Journal of Geographic Information System, 2017, 9, 267-278
http://www.scirp.org/journal/jgis
ISSN Online: 2151-1969
ISSN Print: 2151-1950

A Large Scale GIS Geodatabase of Soil Parameters Supporting the Modeling of Conservation Practice Alternatives in the United States

Mauro Di Luzio[1], Mike J. White[2], Jeffrey G. Arnold[2], Jimmy R. Williams[1], James R. Kiniry[2]

[1]Blackland Research Center, Texas A&M AgriLife Research, Temple, Texas, USA
[2]Grassland Research Center, Agriculture Research Service, United States Department of Agriculture, Temple, Texas, USA
Email: mdiluzio@brc.tamus.edu

How to cite this paper: Di Luzio, M., White, M.J., Arnold, J.G., Williams, J.R. and Kiniry, J.R. (2017) A Large Scale GIS Geodatabase of Soil Parameters Supporting the Modeling of Conservation Practice Alternatives in the United States. *Journal of Geographic Information System*, **9**, 267-278.
https://doi.org/10.4236/jgis.2017.93016

Received: April 17, 2017
Accepted: June 3, 2017
Published: June 6, 2017

Abstract

Water quality modeling requires across-scale support of combined digital soil elements and simulation parameters. This paper presents the unprecedented development of a large spatial scale (1:250,000) ArcGIS geodatabase coverage designed as a functional repository of soil-parameters for modeling and comparison of water quality outcomes in the United States. The set of target models include: SWAT (Soil and Water Assessment Tool), APEX (Agricultural Policy Environmental Extender), and ALMANAC (Agricultural Land Management Alternatives with Numerical Assessment Criteria). This development relies on the Digital General Soil Map (DGSM) as the source of soil information, and leverages on architectural design and associated tools created for a companion product at higher resolution from which also was extended a procedure for refilling a large number of missing derived parameters. Outlined by regional watershed layouts and supported by GIS land use layers, the core product is developed using the File Geodatabase (FGDB) data structure, which brings, via customized Python-based tools, the data directly into geoprocessing workflows. The FGDB implement efficiently stores spatial soil features, tabular model elements and linked relationships, while seamlessly providing the environment for the extraction, spatial analysis, and mapping of the models' parameters. As an alternative, the composing spatial elements, polygons and multi-resolution rasters, and the models' elements are offered as a file-folder system of data with completely Open Source formats. Finally, this geographic database coverage provides support for the traditional large-scale and harmonized application of the models as well as an alternative to the higher resolution companion for areas where this information is still under development.

Keywords

Geodatabase, Soil, SWAT, APEX, ALMANAC

1. Introduction

Modern hydrology-based simulation models require the availability of representative key landscape parameters stored in adequate Geographic Information System (GIS) databases. Soil-related model parameters are traditionally derived from digital records of field-surveys.

In the United States, the most detailed source of such information is provided in extended area of the country by the Soil Survey Geographic Database (SSURGO) [1]. SSURGO is a Taxonomy-based, nationwide digital spatial database developed by the United States Department of Agriculture-Natural Resources Conservation Service (USDA-NRCS) at a range of scales between 1:12,000 and 1:24,000. Derived parameters have been extensively used to provide inputs to various hydrologic models [2] [3] including agricultural hydrology simulation models [4]. The spatially seamless application of SSURGO-based data is currently hindered by its partial incompleteness. In fact, the process of soil survey data collection and seamless completion is intrinsically lengthy and complex. This process could have been delayed, since USDA-NRCS collects, stores, maintains, and distributes soil survey information preferably for privately owned lands. Nevertheless, the development of SSURGO is continuously growing and the publication status updated and shared on line [5]. A basic remedy to the lack of information within incomplete areas is provided by the usage of large-scale source of information. This approach applied to agricultural hydrology models on watersheds and large geographic domains, provides controversial simulation results when compared to those obtained with higher resolution information [4] [6] [7] [8] [9] [10]. Large-scale soil attributes, however, have been successfully applied in hydrology in a large number of studies, and the value and usage of large scale soil data is still considered relevant [11]. It is important to notice that most of these applications were developed using dated data sources, such as the State Soil Geographic (STATSGO) [12], and methods to derive soil parameters for hydrology applications. Generally, there is a deficiency of up-to-date, documented, and functional GIS-based repositories of large scale modeling parameters for agricultural hydrology models.

In this paper we introduce the development and maintenance of a geodatabase coverage built to fulfill these purposes and therefore provide a repository of large scale spatial features and soil parameters for a set of agricultural hydrology models (SWAT, APEX, and ALMANAC). The core geodatabase is here named US-ModSoilParms-TEMPLE250000.

The applied approach is based on the application of a GIS-based data processing workflow to a selected collection of source spatial information. The overall procedure resembles and extends the development accomplished at the

higher resolution [13]. Fundamental differences from such development include the source input data (Section 2.1.1) and the adapted methodology of filling the source data gaps (Section 2.3.2). The overall framework is outlined in **Figure 1** and the following sections.

