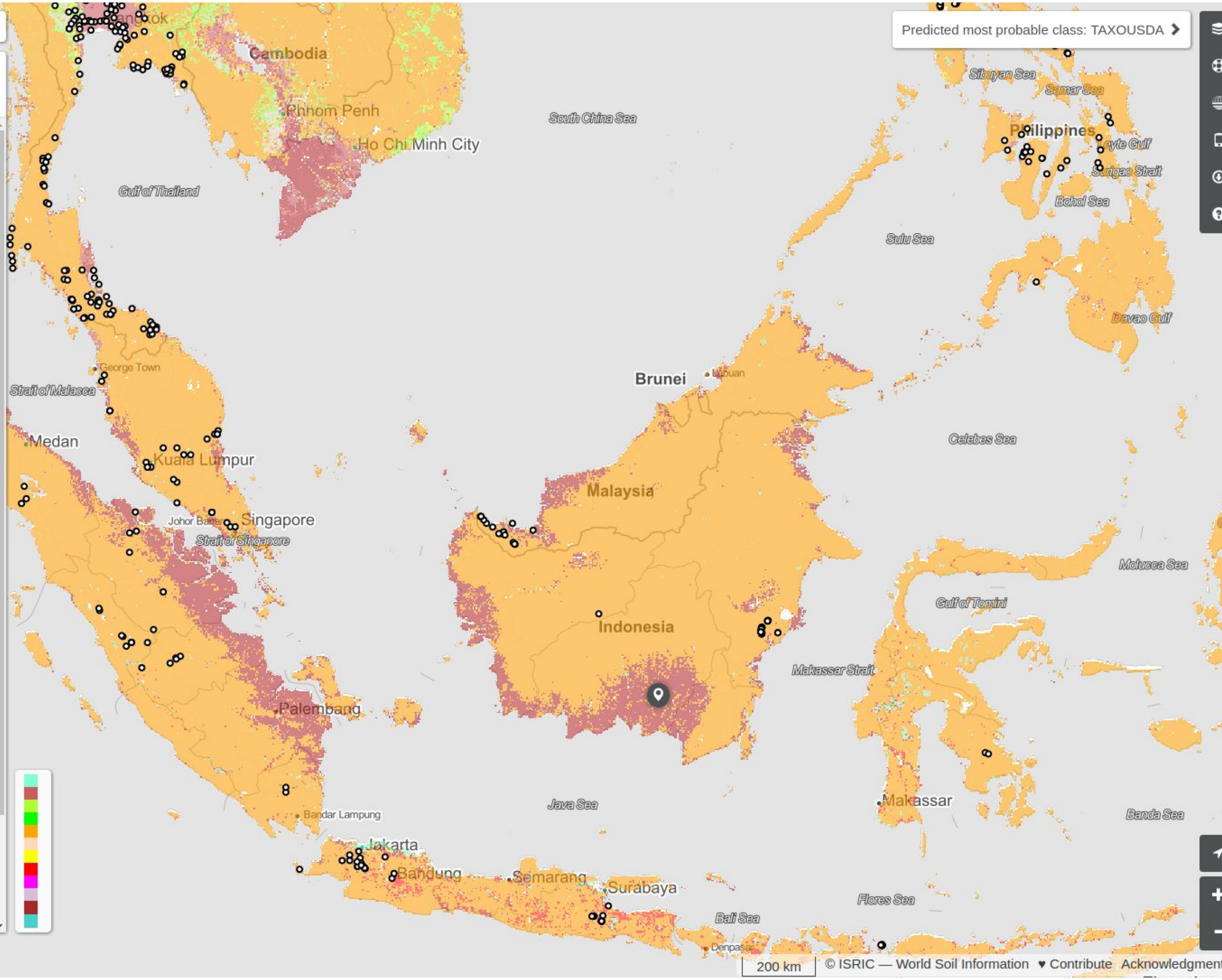
Haplic Gleysols (9%) | Acric Plinthosols (8%)

