Journal of Hydrology 464–465 (2012) 467–476

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Approaches of soil data aggregation for hydrologic simulations

Yuzhou Luo [a,*], Darren L. Ficklin [b], Minghua Zhang [a,*]

[a] Department of Land, Air, and Water Resources, University of California, Davis, CA 95616, USA
[b] Department of Environmental Studies and Sciences, Santa Clara University, Santa Clara, CA 95053, USA

ARTICLE INFO

Article history:
Received 30 December 2011
Received in revised form 10 July 2012
Accepted 21 July 2012
Available online 31 July 2012
This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Ellen Wohl, Associate Editor

Keywords:
Hydrologic simulation
Spatial analysis
Spatial variability
Soil property
Soil taxonomy

SUMMARY

This study presents a comprehensive investigation of the State Soil Geographic Database (STATSGO) and the Soil Survey Geographic Database (SSURGO) soil databases for their applications in hydrologic modeling practices, and provides detailed instructions on soil data aggregation. Two types of soil data aggregations are developed and improved for the preparation of soil input data: (1) spatially-based aggregation for hydrologic models which require one representative soil profile for each spatial units of modeling simulations; and (2) taxonomy-based aggregation to handle the potential edge-matching issues in SSURGO, i.e. the artificial split of a soil type by the boundary of soil survey areas. The developed approach and program were applied at the spatial scales of soil survey area and watershed in California, U.S. representing hydrologic simulation domains with areas at the magnitudes of 1000 km^2 and 10,000 km^2, respectively. Edge-matching issues were identified for more than 20% of the involved soil map-units for both cases, about 90% among which were handled by the proposed approach in this study. The naming inconsistency in soil taxonomy was recognized as one of the causes for remaining issues. The reductions of soil map-units and components numbers before and after the aggregation are less than 10%, indicating that the proposed procedure has the capability to effectively handle the edge-matching issues while maintaining spatial resolution of the soil data.

1. Introduction

Soil data preparation is usually one of the first steps in initializing hydrologic simulations. In addition to the values of soil properties, the spatial resolution of soil input data also has significant effects on the model performance (Geza and McCray, 2007; Peschel et al., 2006; Singh et al., 2011). The Soil Survey Geographic Database (SSURGO) is the most detailed soil survey database for the United States. Developed by the U.S. Department of Agriculture (USDA), the database is available at a range of scales between 1:12,000 and 1:24,000 (USDA, 2011). SSURGO is considered an improved version of the State Soil Geographic Database (STATSGO) which is a generalized soil map at a scale of 1:250,000. Both databases share similar data structures and formats. The landscape is spatially segmented with soil map-units (MU). Soil in each MU is sampled for soil properties in various horizons, and reported by grouping into soil components. Soil components are not geo-referenced, but mainly characterized by the coverage fractions in the respective MU. Data for STATSGO and SSURGO are organized in ESRI shapefiles (for the spatial locations of MUs) and text files (for attribute data of soil properties), and freely available from the Soil Data Mart (http://soildatamart.nrcs.usda.gov/). SSURGO is still under development and does not cover the entire United States. Therefore, SSURGO may be used in conjunction with STATSGO for the full coverage of a large domain of hydrologic simulation (Gatzke et al., 2011).

STATSGO and SSURGO have been widely used in hydrologic models such as the Agricultural Nonpoint Source (AnnAGNPS) model (Polyakov et al., 2007), the Flood Hydrograph Package by the Hydrologic Engineering Center (HEC-1) (Smemoe et al., 2004), the Hydrological Simulation Program-Fortran (HSPF) (Johnson et al., 2003), the European Hydrological System (MIKE SHE) (Sahoo et al., 2006), and the Soil and Water Assessment Tool (SWAT) (Geza et al., 2009; Luo et al., 2008; Santhi et al., 2006; Wu and Johnston, 2007). However, the original data format in STATSGO and SSURGO is not suitable for most hydrologic models, and data extraction and pre-processing are usually required for preparing model input data of soil properties. Many hydrologic models, especially those field-scale or spatially distributed models with relatively small modeling units, ask for a single representative soil profile for each modeling unit which covers one or multiple MU(s). Most existing studies only focus on soil data extraction from STATSGO and SSURGO (Peschel et al., 2003, 2006; Sheshukov et al., 2009; Winchell et al., 2011); however, the geo-referencing of the soil data to the appropriate spatial scale of a hydrologic simulation is not sufficiently discussed. The widely used approach for soil data aggregation is to assume the soil properties in a specific

* Corresponding authors. Tel.: +1 530 754 2447 (Y. Luo), tel.: +1 530 752 4953 (M. Zhang).
E-mail addresses: yzluo@ucdavis.edu (Y. Luo), mhzhang@ucdavis.edu (M. Zhang).