In the first section we present the characteristics of the implemented source data, models, GIS features and code. In the following section we present the results, and in the final section we discuss the highlights.

2. Materials and Methods

2.1. Data Sources

2.1.1. Digital General Soil Map

The USDA-NRCS National Cooperative Soil Survey (NCSS) developed the Digital General Soil Map (DGSM), or STATSGO2 [14], as a Soil-Taxonomy indexed representation of soil patterns in the landscape. DGSM is properly mapped at 1:250,000 scale in the continental U.S. (CONUS), Hawaii, Puerto Rico, and the Virgin Islands and 1:1,000,000 in Alaska. DGSM supersedes the State Soil Geographic (STATSGO) dataset, which included a limited number of soil attributes and outdated spatial features. DGSM includes a broad-based inventory of soils and no-soil areas designed for general planning and management uses covering state, regional, and multi-state areas. Data are distributed in the same packaging format and attributes of the current SSURGO data, which include both spatial and tabular data. Spatial data are delivered in ESRI shape file format and the World Geodetic System 1984 (WGS84) geographic coordinate system. Tabular data are in ASCII text files and pipe delimited fields. Spatial features outline soil general association units or Map Units (MUs), which refer to non-geo-referenced sub-unit groups (soil components, COMPs) accounted as a percentage of the area of the respective MU. Tabular data are logically linked to the spatial features and report physical and chemical soil properties as range and representative values. Information from seven (7) out of sixty-eight (68) tables of soil

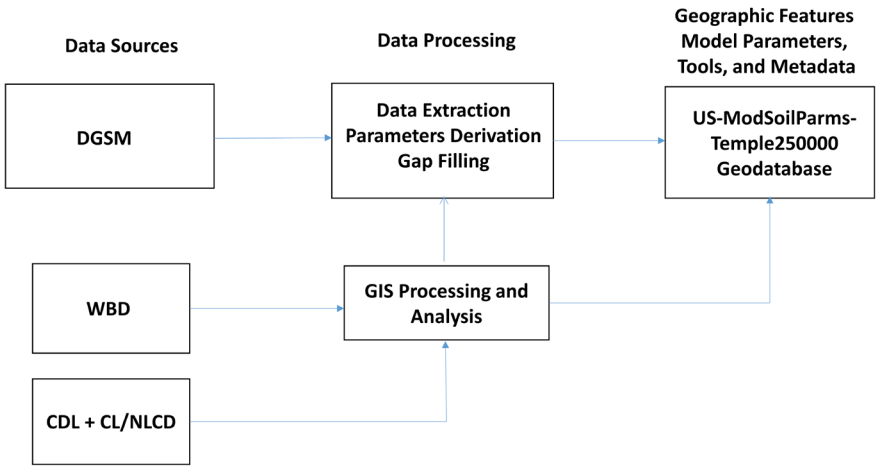

Figure 1. Data flowchart in the development of the US-ModSoilParms-TEMPLE250000 geodatabase. DGSM, Digital General Soil Map; CDL, Crop Data Layer; CL, Cultivated Land; WBD, Watershed Boundary Dataset; NLCD, National Land Cover Data Set.

attributes were used in this work, namely: 1) *Legend*; 2) *Mapunit*; 3) *Component*; 4) *Chorizon*; 5) *Chfrags*; 6) *Chtexturegrp*; and 7) *Muaggatt*. DGSM was downloaded as a single seamless national spatial and tabular dataset from the Internet at http://websoilsurvey.sc.egov.usda.gov.

2.1.2. Supporting Spatial Layers

The National Watershed Boundary Dataset (WBD) [15] was implemented to define the topographically-based hydrologic unit boundaries characterizing the domains of surface water flow. The WBD features used in this work include: a) Boundaries of 21 Regions (identified by 2-digit numbers): Regions 01-18 compose the CONUS, Alaska (19), Hawaii (20), Caribbean (21), whereas the South Pacific Islands (Region 22) are not covered by the DGSM layer; and b) A total of 2297 sub-basins identified by 8-digit numbers. The entire WBD GIS dataset was obtained from http://www.nrcs.usda.gov/wps/portal/nrcs/main/national/water/watersheds/dataset.

Land Use Land Cover (LULC) spatial layers were used to build local spatial statistics (MU level) and bias originally surveyed parameters when these most likely evolved, since the original collection date (e.g. Organic Matter). Cropland Data Layer (CDL) is a land cover product with more than one hundred (133) classes, 30 m resolution raster-based grid spanning the CONUS, with agricultural cover types in fine detail and with the remaining classes in less detail [16] [17]. These data sets were obtained from the NASS (National Agriculture Statistics Service) data server at http://nassgeodata.gmu.edu/CropScape along with the Cultivated Layer (CL), which explicitly distinguishes and reviewed the cultivated from non-cultivated land. The National Land Cover Data Set (NLCD) for the year 2001 [18], is a 16-class (additional four classes are used only in Alaska) land cover classification at a spatial resolution of 30 m obtained from the Multi-Resolution Land Characteristics Consortium (MRLC) at www.mrlc.gov to characterize the land use land cover in areas outside the CONUS, such as regions 19-21.