Site characteristics

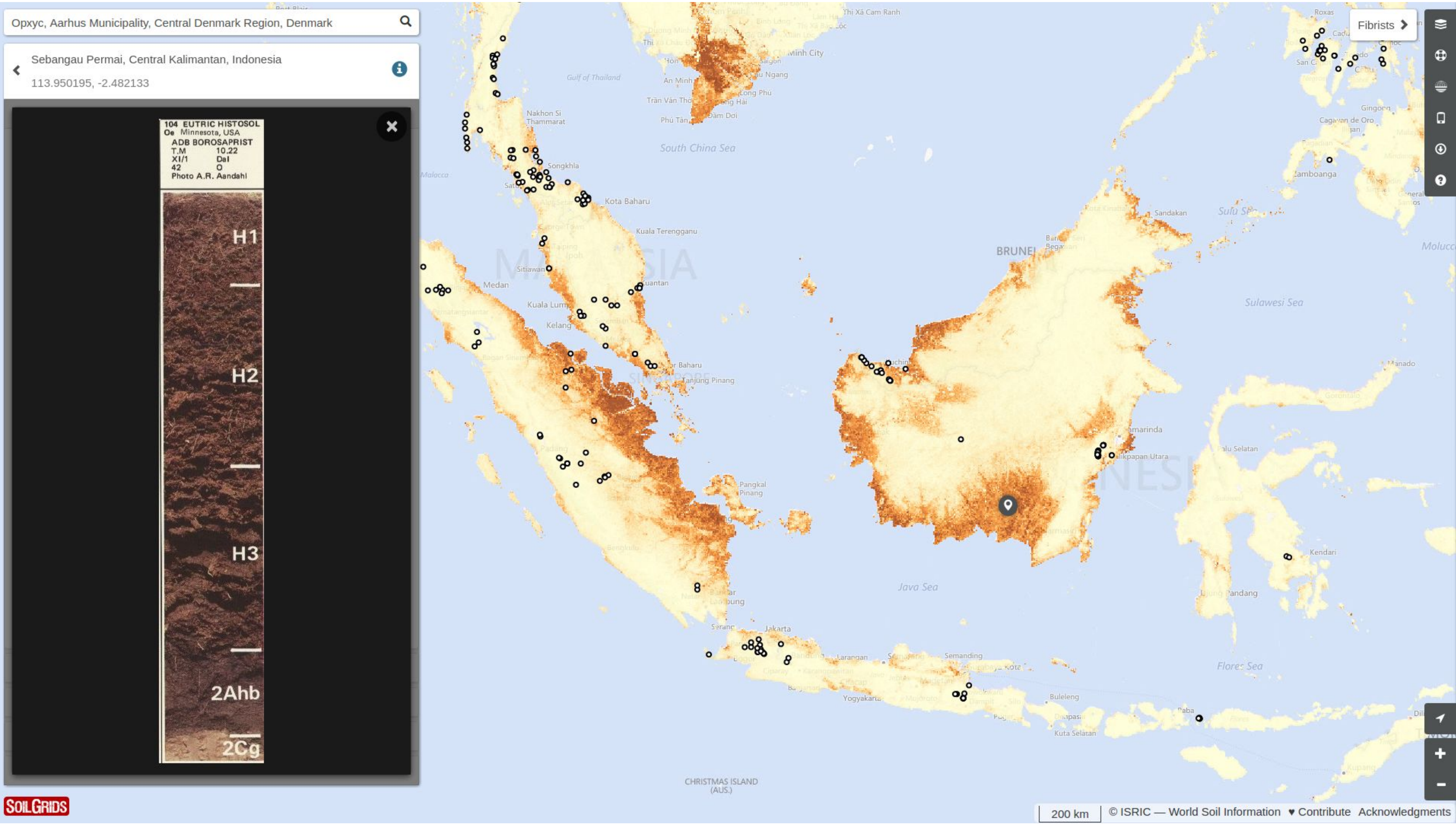
Physical soil properties

Chemical soil properties

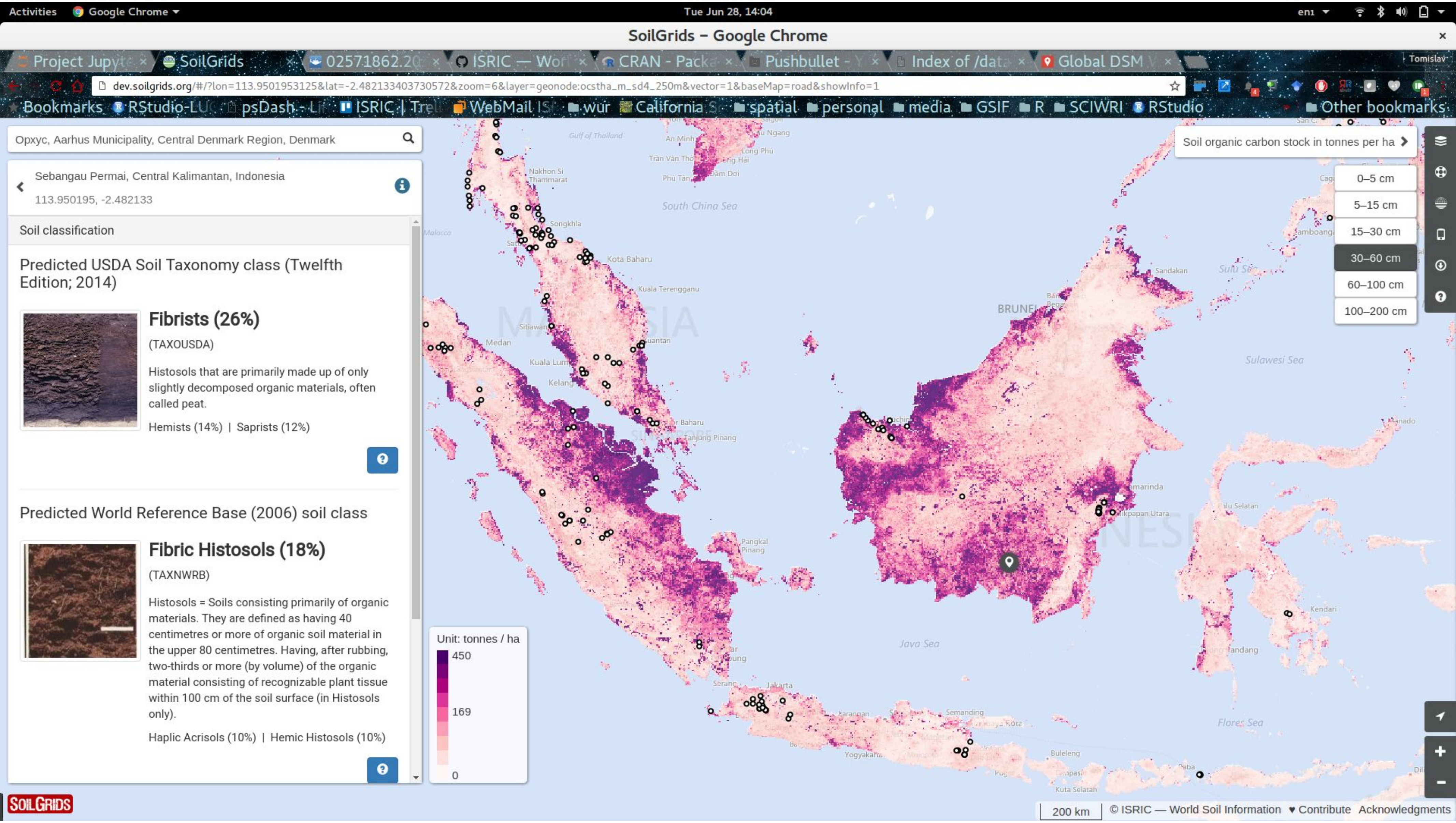
SoilGrids



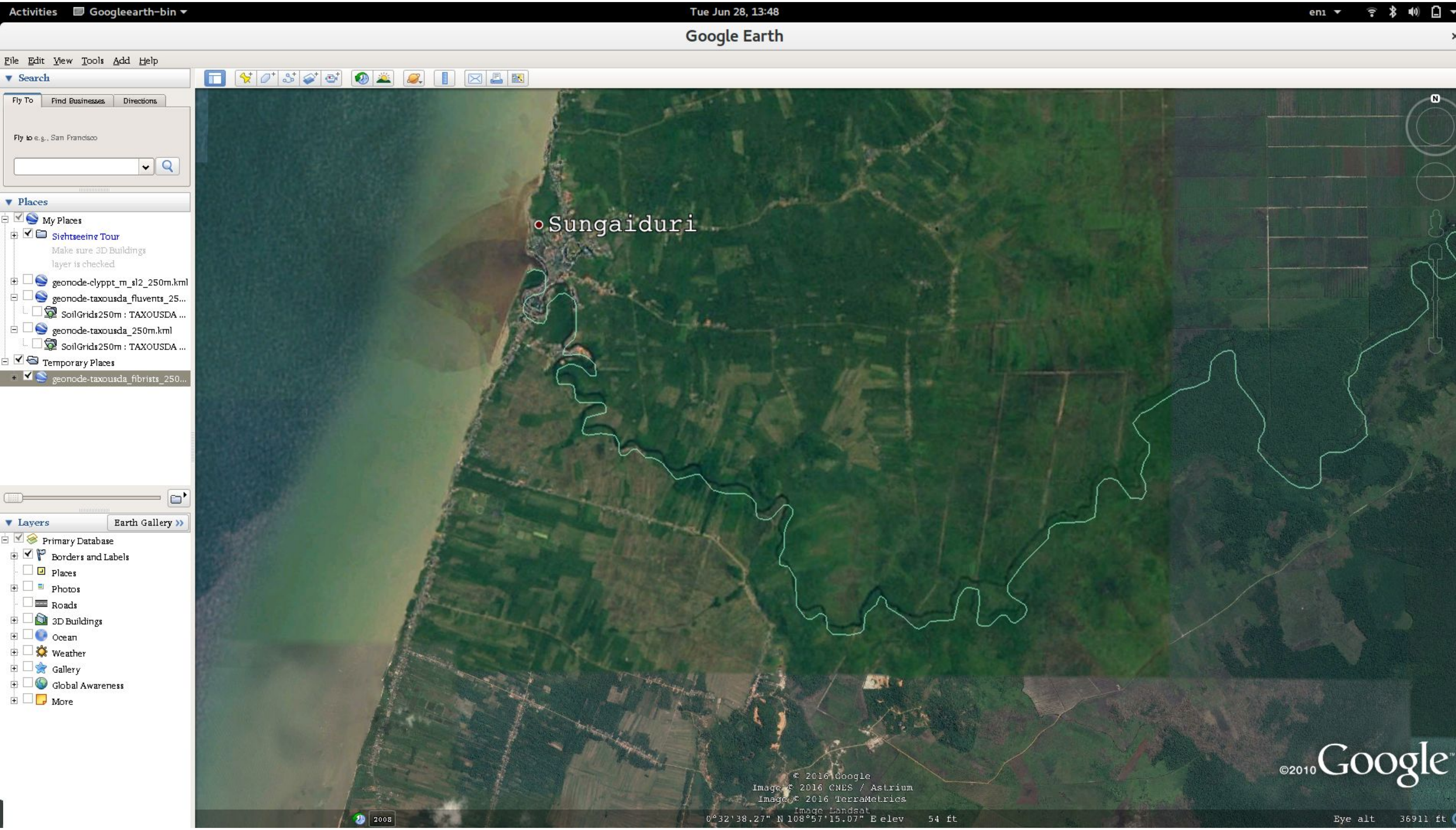
ISRIC World Soil Information

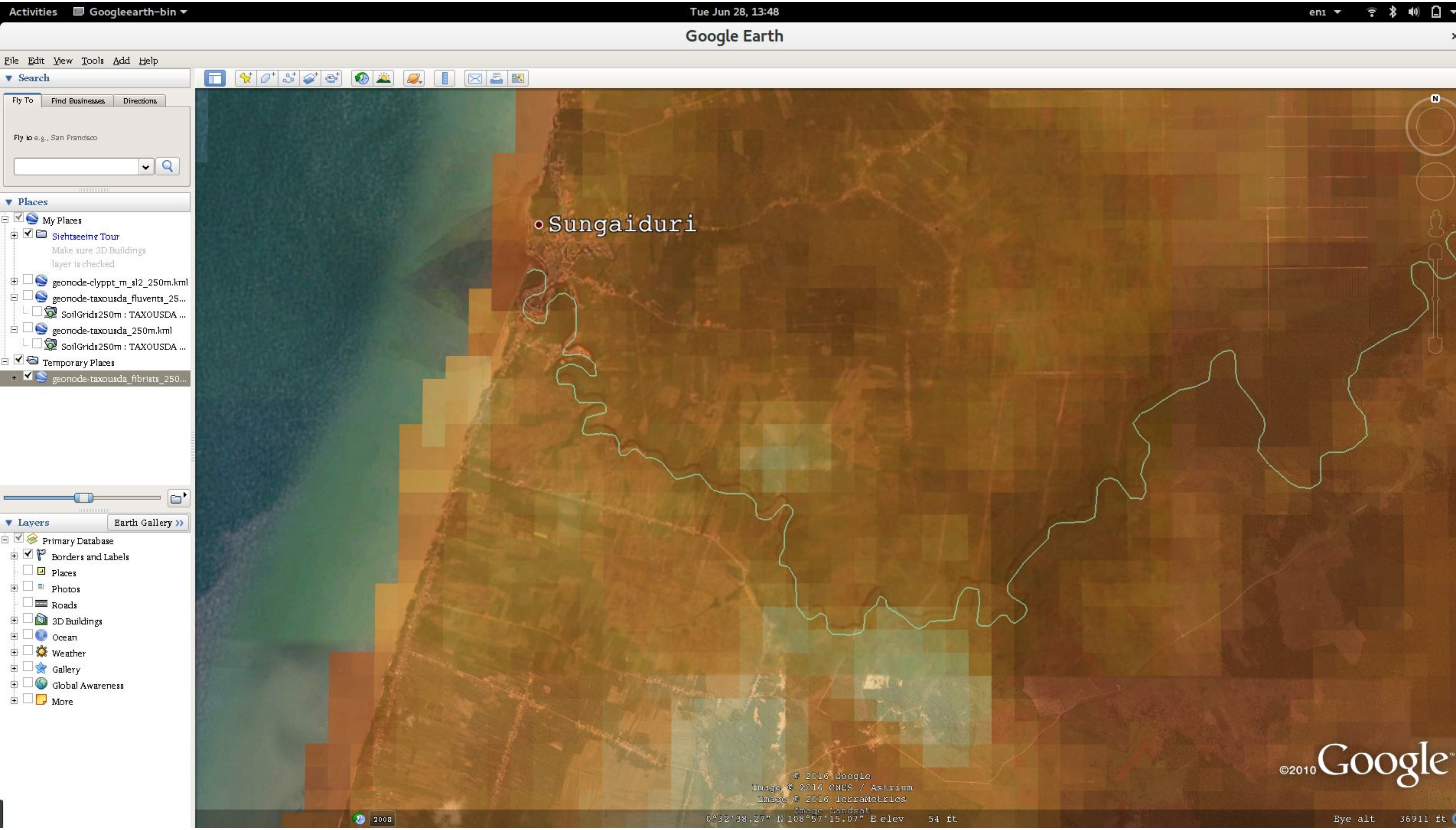


ISRIC World Soil Information

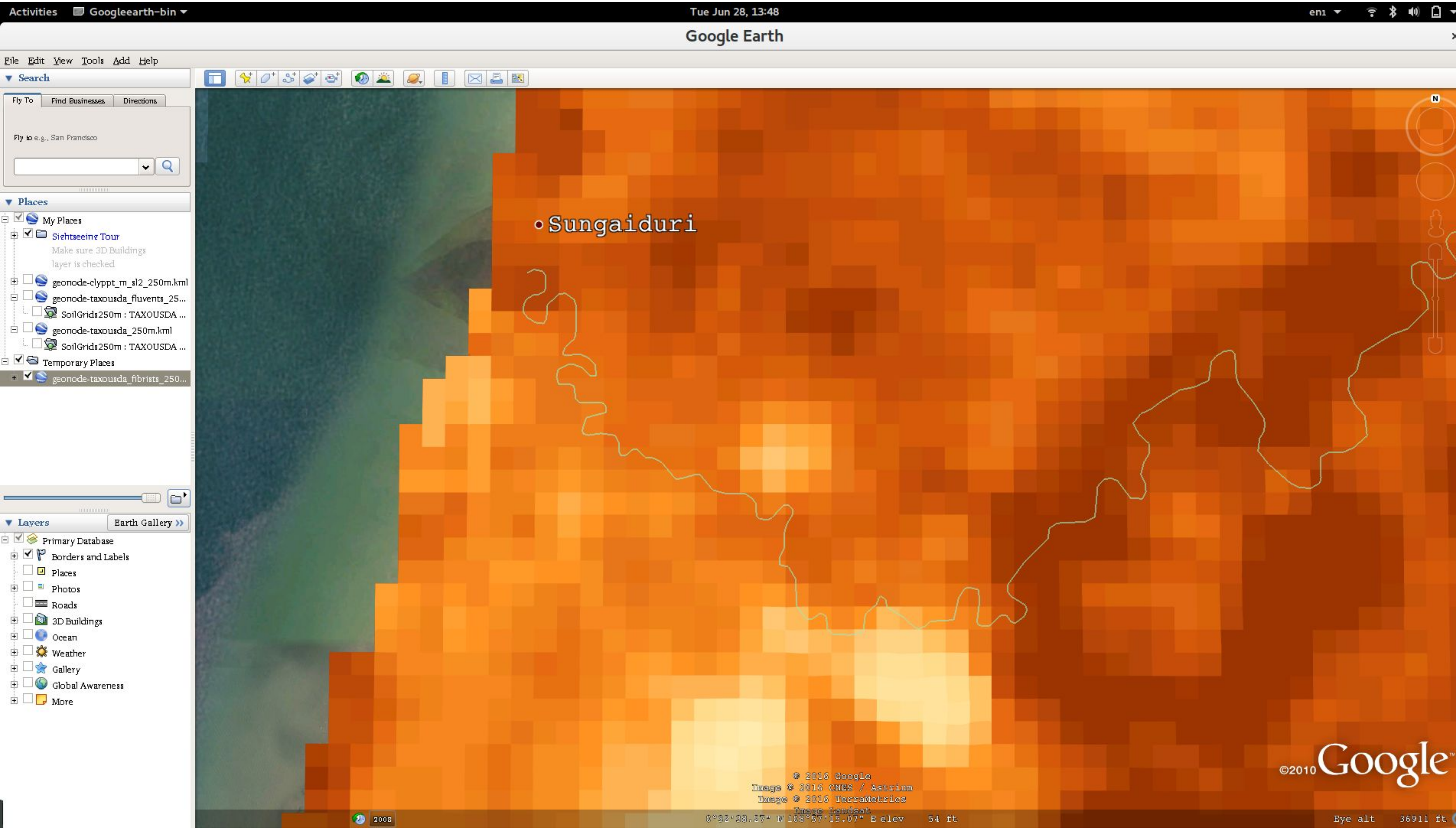


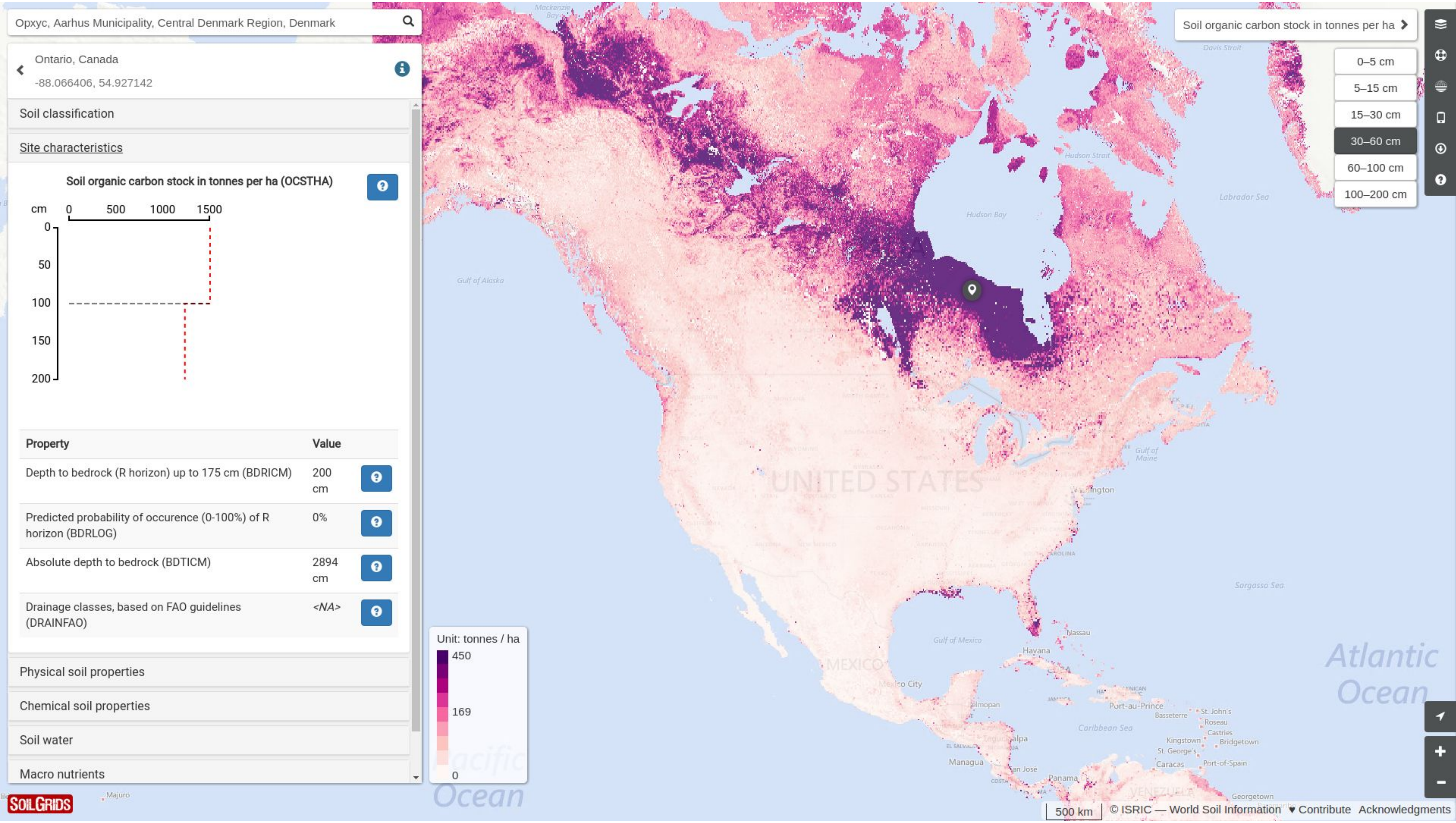
World Soil Information


Activities
Googleearth-bin
Tue Jun 28, 13:48
Google Earth
File Edit View Tools Add Help
Search
Fly To
Find Businesses
Directions
Fly to e.g., San Francisco
Places
My Places
Sightseeing Tour
Make sure 3D Buildings layer is checked
geonode-clyppt_m_sl2_250m.kml
geonode-taxousda_fluvents_25...
SoilGrids250m : TAXOUSDA ...
geonode-taxousda_250m.kml
SoilGrids250m : TAXOUSDA ...
Temporary Places
geonode-taxousda_fibrists_250...
Layers
Earth Gallery
Primary Database
Borders and Labels
Places
Photos
Roads
3D Buildings
Ocean
Weather
Gallery
Global Awareness
More
Sungaiduri
© 2016 Google
Image © 2016 CNES / Astrium
Image © 2016 TerraMetrics
Image Landsat
2008
0°32'38.27" N 108°57'15.07" E elev 54 ft
Eye alt 36911 ft
©2010 Google