0022-1694/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2012.07.036

area could be represented by the soil profile of the most extensive component, or "dominant component", which is the one with maximum fraction area in the MU. For example, Miller and White (1998) used STASTGO dominant soils texture grids for eleven soil layers for the conterminous United States in a continental distributed modeling system. A similar method was applied to SSURGO for deriving soil parameters for SAC-SMA (Sacramento Soil Moisture Accounting) model (Anderson et al., 2006; Zhang et al., 2011). However, each MU in both STATSGO and SSURGO may have multiple components, and sometimes the dominant component may only cover a small fraction of the total area of the unit. For example, the MU of CA355 in STATSGO has 21 components and the fraction area of the dominant component is only 8%. In this case, soil data taken from the dominant component is not necessarily representative to the entire MU.

A more reasonable soil profile can be generated based on area-weighted average of soil data in all available components within a certain spatial area. This process is referred as "soil data aggregation" (FGDC, 1997), and can be applied to a single MU or multiple MUs representing a spatial unit in the modeling domain. For example, USDA developed the "Soil Data Viewer" (http://soils.usda.gov/sdv/) for soil data aggregation for typical soil parameters based on area-weighted averaging. In the program, data processing is only performed for the top soil horizon while multiple horizons are usually required for hydrologic modeling. Similar methods have been previously applied to soil and hydrologic studies (Buell and Markewich, 2004; Davidson and Lefebvre, 1993; Smith et al., 2005). However, a consistent and systematic procedure for the aggregation of soil data is not yet available for the input data preparation of a general model. Additionally, some of the important issues in the data processing, e.g. the construction of soil horizons and the determination of accumulative soil depth, were not discussed in existing studies.

In addition to soil data aggregation, another hindrance of the application of SSURGO in hydrologic models has been identified recently. For the development of SSURGO, the United States is divided into soil survey areas (SSAa), and soil data for SSAs are incorporated into the SSURGO database as completed. Therefore, one soil type across the boundaries of SSAs would be artificially separated the segments with different MU identifications. For adjacent SSAs, soil surveys may be conducted in different time periods with different personale and surveying methods, leading to inconsistencies between soil surveys. This problem was first introduced in our previous study (Gatzke et al., 2011) and referred as the "edge-matching issue". Even with efforts to minimize the discontinuities along SSA boundaries (USDA, 1999), the issue remains in SSURGO version 2.0. To minimize the impacts of edge-matching problem in hydrologic simulations, Gatzke et al. (2011) proposed an approach to aggregate soil data by great group taxonomy. The method reduced edge matching errors between SSAs by 41% in the testing area of California's San Joaquin Valley watershed. However, significant reduction on spatial resolution of the soil data was also observed: the resultant soil data after taxonomy-based aggregation resulted in only 44 unique soil types (or 257 by intercepting with 15 delineated sub-basins), while in the original SSURGO there are more than 1000 MUs and components. After the processing, SSURGO is actually degraded to a spatial resolution even lower than STATSGO, for which 67 MUs and about 1000 components are included in this area. While the taxonomy-based aggregation is promising in minimizing the edge-matching errors, therefore, application of this approach to the entire study area of the hydrologic simulation may significantly compromise the advantage of SSURGO in representing the spatial variability on soil properties.

This study aims to develop an approach for soil data aggregation and for the appropriate handling of the edge-matching issues. Specific study objectives include: (1) to develop general method for soil data extraction and aggregation from STATSGO and SSURGO databases; (2) to investigate the potential impacts of the edge-matching issues on hydrologic simulations and develop soil aggregation approach to minimize the impacts; and (3) to implement the proposed approaches and test them at both SSA and watershed scales. Although SSURGO and STATSGO are used as data sources for the development and demonstration in this study, the resultant approaches are also anticipate to provide useful information for the soil parameter preparation based on other data sources such as the Canada National Soil Database, China Soil Scientific Database, and European Soil Database.