2.2. Models

The set of agricultural-hydrology simulation models include: 1) SWAT (Soil and Water Assessment Tool) model [19] designed for river basin and watershed hydrology simulation of water, sediment, nutrient, pesticide and fecal bacteria yields in agriculture-dominated landscapes and draining channels; 2) APEX (Agricultural Policy Environmental EXtender) [20], is designed for field- and farm-scale simulation of all the basic hydrological and chemical processes of farming systems and their interactions; and 3) ALMANAC (Agricultural Land Management Alternatives with Numerical Assessment Criteria) [21] is designed for field-scale simulation of the crop growth of a wide range of plant species and their competition. Commonly, these models require two types of input parameters: the first one (component level) represents the soil as a whole, while the second one depicts the soil across the vertical profile (layer level).

2.3. Data Processing

2.3.1. Geodatabase and Python

The ESRI ArcGIS File Geodatabase (FGDB) [22] version 10.1 provided the capability to handle and optimize the performances of the hosting data sets, while reducing the feature geometry and raster storage when compared to traditional shape files and personal geodatabases. Python language version 2.7 [23] and the ArcPy module provided by ArcGIS were utilized to access and operate the built-in geoprocessing routines and other tools offered by the Spatial Analyst extension [24] and ArcGIS 10.1. In this way, the compatibility with all the later versions was preserved.

2.3.2. Gap Filling

The companion development at high resolution identified a relatively large number of voids in the source data, which resulted in a large number of gaps in the compiled database records [13]. The procedure allowed the provision of an indexed set of scored-replacement parameters for the three models (SWAT, APEX, and ALMANAC) at the component and layer level.

At the first level, this was accomplished using a hierarchically-based methodology leveraging upon the Soil Taxonomy information and the geographic locations of the gaps. Texture-based replacement records were constructed and provided replacement at the layer level. In addition, proper default parameter records were consolidated for components referring to non-soil categories (e.g. badland, gullied land, lava flow, pits, and water). The overall set of replacements composed a database of Soil Taxonomy and Soil Texture indexed High Resolution Representative Values. This database was used to fill in the models' parameter gaps derived from original gaps contained in the source DGSM information.

The representative value (highest-scored) of each missing model parameter was retrieved by matching: a) the available Soil Taxonomy attribute from DGSM in a down-top search across the Soil Taxonomy-organized database (component level parameter); and b) the available Texture attribute (layer level parameter).

3. Results

The application of the procedure outlined in section 2.3.2 refilled the total number of parameter voids shown for each model in **Table 1**.

This step led to a spatial and tabular seamless outcome, which is provided in three means:

Table 1. Percentage and total model parameter voids refilled at the component and layer level.

Model	Component		Layer	
	All	Dominant	All	Dominant
SWAT	17.9% (74,198)	17.8% (6,820)	26.7% (1,084,633)	26.4% (99,678)
APEX	14.1% (277,363)	14.0% (25,512)	16.2% (2,139,258)	15.9% (194,943)
ALMANAC	30.5% (568,612)	29.9% (51,518)	24.9% (2,020,081)	25.1% (189,456)

1) 21-region (2-digit WBD HUC)-wide FGDBs composed by tiles outlined by the respective 1-km buffered 8-digit WBD polygon. Each tile includes the following elements: a) spatial part as Feature Class (ArcGIS format for vector data) and Geographic Coordinate System (GCS) WGS84 coordinates; b) spatial part as raster (Raster Datasets) at two resolutions (10-meter and 30-meter) in a locally proper Projected Coordinate System (PCS).; c) Three-Model attributes as FGDB tables with related component and layer level, and in relationship with the MU features (Figure 2); d) Metadata as Federal Geographic Data Committee (FGDC) Extensible Markup Language XML file, as detailed technical documentation containing User Guide and Tutorial document; and e) A set of Python-based tools, namely *SoilDatabases* toolset, grouped in an ArcGIS Toolbox, namely *GeoTEMPLE*, that can be used to manipulate and export the data as needed. The pool of constructed databases includes a total number of approximately 9,569 MUs and 103,626 components/soil series phases. The distribution of the number of components within the respective MUs for the entire set of geodatabases is depicted in Figure 3. The skewed geo-physical distribution of components does not affect the functionality of the geodatabase. The resulting total storage volume for the 21 regional FGDB is 6.2 GB for the complete version and slightly less (5.8 GB) for the *Lite* version, which includes only the dominant components (highest areal occupancy within the respective MU polygon) and the associated layers. By design, the structure and elements resemble and share the tools with the development at high resolution [13] and the data linkage depicted in Figure 4.

2) Two single FGDBs covering the CONUS (1 - 18) and all the features and internal organization listed above in point "1". One FGDB is provided in the *Lite* version (dominant components) with a storage volume of 2.66 GB and the second one with the complete set of components (2.69 GB).