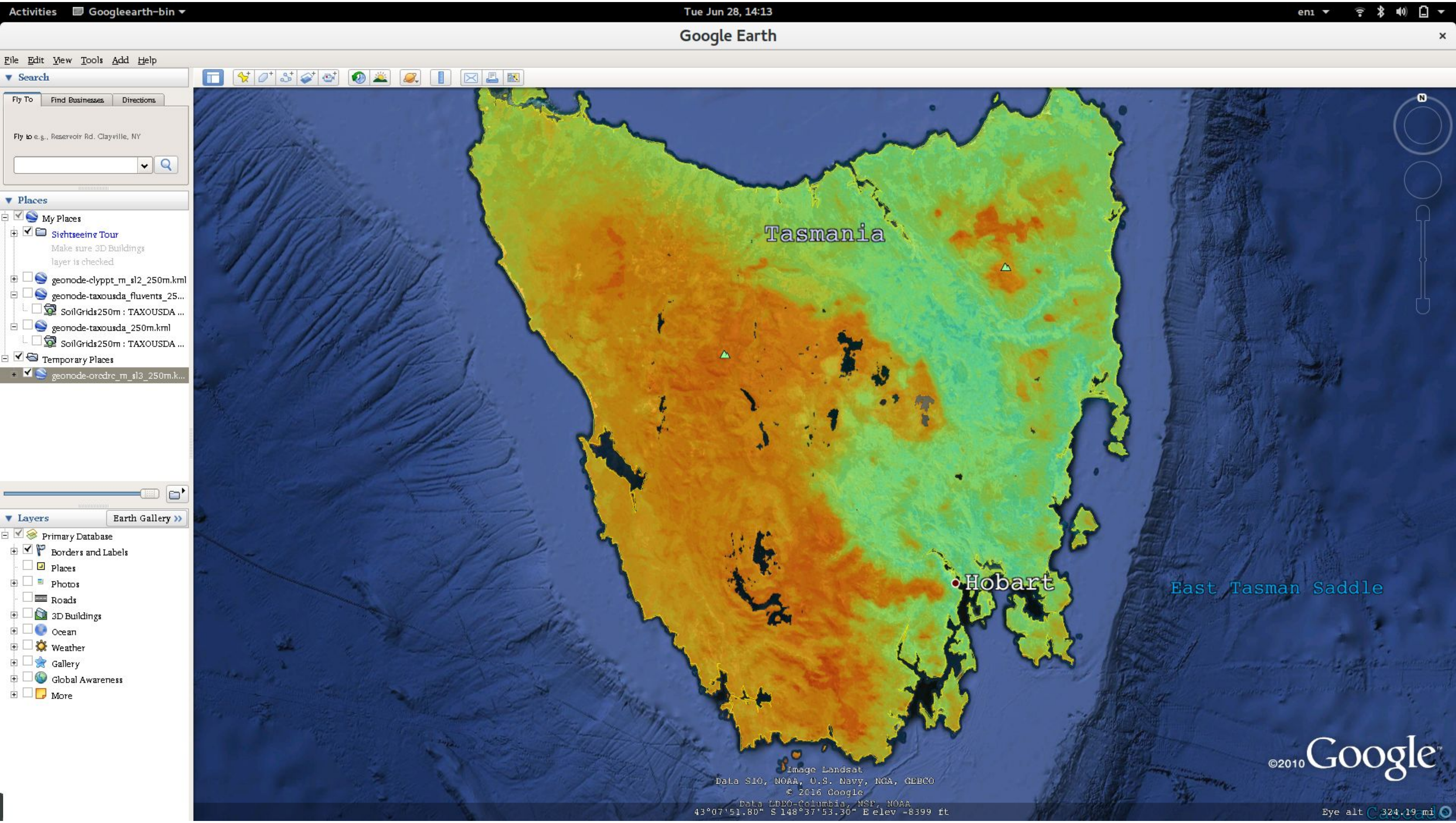


World Soil Information

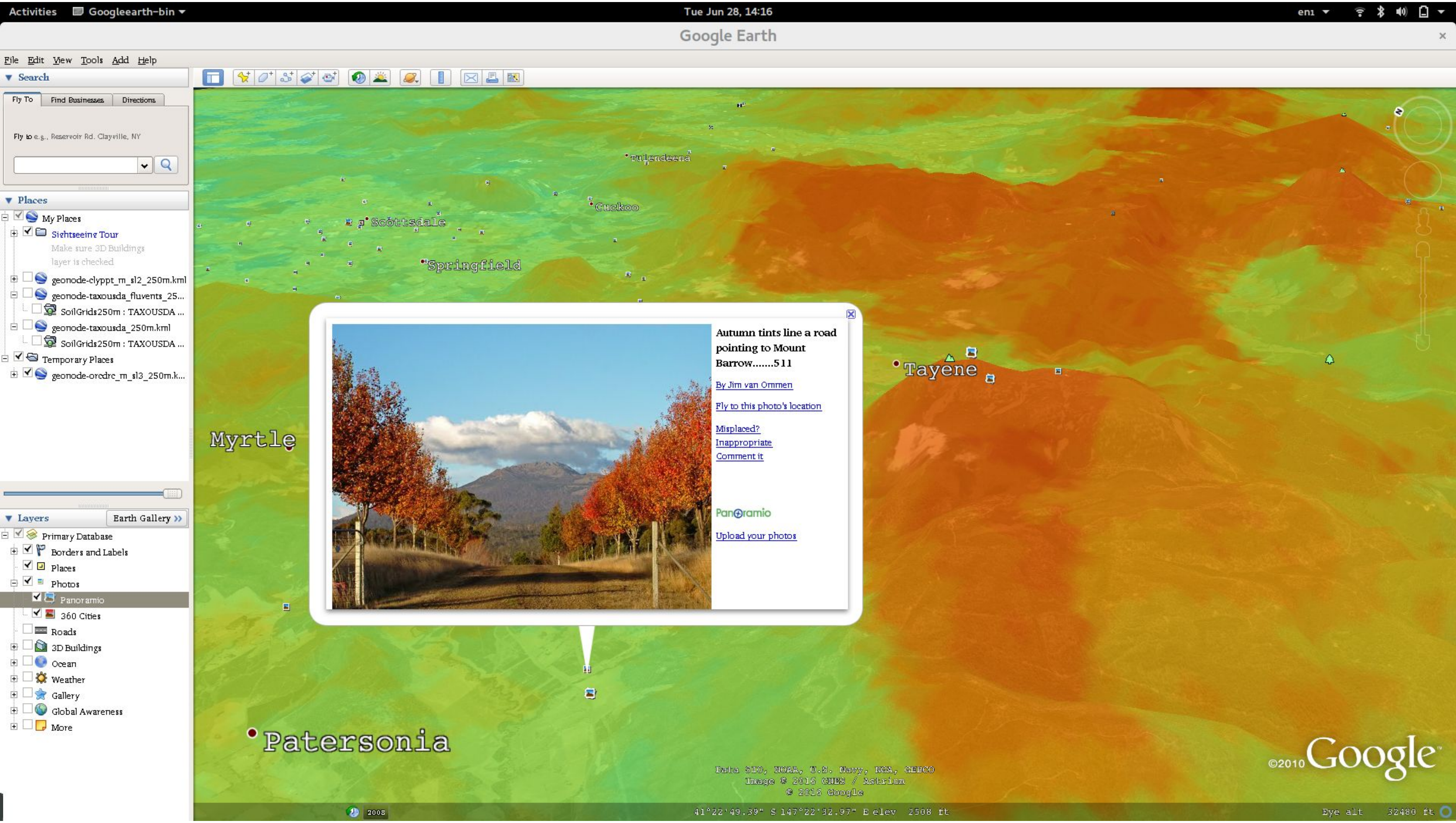

Activities
Googleearth-bin
Tue Jun 28, 13:48
Google Earth
File Edit View Tools Add Help
Search
Fly To
Find Businesses
Directions
Fly to e.g., San Francisco
Places
My Places
Sightseeing Tour
Make sure 3D Buildings layer is checked
geonode-clyppt_m_sl2_250m.kml
geonode-taxousda_fluvents_25...
SoilGrids250m : TAXOUSDA ...
geonode-taxousda_250m.kml
SoilGrids250m : TAXOUSDA ...
Temporary Places
geonode-taxousda_fibrists_250...
Layers
Earth Gallery
Primary Database
Borders and Labels
Places
Photos
Roads
3D Buildings
Ocean
Weather
Gallery
Global Awareness
More
Sungaiduri
©2010 Google
Eye alt 36911 ft