2. Soil data aggregation

In STATSGO and SSURGO, one MU is geo-referenced to a polygon in the spatial dataset, while the geographic locations of soil components within a MU are not explicitly documented. Data of soil properties are organized by component (table "*component*") and its horizons (table "*chorizon*") in the attribute tables. Taking the MU with key of 462782 in Merced County of California as an example, the geographic location of the MU can be easily determined based on the spatial data provided in SSURGO (Fig. 1). Eight components are identified in the MU, but the respective geographic information is not available. The components are characterized by their key, name, and percent coverage in the data table "*component*". Two of the components (key of 636287 with 50% coverage, and 636288 with 35% coverage) are considered as major components while other minor components are called "inclusions". Only major components are provided with soil horizon data in the data table ("*chorizon*"). These data are frequently used in hydrological simulations, and could be categorized as (a) horizon-dependent properties (available water content, bulk density, saturated hydraulic conductivity, clay/sand/silt contents, organic matter content, erodibility, and electrical conductivity), and (b) surface properties (hydrologic soil group and soil albedo).

Soil data aggregation is the process in which soil data from all components of one or multiple MU(s) representing a spatial unit of the modeling domain was aggregated into one soil profile, so that the resultant soil data could be geo-referenced and used in spatially distributed modeling. Soil data aggregation in this study is based on the depth-slicing algorithm, which has been previous used for the data processing of STATSGO and SSURGO (Beaudette and O'Geen, 2011; Gatzke et al., 2011; Luo, 2006; Miller and White, 1998; USGS, 1995). In general, the representative value of a soil property at a given depth, $s(z)$, is calculated as area-weighted average of the values at the same depth in all involved components:

$$S(z) = \frac{\sum_i [S_i(z) \cdot f_i]}{\sum_i f_i} \tag{1}$$

where i is a running index for the components with data at the depth z, $s_i(z)$ is the corresponding properties at depth z in component i, and f_i is the fractional area of i. Eq. (1) is also appropriate for soil properties only defined at the surface, by setting $z = 0$. For hydrologic soil groups which indicate the minimum rate of infiltration of the soil type, the descriptive grouping are aggregated based on the numerical conversion with group $A = 1$, $B = 2$, $C = 3$, and $D = 4$. These values are averaged by following Eq. (1), and then converted back to letters using the same conversion (Burns et al., 2004; USGS, 1995).

The general depth-slicing algorithm for soil data aggregation was improved in this study by investigating the cumulative soil depth in the soil profile after aggregation. Most existing studies set this depth as the maximum value of cumulative soil depth in