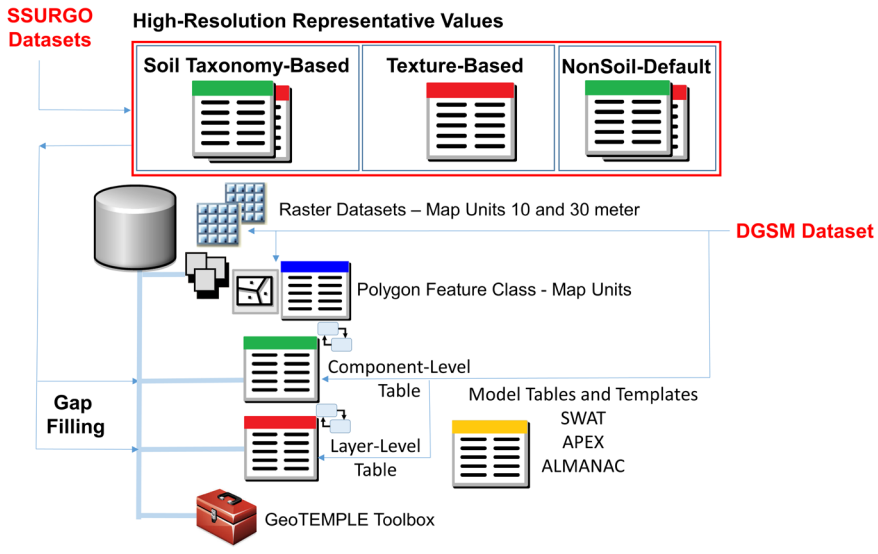

Figure 2. Elements of the US-ModSoilParms-TEMPLE250000 geodatabase and outline of the gap filling procedure.

3) A single file-folder-based framework (Figure 5), which hosts Open Source formatted 8-digit-tile spatial features and the associated model tables. We used raster GeoTIFF (Geographic Tagged Image File Format) files at the 10-meter and 30-meter cell size to represent the PCS MU rasters and ESRI Shapes files correspond to the geodatabase map unit Feature Classes in GCS. The model attributes were stored using dBASE tables. The complete system occupies a 24 GB storage volume.

4. Discussion

In this work, a geoprocessing work flow, previously developed using soil survey

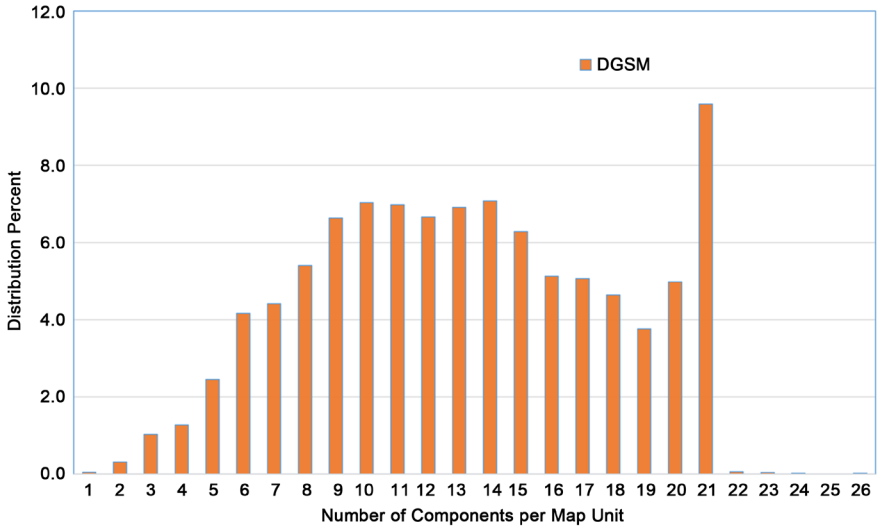

Figure 3. Percent distribution of the number of components per map unit.