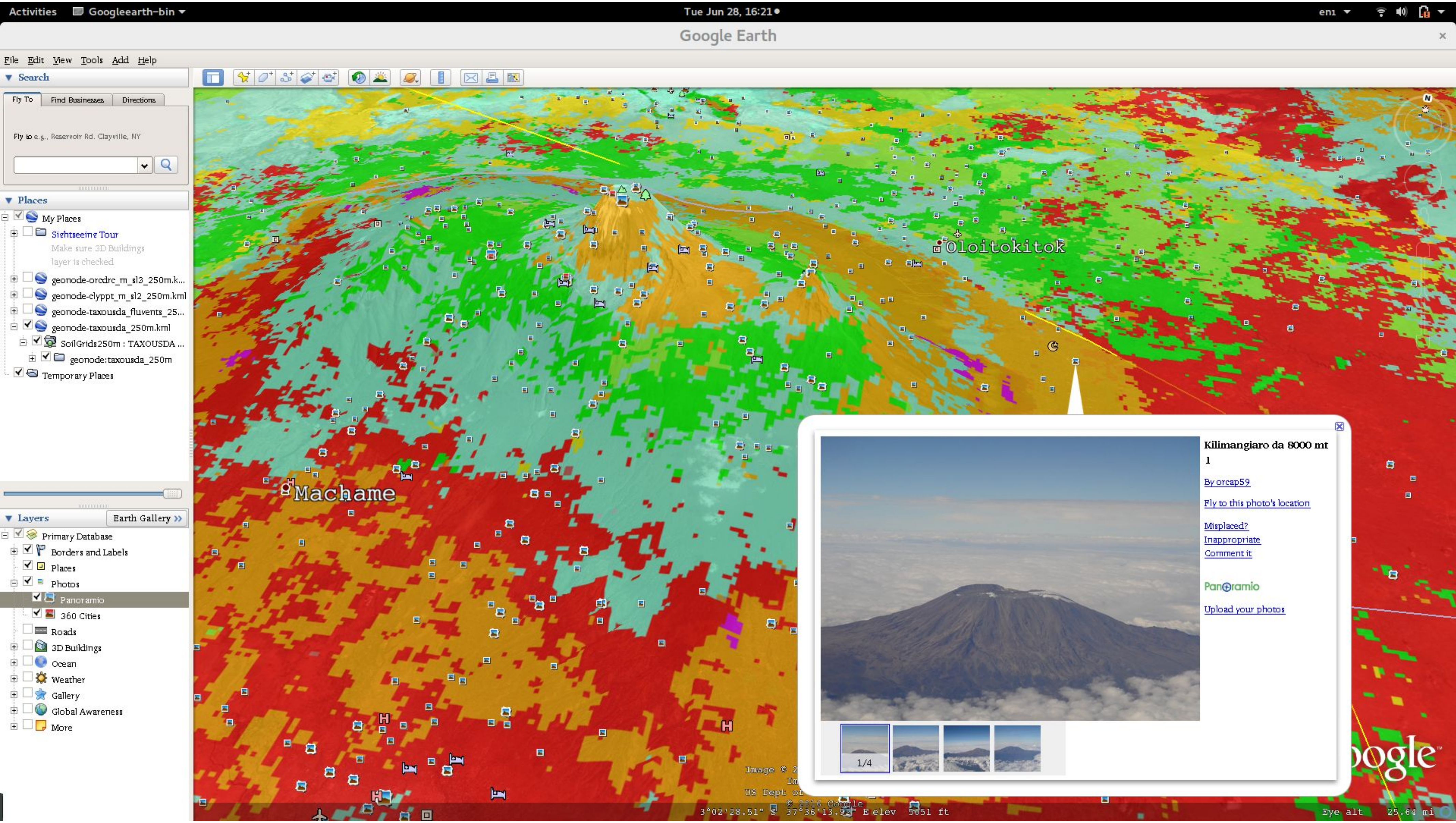
ISRIC World Soil Information



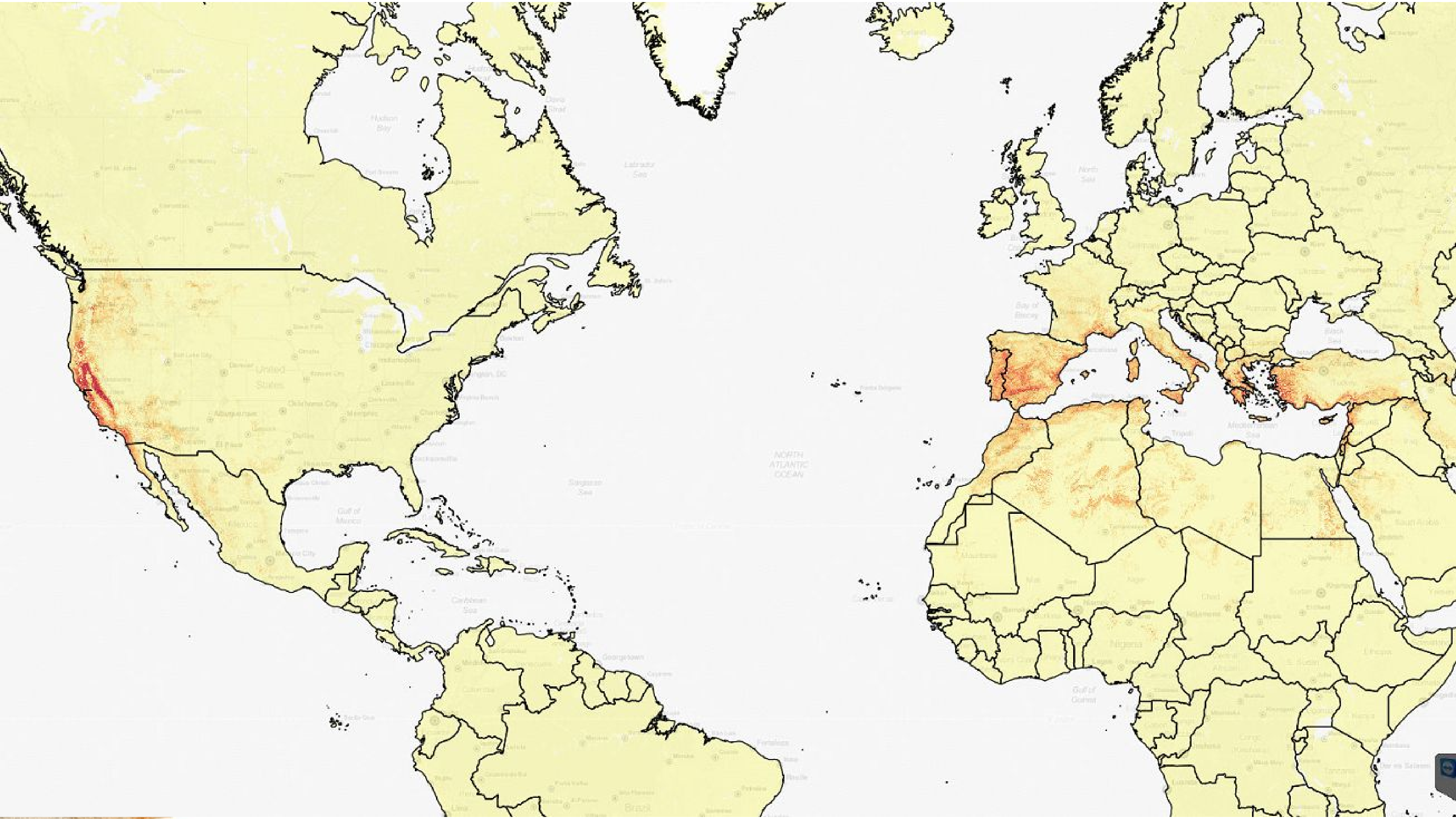
World Soil Information



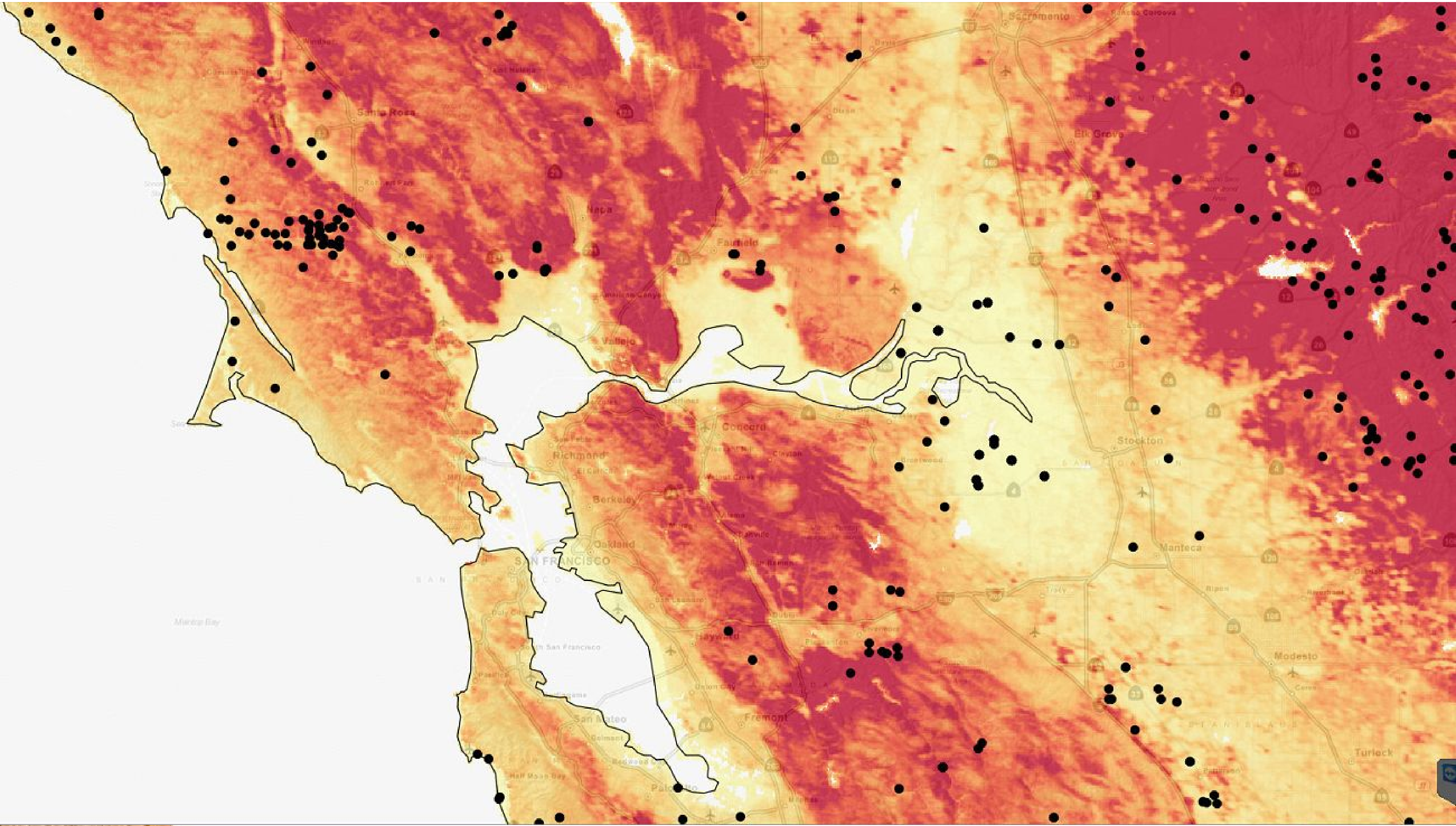
World Soil Information


Activities
Googleearth-bin
Tue Jun 28, 16:21
en1
Google Earth
File Edit View Tools Add Help
Search
Fly To
Find Businesses
Directions
Fly to e.g., Reservoir Rd, Clayville, NY
Places
My Places
Sightseeing Tour
Make sure 3D Buildings layer is checked
geonode-orcdrc_m_sl3_250m.k...
geonode-clyppt_m_sl2_250m.kml
geonode-taxousda_fluvents_25...
geonode-taxousda_250m.kml
SoilGrids250m : TAXOUSDA ...
geonode:taxousda_250m
Temporary Places
Layers
Earth Gallery
Primary Database
Borders and Labels
Places
Photos
Panoramio
360 Cities
Roads
3D Buildings
Ocean
Weather
Gallery
Global Awareness
More
Oloitokitok
Machame
Kilimangiaro da 8000 mt 1
By orcap59
Fly to this photo's location
Misplaced?
Inappropriate
Comment it
Panoramio
Upload your photos
1/4
3°02'28.51" S 37°36'13.92" E elev 5051 ft
Eye alt 25.64 mi

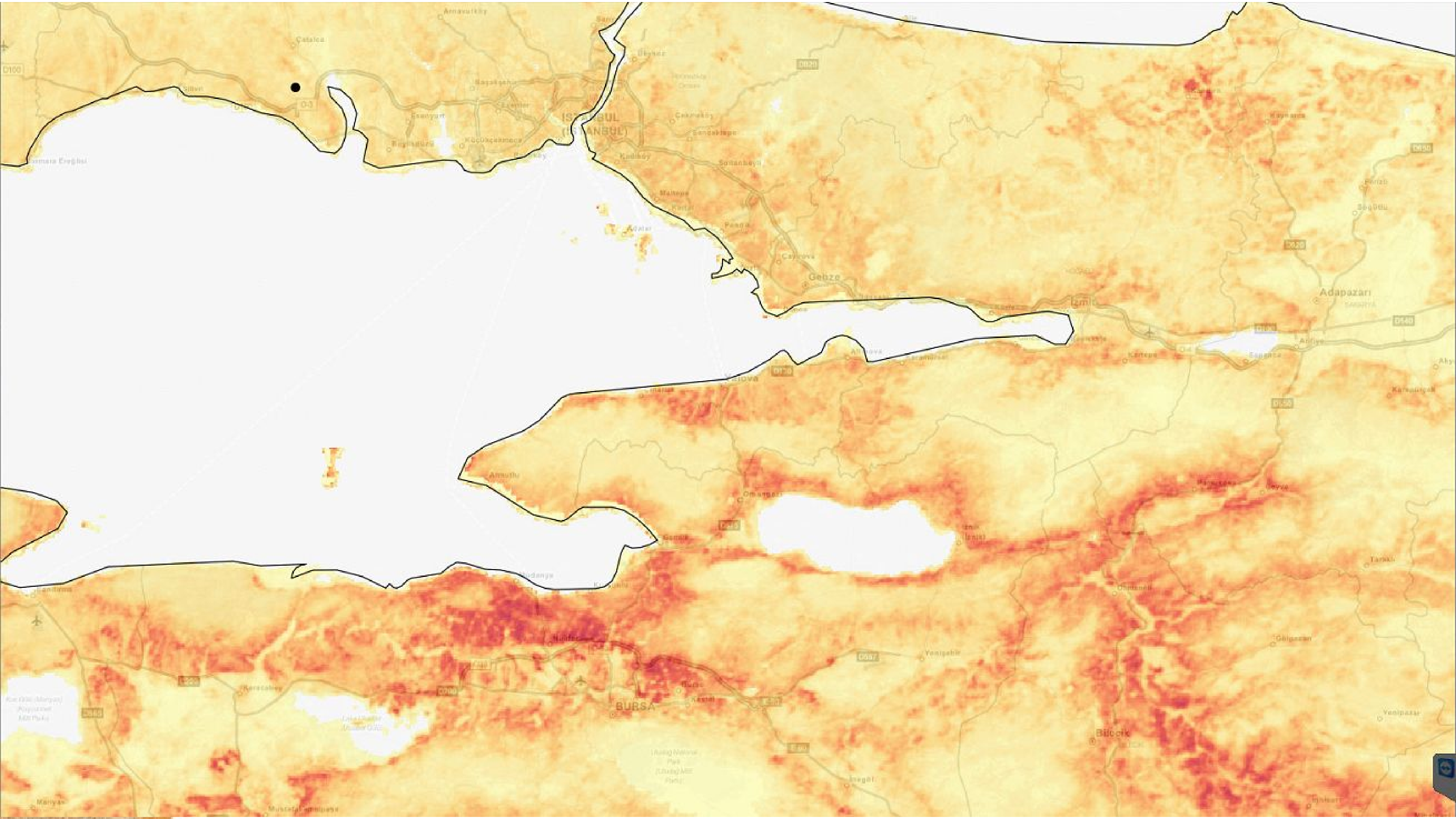
Xeralfs



Thanks to enough soil profiles from USA...



... it is possible to map areas of similar soil/climate in Turkey!



ISRIC World Soil Information

Conclusions

Conclusions

- ➔ Traditional soil surveyors got it right! — distribution of soil classes is mainly controlled by DEM morphometry (especially hydrological parameters).