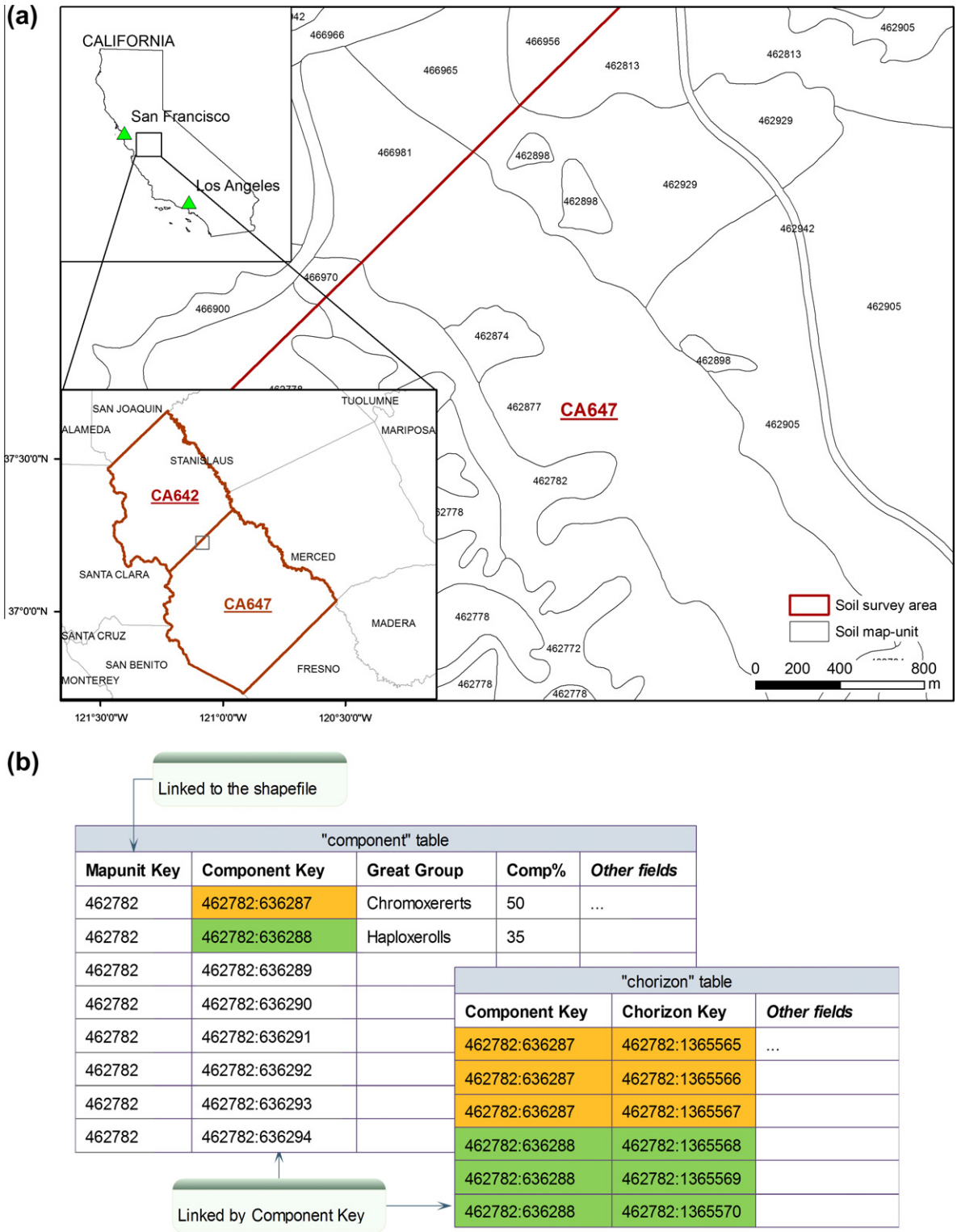

Fig. 1. Relationships of soil map-unit (MU), components, and horizon data of soil properties, with MU of 462782 in soil survey area of CA647 as an example. (a) Shows the geographic location of the MU; (b) shows attribute data associated with the MU.

all involved components. This setting is designed to completely capture the available data in all soil horizons, but may artificially increase the water storage capacity in the resultant soil profile compared to the original components. For example, when two components with 500 cm and 1000 cm soil depths, respectively, are aggregated, the resulting soil profile has a depth of 1000 cm. By applying this algorithm to a large region, the overestimation on the water storage capacity could be accumulated to have significant effects on hydrologic simulations. Previous studies indicated that hydrologic models are highly sensitivity to the soil parameters describing soil–water capacity, including field capacity, available water content, and wilting point. The use of soil input data with higher soil–water storage potentially decreases surface and subsurface runoff, especially at larger spatial scales.

In this study, the cumulative soil depth of the representative soil profile is determined by matching the total saturated water content in the aggregated soil profile to that in the original components to be aggregated:

$$\int_0^{zmx} s_\theta(z) = \frac{\sum_i \left[\int_0^{zmx_i} (s_{\theta,i}(z) f_i)\right]}{\sum_i f_i} \tag{2}$$

$$S_{\theta,i}(z) = 1 - s_{BD,i}(z)/\rho_g \tag{3}$$

where s_θ (dimensionless) is the porosity, zmx_i (cm) is the cumulative soil depth in component i, zmx (cm) is the cumulative soil depth to be determined for the aggregated soil profile, and s_{BD} and ρ_g (g cm^{-3}) are soil bulk density and grain density (ρ_g = 2.65), respectively. There is no analytical solution in a general form for zmx in Eq. (2), but the value of zmx can be determined by trial tests with 1 cm increment along the soil profile. The proposed algorithm in Eq. (2) generates an appropriate zmx by remaining the same total soil–water storage as that reported in all involved components during soil data aggregation.