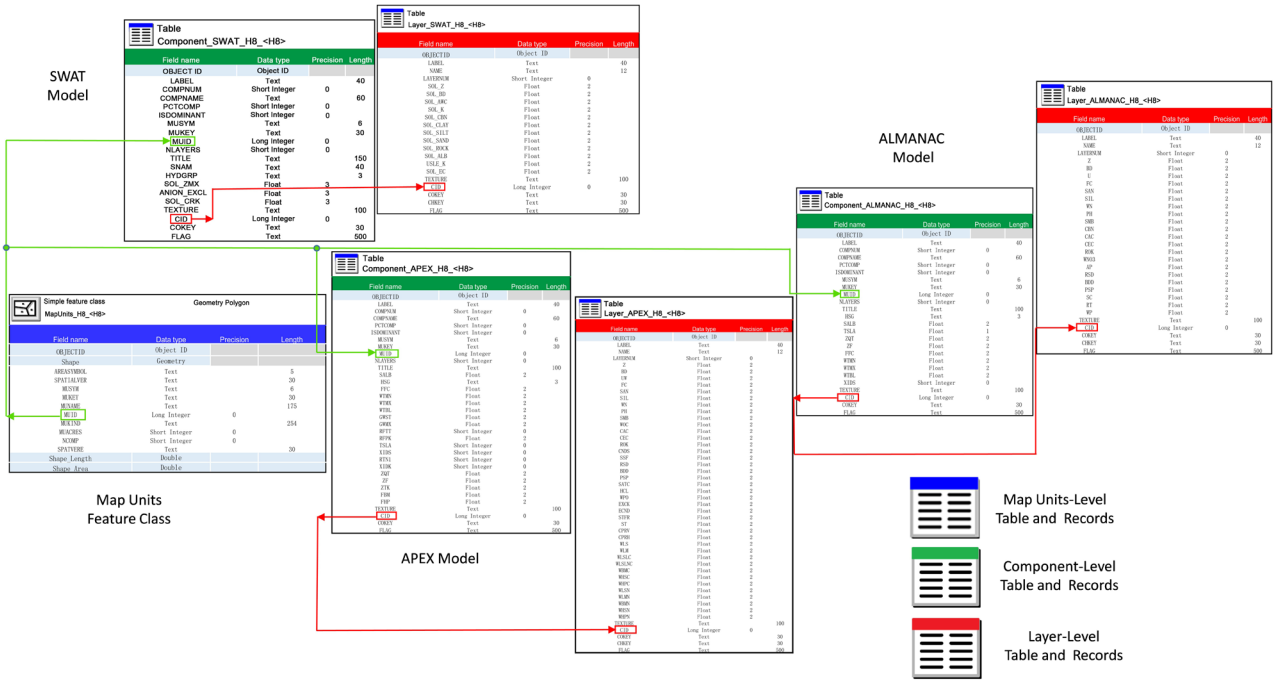

Figure 4. Database schema of the US-ModSoilParms-TEMPLE250000 geodatabase.

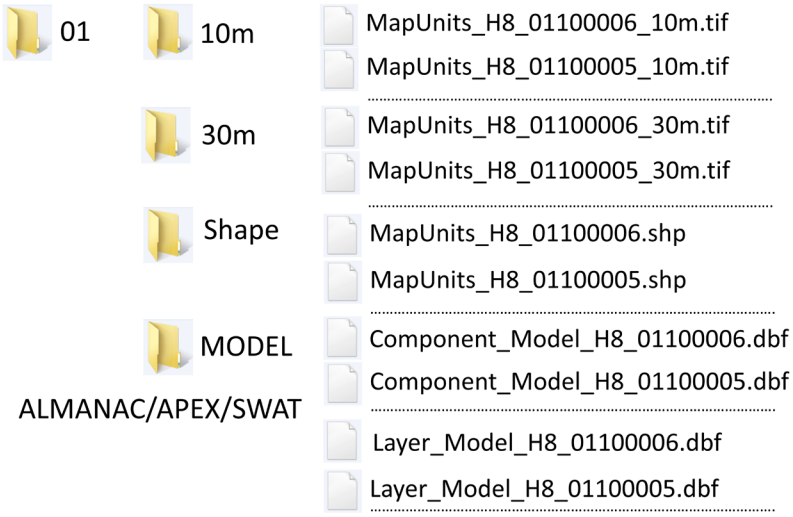

Figure 5. Region 01 excerpt of the file folder system hosting Open Source format elements of the US-ModSoilParms-TEMPLE250000 geodatabase.

data at high resolution (SSURGO), was implemented and extended using information at larger scale (DGSM). The resulting coverage provides seamless attributes to three agricultural hydrology models with geo-spatial functionality at the 1:250,000 spatial scale across the United States. The outcomes are provided in three alternative ways, each one with specific goals and functionalities.

The first alternative, core product of this development, is composed of a set of twenty one (21), drainage outlined regional FGDBs, tiled at the sub-basin level (8-digit), each one including essential items for geo-processing applications, such as: soil MU as Feature Classes polygons, Raster Datasets at two resolutions (10-meter and 30-meter) and three-model tables of specific parameters. The segmentation facilitates the management and application of the data organized in a framework inherited and shared by the same architecture and schema of the companion development at high resolution. The records earlier processed at high-resolution provided a Soil-Taxonomy database system of ranked replacement groups and records which turned out to be effective to fill numerous missing parameters originated at the DGSM level. Although the scale and density of the MU features did not required it, the tiled strategy was maintained to provide all-in-one across-scale solutions. In fact, while tile data sets are easier to maintain and to be included in geoprocessing frameworks, previously developed naming conventions and geodatabase items were directly portable and equally usable within this new development (**Figure 6**). Such items include in particular: a) A toolset for ArcGIS and associated referenced Python scripts to aggregate multiple tiles and/or any subset of the Feature Classes along with model attributes and relationships and/or extract and transfer to external model interface environments; b) Metadata and User Guide with tutorials.