- ➔ Soil classification and polygon models of soils seem to make sense — in many parts of the world we see "soil groupings i.e. **soil bodies**"… but there are also many transition zones and individual patches… so it is really a hybrid model that we need to use to represent spatial variation.
- ➔ In the machine learning framework, much more time needs to be spent on preparing data / experimental design.

Conclusions II

- ➔ Our predictions could still be improved: **the most critical is to prepare a better (covariate) map of parent material and drainage classes**.
- ➔ We could also now ”easily” go beyond 250 m (100 m, 30 m) because there is so much remote sensing data in the public domain.
- ➔ SoilGrids250m (global models) can be merged with local predictions to produce best unbiased predictions of soil properties.

NATURE COMMUNICATIONS | ARTICLE OPEN

Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations

Christian Folberth, Rastislav Skalský, Elena Moltchanova, Juraj Balkovič, Ligia B. Azevedo, Michael Obersteiner & Marijn van der Velde

Affiliations | Contributions | Corresponding author

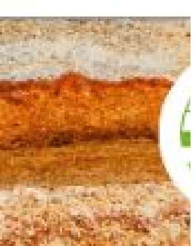
Nature Communications **7**, Article number: 11872 | doi:10.1038/ncomms11872
Received 29 July 2015 | Accepted 04 May 2016 | Published 21 June 2016