As a demonstration of the depth-slicing approach and the improvement on determining the cumulative soil thickness proposed in this study, soil data in the two major components in the MU462782 (Fig. 1 and Table 1) are aggregated. Based on the reported bulk density, the total saturated water content in all involved components can be determined as 39.34 cm (the right-hand side of Eq. (2), details in Table 1). Illustrated in Fig. 2a is the area-weighted averages of the bulk density determined by Eq. (1), while in Fig. 2b is the corresponding profile of porosity by Eq. (3). By matching the total saturated water content of the aggregated soil profile to the reported value of 39.34 cm, the cumulative soil depth for aggregation (zmx) was determined to be 101 cm (Fig. 2c). Compared to the maximum soil depth with reported data in the original components (i.e. 119 cm Table 1), the proposed approach in soil data aggregation avoided an overestimation of total saturated water content in the representative soil profile by 19% (Fig. 2b).

The soil property aggregated with Eq. (1) is a continuous step function (Fig. 2a). Some hydrologic simulators only allow a limited number of soil horizons as input data. For example, 10 soil horizons are the maximum number of soil horizons in SWAT. Therefore, the resultant soil properties from Eq. (1) may have to be discretized according to the prescribed soil horizon structure. In the development of CONUS-SOIL database (Miller and White, 1998), the STATSGO data were distributed into a set of 11 standard soil layers with depths of 5, 5, 10, 10, 10, 20, 20, 20, 50, 50, and 50 cm. Similarly, the soil horizon structure used in Gatzke et al. (2011) had 8 layers with depths of 5, 5, 5, 15, 30, 30, 60, and 100 cm for SSURGO data. For each layer, the representative values of soil properties are calculated as average of the properties within the layer:

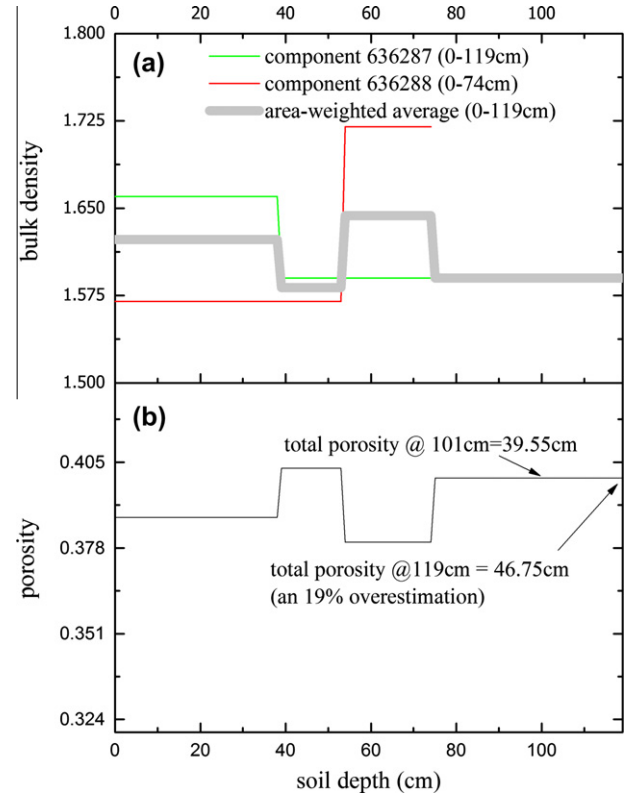

Fig. 2. Demonstration of the soil aggregation approach for the soil map-unit of 462782 with two major components, (a) bulk density profiles in the two components and as the area-weighted average; (b) porosity profile calculated from the area-weighted average of bulk density.

$$S(k) = \frac{\int_{z_{k1}}^{z_{k2}} s(z)}{z_{k2} - z_{k1}} \tag{4}$$

where k is layer index with soil depth from z_{k1} to z_{k2}, and $S(k)$ is the layer-averaged soil property.