The second alternative provides the composing data items (Feature Class, Raster Datasets, and Model Tables) within single ArcGIS FGDBs, each one cov-ering the CONUS. This option provides the same, yet lumped, items of the

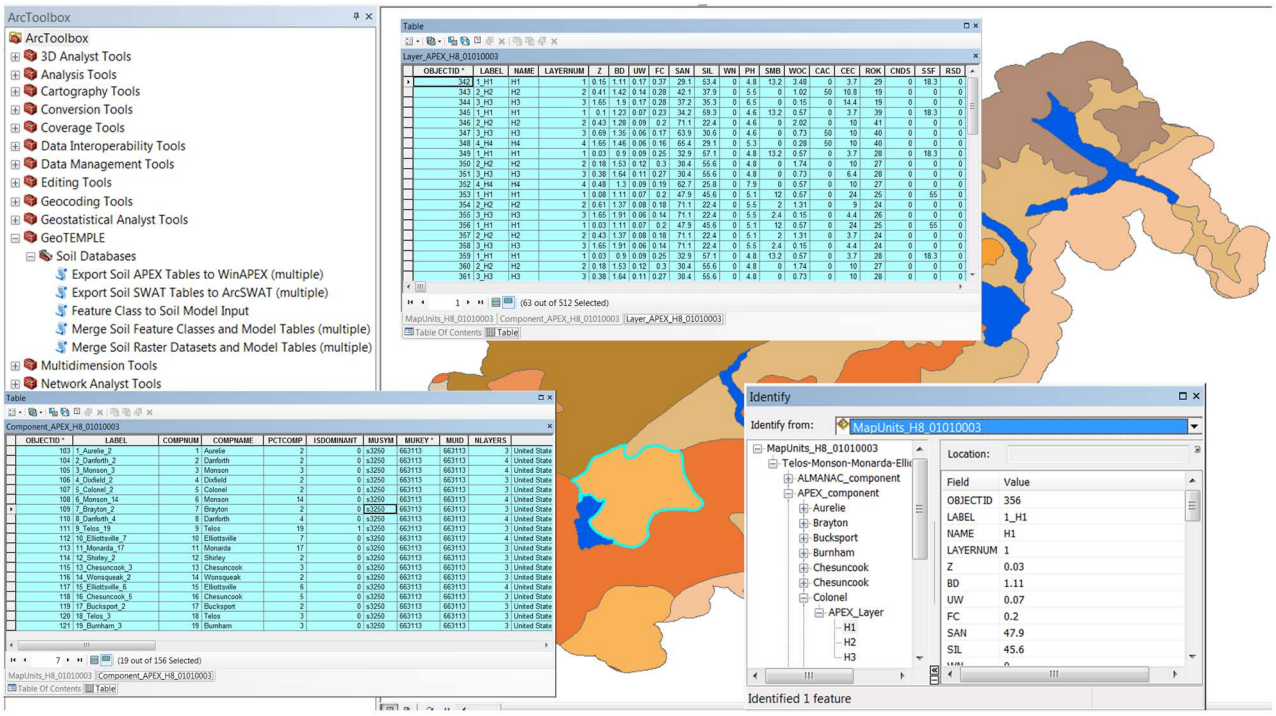

Figure 6. Spatial layers, models' parameters and GeoTEMPLE tools at the user's fingertip within the US-ModSoil Parms-TEMPLE250000 geodatabase.

regional-tiled geodatabases. This structure is advantageous for the quick interactive applications and or/analysis at the geographic extent of the entire CONUS. A simple example is shown in **Figure 7** for the top-soil Bulk Density parameter of SWAT, but any model parameter, both at the component and layer level, can be conveniently mapped and its distribution immediately evaluated and/or exported for further analysis and/or geoprocessing.

The third option offers to the Open Source software community accessibility within US-ModSoilParms-TEMPLE250000. Indeed, ESRI's software provides to programming languages such as Python and R (http://www.r-project.org) the capability to access and edit the FGDBs using ArcGIS site-packages (e.g. ArcPy and Bridge). However, the companion folder-based database framework developed using Shape files, GeoTIFF rasters, and dBASE tables, provides a comparable yet with expanded storage, offering direct access to the core content of this development.

5. Conclusion

Our work provides an unprecedented, large spatial scale, seamless and functional geographic database repository of soil parameters for three widely-used agriculture-hydrology simulation models in the United States. The data, assembled in three different fashions, along with customized tools, User Guide and details of this development, are planned to be available and continuously updated at http://soilandwaterhub.org/GeoTEMPLE.