PDF | Citation | Reprints | Rights & permissions | Article metrics

Abstract

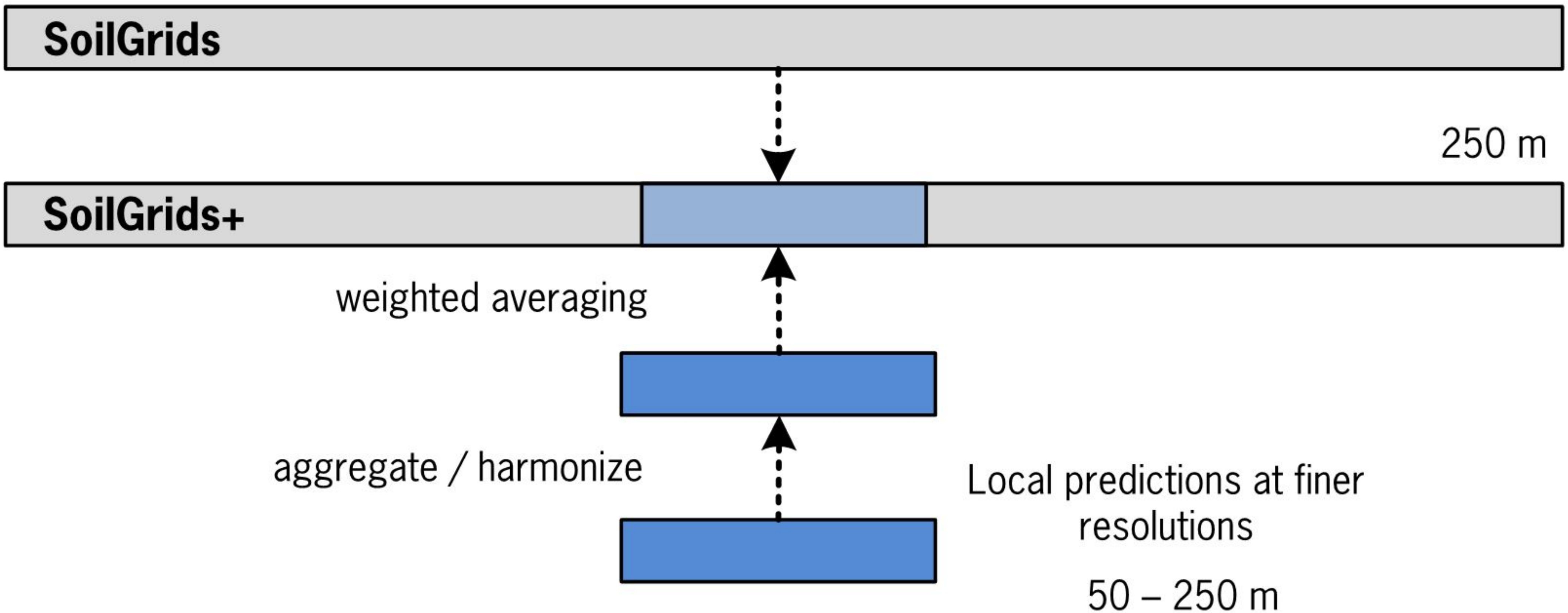
Abstract • Introduction • Results • Discussion • Methods • Additional information • References • Acknowledgements • Author information • Supplementary information

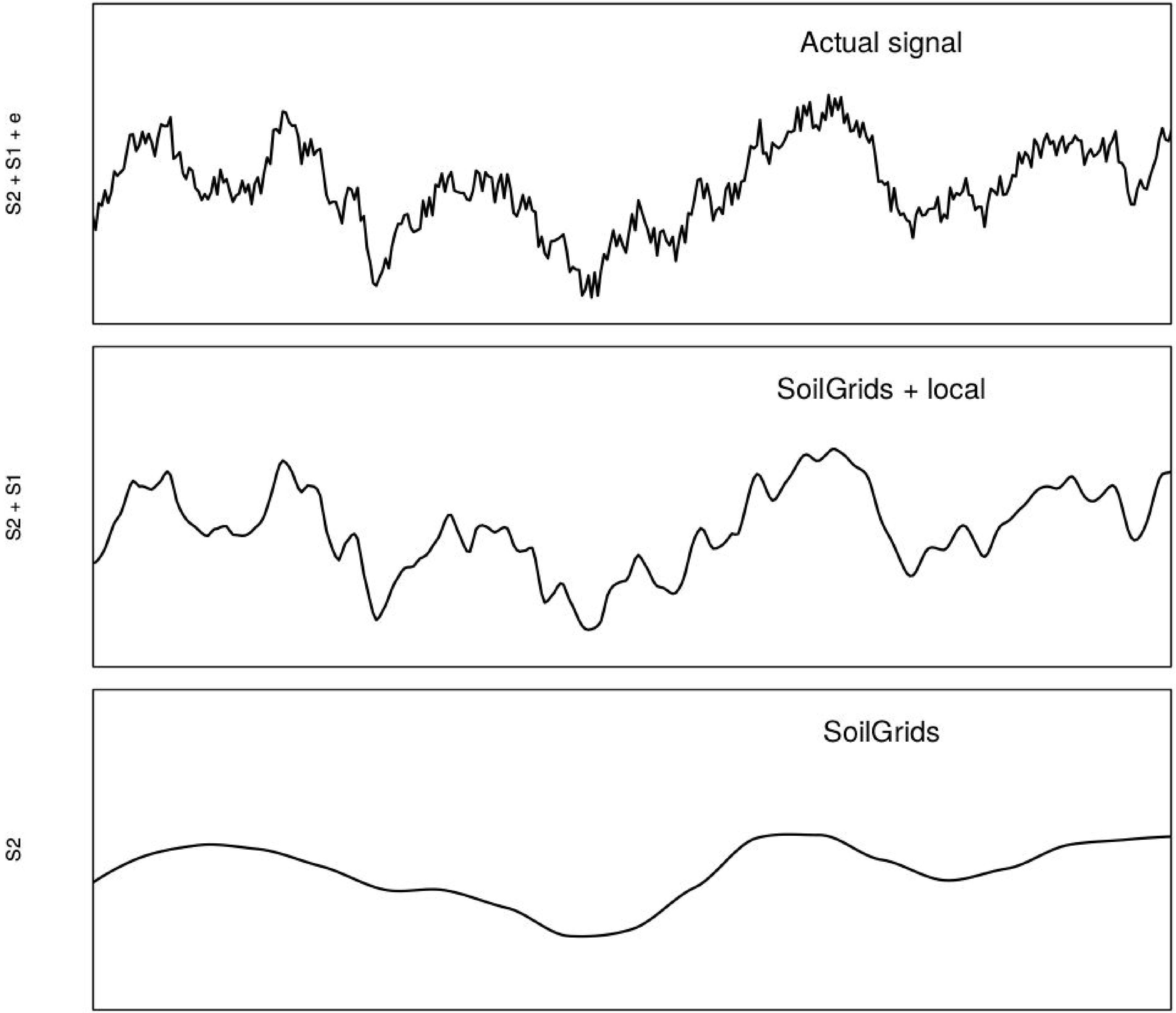
Global gridded crop models (GGCMs) are increasingly used for agro-environmental assessments and estimates of climate change impacts on food production. Recently, the influence of climate data and weather variability on GGCM outcomes has come under detailed scrutiny, unlike the influence of soil data. Here we compare yield variability caused by the soil type selected for GGCM simulations to weather-induced yield variability. Without fertilizer application, soil-type-related yield variability generally outweighs the simulated inter-annual variability in yield due to weather. Increasing applications of fertilizer and irrigation reduce this variability until it is practically negligible. Importantly, estimated climate change effects on yield can be either negative or positive depending on the chosen soil type. Soils thus have the capacity to either buffer or amplify these impacts. Our findings call for improvements in soil data available for crop modelling and more explicit accounting for soil variability in GGCM



SoilGrids+

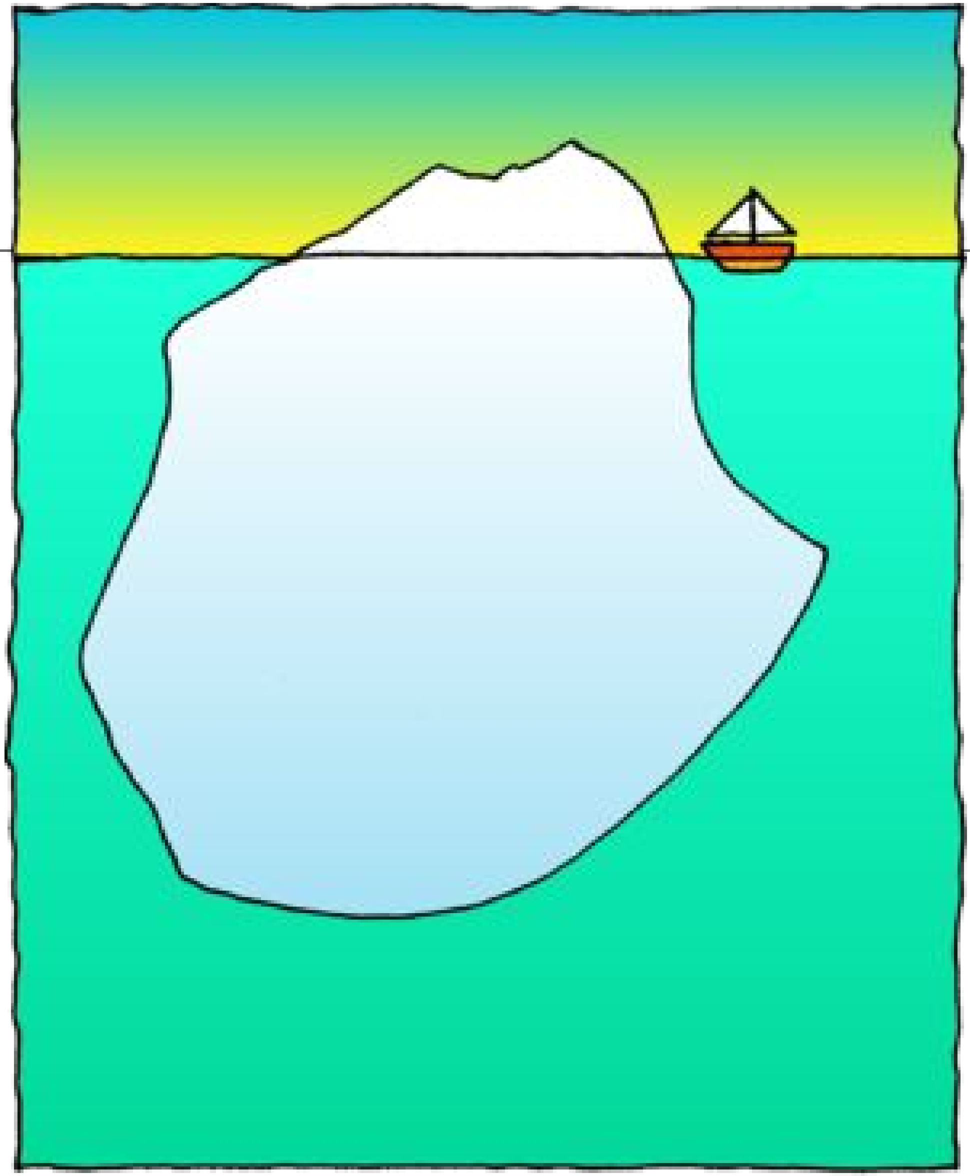
Global, consistent, complete and up-to-date gridded soil information