Based on the depth-slicing algorithm and the improvements proposed in this study, the detail processes in soil data aggregation are described as follows:

(1) To identify MUs according to the spatial units of the modeling domain, and extract soil data for all major components of the involved MUs.
(2) To determine the cumulative soil depth for the representative soil profile (zmx) by Eq. (2).
(3) To aggregate required soil properties for each centimeter of the soil profile, based on Eq. (1), until reaching zmx.

Table 1
Bulk density and porosity for the MU462782 in the SSURGO database.

Horizon	Component 636287 (coverage = 50%)			Component 636288 (coverage = 35%)		
	Depth (cm)	Bulk density (g cm^{-3})	Porosity	Depth (cm)	Bulk density (g cm^{-3})	Porosity
1	0–38	1.66	37.36%	0–53	1.57	40.75%
2	38–119	1.59	40.00%	53–74	1.72	35.09%
3	119–130	NA		74–84	NA	

Notes: Bulk density was retrieved from SSURGO while porosity was calculated by Eq. (3). The total porosity (or saturated water content) reported in component 636287 is calculated as 37.36%*38 + 40.00%*(119–38) = 46.60 cm. Similarly, the value in 636288 is 28.97 cm. Therefore, the area-weighted average of the saturated water content in the two components is calculated by Eq. (1) as (46.60*50% + 28.97*35%)/(50% + 35%) = 39.34 cm.

(4) (If applicable) to compute representative soil properties for each soil layer according to the prescribed layer structure up to *zmx*.

3. The edge-matching issue

As previously mentioned, the edge-matching issue is observed in SSURGO when one soil type is divided by SSA boundaries and assigned as multiple MUs. The issue is illustrated in Fig. 1a. MU pairs such as [466970, 462782], [466981, 462877], [466965, 462929], [466956, 462813] represent landscapes artificially dissected along the boundary between SSAs of CA642 ("Stanislaus County, California, Western Part") and CA647 ("Merced County, California, Western Part"). The issue is potentially associated with any pair of MUs which are adjacent and across SSA boundary. In spatial analysis, the split MUs can be defined as two polygons sharing the line segment of the SSA boundary. Even though soil data for both sides of the boundaries are reliable, the artificial split has potential effects on hydrologic simulation, especially for the modeling projects with a simulation domain over multiple SSAs. One example is the Hydrologic Response Unit (HRU) distribution. HRU is a concept widely used in watershed and catchment scale models to reduce modeling complexity. In a sub-basin, areas with similar hydrologic characteristics of land use, soil, and/or slope are lumped into a single unit. This process requires continuous and consistent soil data coverage for determining the expansive soil types in the sub-basin, since only major soil types with spatial coverage larger than a threshold would be considered in the HRU distribution. The artificial splits across SSA boundaries reduce the fractional area of a soil type and thus potentially have effects on the soil types selected in subsequent HRU distribution and model simulation. The edge-matching issues may have serious implications, dependent on the total fraction of involved MUs, when SSURGO data is used in modeling surface hydrology at large scales.

Soil aggregation by great group taxonomy was proposed in our previous study to minimize the impact of edge-matching issues on the parameterization and simulation of hydrologic models (Gatzke et al., 2011). The basic assumption is that, the artificially dissected MUs could be recovered by aggregating their components with common great group taxonomy. Therefore, the components of one soil type across a SSA boundary can be programmatically identified and aggregated based on the common taxonomy. The demonstration in Fig. 3 and Table 2 explains this process with the adjacent MU pair [466970, 462782] across SSAs of CA642 and CA647 as an example. A common taxonomy of "Haploxerolls" was identified in both MUs and aggregated as one soil type. The processed data indicated that "Haploxerolls" is the dominant taxonomy (fraction area of 35%) in the combined MU, while it is not dominant in either MU of the original data (Table 2). This example demonstrated the impacts of artificial splitting of a soil type by SSAs on the soil data interpretation, as well as the capability of the proposed soil data aggregation in handling the edge-matching issues.