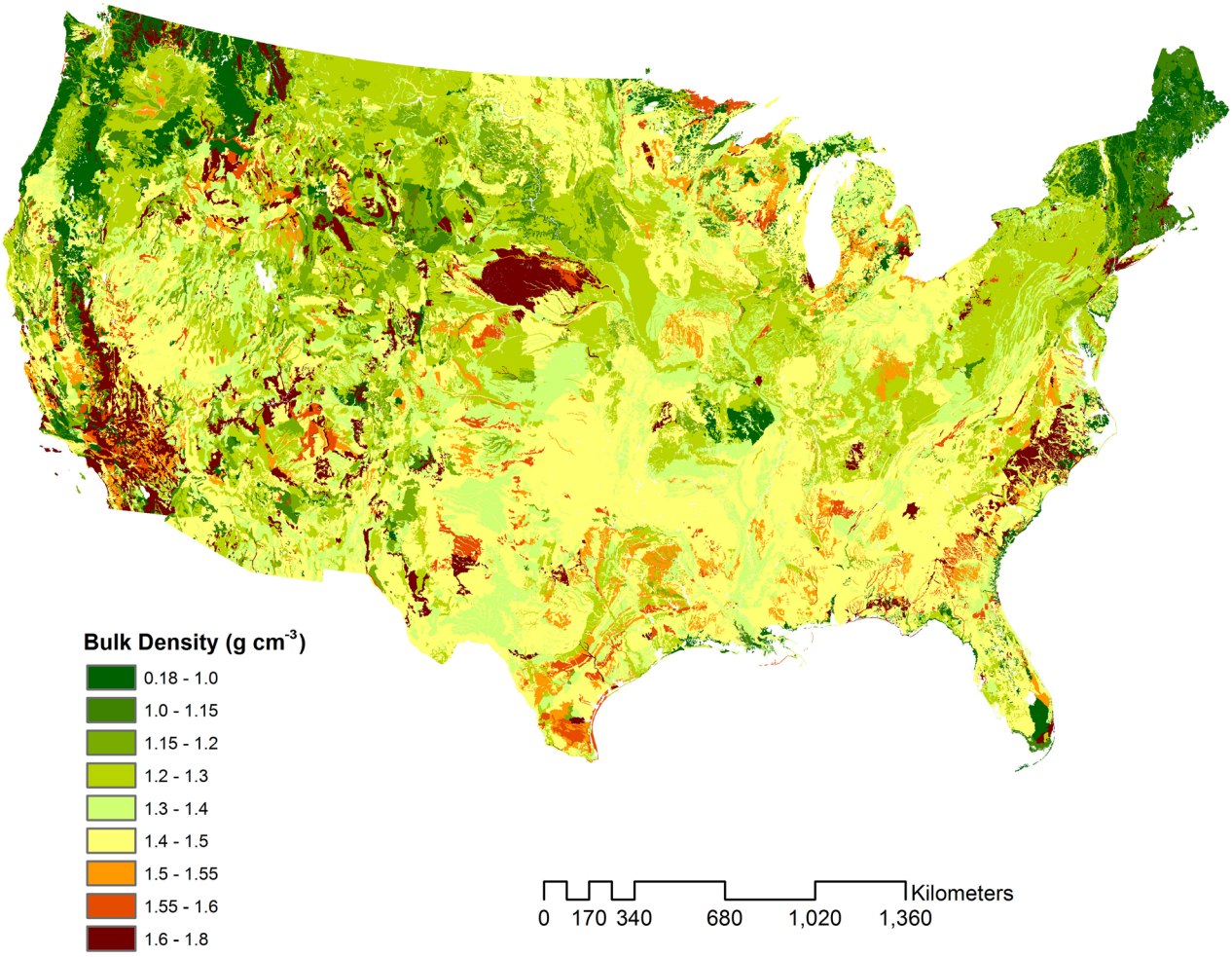

Figure 7. Map of the top soil-layer Bulk Density parameter for the SWAT model across the CONUS.

Acknowledgements

Soil Survey Staff, Natural Resources Conservation Service (NRCS) United States Department of Agriculture (USDA), for providing the DGSM dataset.

National Agricultural Statistics Service (NASS) United States Department of Agriculture (USDA), for providing the CDL and CL dataset.

Natural Resources Conservation Service (NRCS) United States Department of Agriculture (USDA), United States Geological Survey (USGS), and the Environmental Protection Agency (EPA), for providing the Watershed Boundary Dataset (WBD) for the United States.

United States Geological Survey (USGS) for providing the National Land Cover Database (NLCD).

References

[1] USDA-NRCS (2016) Soil Survey Geographic (SSURGO) Database. United States Department of Agriculture, Natural Resources Conservation Service: Washington DC, USA. http://websoilsurvey.sc.egov.usda.gov

[2] Anderson, R.M., Koren, V.I. and Reed, S.M. (2006) Using SSURGO Data to Im-

prove Sacramento Model a Priori Parameter Estimates. *Journal of Hydrology*, **320**, 103-116.

[3] Zhang, Y., Zhang, Z., Reed, S. and Koren, V. (2011) An Enhanced and Automated Approach for Deriving a Priori SAC-SMA Parameters from the Soil Survey Geographic Database. *Computers and Geosciences*, **37**, 219-231.

[4] Mednick, A.C., Sullivan, J. and Watermolen, D.J. (2008) Comparing the Use of STATSGO and SSURGO Soils Data in Water Quality Modeling: A Literature Review. *Research/Management Findings*, **60**, Wisconsin Department of Natural Resources. Bureau of Science Services, Madison, WI.