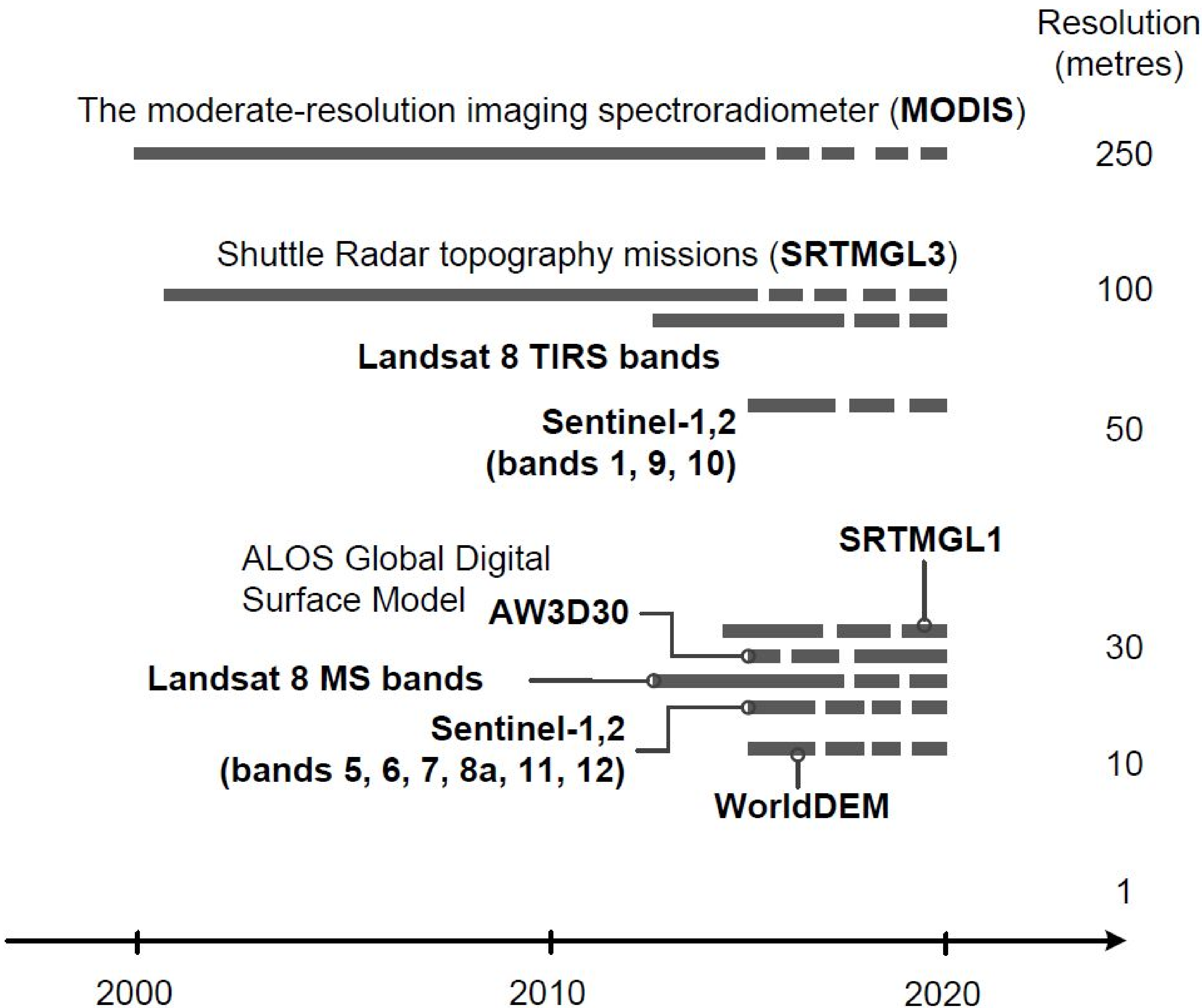
Actual signal
S2 + S1 + e
SoilGrids + local
S2 + S1
SoilGrids
S2

SoilGrids250m



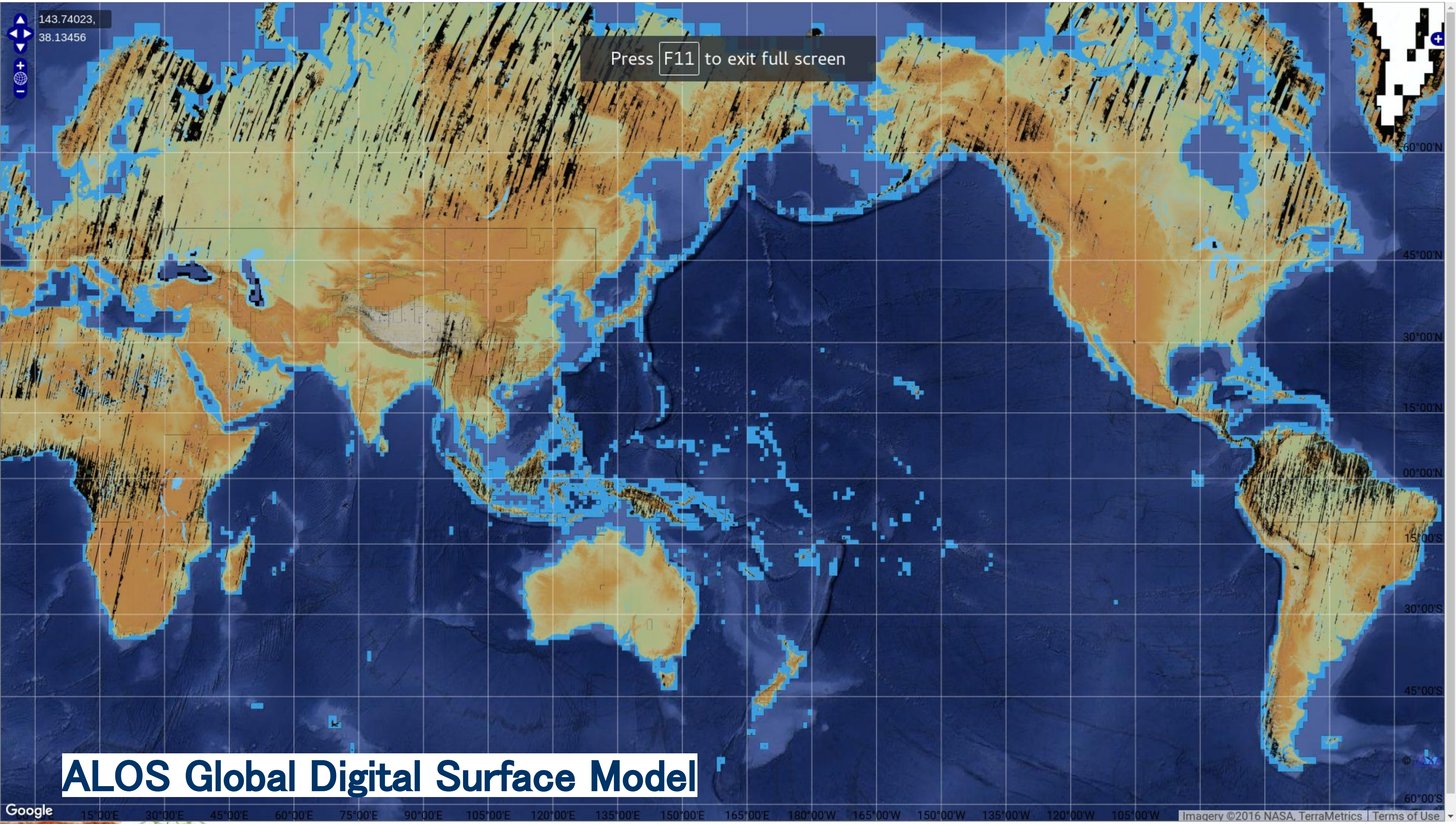
God's Word is like an iceberg-
there is more truth unseen than seen

We still know very little about world soils!



Resolution (metres)
The moderate-resolution imaging spectroradiometer (MODIS)
250
Shuttle Radar topography missions (SRTMGL3)
100
Landsat 8 TIRS bands
Sentinel-1,2 (bands 1, 9, 10)
50
SRTMGL1
ALOS Global Digital Surface Model AW3D30
30
Landsat 8 MS bands
Sentinel-1,2 (bands 5, 6, 7, 8a, 11, 12)
10
WorldDEM
1
2000
2010
2020

Towards 100 m, 30 m resolution...



ISRIC World Soil Information

Get ready for the Soil Data Revolution!