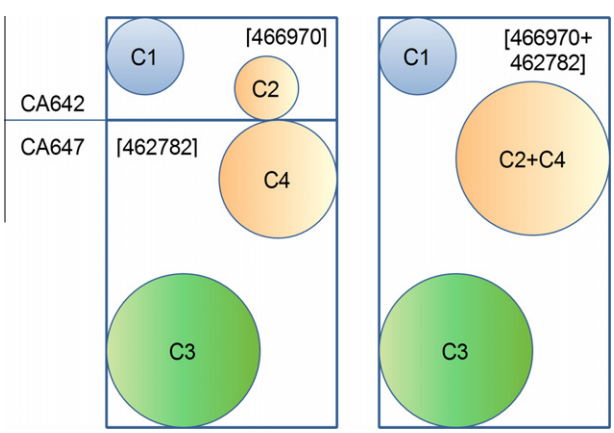

Fig. 3. Demonstration (not drawn to scale) of soil data aggregation for handling the edge-matching issue. Shown in the example are two adjacent soil map-units (MUs) on the boundaries of two soil survey areas: 466970 in CA642 and 462782 in CA647 (Fig. 1a). Major components include 646276 (C1) and 646277 (C2) for MU466970, and 636287 (C3) and 636288 (C4) for MU462782. C2 and C4 share the common great group taxonomy of "Haploxerolls" and combined as a one soil type in this study. See Table 2 for more details.

Soil taxonomy provides a robust framework for soil grouping by physical and chemical properties (USDA, 1999). For hydrologic modeling, the utilization of taxonomy-based aggregation refines the operational definition of soil type with soil classification rather than artificial structure of soil database or soil sampling. Aggregation of soil data by taxonomy generates continuous and consistent soil coverage to ensure accurate distribution of HRU and other soil data related configurations for hydrologic simulations. Appropriate characterization of soil properties within the context of hydrologic modeling have large impact on the model performance on water and water quality processes. Gatzke et al. (2011) indicated that, for instance, SWAT predictions on surface hydrologic processes including surface runoff and sediment yield from the San Joaquin Valley watershed could be improved with soil data aggregated by taxonomy, compared to that by MU.

The primary problem in using taxonomy to solve edge-matching issue is the loss of spatial resolution in soil data. In a large area of hydrologic simulation domain, the total number of distinct taxonomies is usually significantly less than that of MUs. For example, in our previous case study (Gatzke et al., 2011), taxonomy-based aggregation of soil data was applied to all MUs in the study area even though the majority of MUs were not located on SSA boundaries. Therefore, the processing could be significantly simplified by ignoring the spatial analysis such as the determination of whether a MU is on the SSA boundaries and whether two MUs are adjacent. However, taxonomy-based aggregation for all components in the domain of hydrologic simulation significantly reduced the spatial variability in soil properties. By applying this method to the San Joaquin Valley watershed, the original >1000 MUs in SSURGO were aggregated into only 44 taxonomies, or 257 based on the intersection with subbasins (Gatzke et al., 2011). Consequently, the advantage of high resolution soil data in SSURGO was actually not reflected in this case.

Therefore, the major limitation of the taxonomy-based soil data aggregation in the previous study is that spatial locations and relationships of MUs were not considered. In order to take advantage of the taxonomy-based aggregation in handling the edge-matching issue while to remain the high spatial resolution of SSURGO, the following improvements are applied to the previous approach: (a) only MUs located on the SSA boundary are considered in the soil data aggregation; and (b) aggregation is conducted for each group of adjacent MUs across the SSA boundaries. Three steps are involved in the new approach:

(1) to identify pairs of MUs which share the line segments of SSA boundaries;
(2) to group the identified MU pairs with common boundaries as spatial units for potential soil data aggregation. Grouping is necessary because a MU may be adjacent to multiple MUs on the other side of a SSA boundary (one-to-multiple), and even multiple-to-multiple cases may exist. For example, if MU1 in SSA1 shares the SSA boundary with MU2 and MU3 in SSA2, there will be two MU pairs of [MU1, MU2] and [MU1, MU3] but considered only as 1 group; and

Table 2
Data and statistics for the components involved in the demonstration (Fig. 3) of soil data aggregation for handling the edge-matching issue. Dominant components in the respective MU are highlighted.