[5] (2017) Status Map of Soil Surveys Available from the Web Soil Survey. http://Websoilsurvey.nrcs.usda.gov/DataAvailability/SoilDataAvailabilityMap.pdf

[6] Peschel, J.M., Haan, P.K. and Lacey, R.E. (2006) Influences of Soil Dataset Resolution on Hydrologic Modeling. *Journal of the American Water Resources Association*, **42**, 1371-1389.

[7] Geza, M. and McCray, J.E. (2008) Effects of Soil Data Resolution on SWAT Model Stream Flow and Water Quality Predictions. *Journal of Environmental Management*, **88**, 393-406.

[8] Kumar, S. and Venkatesh, M. (2009) Impact of Watershed Subdivision and Soil Data Resolution on SWAT Model Calibration and Parameter Uncertainty. *Journal of the American Water Resources Association*, **45**, 1179-1196. https://doi.org/10.1111/j.1752-1688.2009.00353.x

[9] Mednick, A.C. (2010) Does Soil Data Resolution Matter? State Soil Geographic Database versus Soil Survey Geographic Database in Rainfall-Runoff Modeling across Wisconsin. *Journal of Soil and Water Conservation*, **65**, 190-199. https://doi.org/10.2489/jswc.65.3.190

[10] Sheshukov, A.Y., Daggupati, P., Douglas-Mankin, K.R. and Lee, M. (2011) High Spatial Resolution Soil Data for Watershed Modeling: 2. Assessing Impacts on Watershed Hydrologic Response. *Journal of Natural and Environmental Sciences*, **2**, 32-41.

[11] Williamson, T.N., Taylor, C.J. and Newson J.K. (2013) Significance of Exchanging SSURGO and STATSGO Data When Modeling Hydrology in Diverse Physiographic Terranes. *Soil Science Society of America Journal*, **77**, 877-889. https://doi.org/10.2136/sssaj2012.0069

[12] USDA-SCS (1994) State Soil Geographic (STATSGO) Database. Publication No. 1492. U.S. Department of Agriculture, Soil Conservation Service, National Soil Survey Center, Washington DC, USA.

[13] Di Luzio, M., White, J.M., Arnold, J.G., Williams, J.R. and Kiniry, J.R. (2017) Advancement of a Soil Parameters Geodatabase for the Modeling Assessment of Conservation Practice Outcomes in the United States. *International Journal of Geospatial and Environmental Research*, **4**, 1-13. http://dc.uwm.edu/ijger/vol4/iss1/2

[14] USDA-NRCS (2017) United States Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Staff. Web Soil Survey. http://websoilsurvey.nrcs.usda.gov/

[15] USDA-NRCS, USGS and USEPA (2016) United States Department of Agriculture (USDA) Natural Resources Conservation Service (NRCS), United States Geological Survey (USGS), and The United States Environmental Protection Agency (USEPA) Watershed Boundary Dataset for the United States, Washington DC, USA. http://datagateway.nrcs.usda.gov

[16] Boryan, C.G., Yang, Z., Di, L. and Hunt, K. (2014) A New Automatic Stratification

Method for U.S. Agricultural Area Sampling Frame Construction Based on the Cropland Data Layer. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **7**, 4317-4327. https://doi.org/10.1109/JSTARS.2014.2322584

[17] Johnson, D.M. (2016) A Comprehensive Assessment of the Correlations between Field Crop Yields and Commonly Used MODIS Products. *International Journal of Applied Earth Observation and Geoinformation*, **52**, 65-81.

[18] Homer, C., Huang, C., Yang, L., Wylie, B. and Coan, M. (2004) Development of a 2001 National Landcover Database for the United States. *Photogrammetric Engineering and Remote Sensing*, **70**, 829-840. https://www.mrlc.gov/pdf/July_PERS.pdf https://doi.org/10.14358/PERS.70.7.829

[19] Arnold, J.G. and Fohrer, N. (2005) SWAT2000: Current Capabilities and Research Opportunities in Applied Watershed Modeling. *Hydrological Processes*, **19**, 563-572. https://doi.org/10.1002/hyp.5611

[20] Williams, J.R., Harman, W.L., Magre, M., Kizil, U., Lindley, J.A., Padmanabhan, G. and Wang, E. (2006) APEX Feedlot Water Quality Simulation. *Transactions of the ASAE*, **49**, 61-73. https://doi.org/10.13031/2013.20244

[21] Kiniry, J.R., Williams, J.R., Gassman, P.W. and Debaeke, P. (1992) A General Process Oriented Model for Two Competing Plant Species. *Transactions of the ASAE*, **35**, 801-810. https://doi.org/10.13031/2013.28665

[22] Environmental Systems Research Institute (ESRI) (2009) The Top Nine Reasons to Use a File Geodatabase. https://www.esri.com/news/arcuser/0309/files/9reasons.pdf

[23] van Rossum, G. (2017) Guido van Rossum—Personal Home Page. http://www.python.org/~guido/

[24] Environmental Systems Research Institute (ESRI) (2013) ArcGIS Spatial Analyst. http://www.esri.com/software/arcgis/extensions/spatialanalyst