Map-unit key (area in acre)	Component key	Great group	Component % coverage in the MU
466970 (1000)	C1 = 646276	Haploxererts	50
	C2 = 646277	Haploxerolls	35
462782 (1670)	C3 = 636287	Chromoxererts	50
	C4 = 636288	Haploxerolls	35
Combined (2670)	C1	Haploxererts	19
	C3	Chromoxererts	31
	C2 + C4	Haploxerolls	35

(3) to perform soil data aggregation. Spatially, the MU groups identified in (1) are merged together to be a new MU. For attribute data of soil parameters, the components with common taxonomy are aggregated for a single representative soil profile as a new component. Finally, fraction areas of all components in the new MU are recalculated.

The MUs grouped in the step (2) are considered to be associated with the edge-matching issues, while only a portion of them could be actually handled in the subsequent step (3). The fraction (number of handled over identified MUs) is reported as an indicator for the performance of the proposed approach in handling edge-matching issues. One special case is for the MUs of water. Some SSA boundaries are actually delineated by major rivers. For example, the boundaries of CA642 and CA644, CA647 and CA648, CA651 and CA653, and CA651 and CA654 in California are along the San Joaquin River. In the proposed data processing, MUs representing water are excluded from the identification of MU pairs for taxonomy-based aggregation, so that the SSA boundaries by streams are automatically excluded in handling edge-matching issues.

4. Computer implementation

For automating the soil data aggregation processes proposed in this study, a program was developed based on ArcGIS ArcObjects for Microsoft .NET framework (Fig. 4). The program was designed for the following processes:

(1) Soil data extraction: To extract data of soil properties from the tabular source data of STATSGO or SSURGO into an output database in Microsoft Access format. The extracted data is organized by component, i.e. each record is for one component with both surface properties (MU key, component key, great group taxonomy, area, fraction area in the MU, number of horizons, hydrological soil group, albedo, and cumulative soil layer thickness) and horizon-dependent properties (horizon thickness, bulk density, available water content, saturated hydraulic conductivity, erodibility, and percent contents of organic carbon content, clay, silt, sand, and rock fragment content). Included soil data and data structure in this study are following the template in ArcSWAT (Winchell et al., 2011), a pre-processor for SWAT.
(2) Edge-matching issue handling: If the input data includes multiple SSAs, this option is available for handling potential edge-matching issues. MU pairs which are adjacent with each other and belong to different SSAs are identified by spatial analysis tools provided in the ArcObjects. All identified pairs are processed based on the approach discussed in the last section; and
(3) Soil data aggregation: Soil data can be aggregated with the provided map for the delineation of modeling domain. A representative soil profile will be generated for each of the hydrologic modeling units such as catchments and fields.

The program generates data in Microsoft Access format by following the soil data structure required by ArcSWAT. Therefore, the program output can be directly applied as user soil data for HRU distribution with ArcSWAT to develop a SWAT project. It is worthy to note that application of the soil data processing documented in this study is not limited by a specific hydrologic model. In fact, a comprehensive set of soil properties is extracted and aggregated in the program (listed above). Those properties are generally sufficient for parameterizing any hydrologic simulations. The program will be available on the developers' website (http://agis.ucdavis.edu/) for public access. Meanwhile, individual requests could be fulfilled by the corresponding authors.

5. Case studies

5.1. Two adjacent SSAs

Two adjacent SSAs of CA642 (with total coverage of 1580 km^2) and CA647 (2410 km^2) (Fig. 1a) are used as the first demonstration of the aggregation of soil data and the handling of edge-matching issue. There are 106 and 189 distinct MUs in the two SSAs, respectively, and some are represented as multiple-part polygons. Both the SSAs have 25 MUs located at their shared boundaries, accounting for 18% of the area in CA642 and 25% of CA647. For the two SSAs as a whole, MUs which are associated with the edge-matching issues account for 22% areas of the total area. The program developed in this study identified 27 MU pairs of MUs which are across the SSA boundary and adjacent to each other and generated 22 potential groups of MUs for soil data aggregation (Table 3). Finally, common great group taxonomies in different MUs were observed in 19 MU groups, for which the components with common taxonomies were aggregated. We conclude that 86% (19 divided by 22) of the edge-matching issues in the study area of CA642 and CA647 are handled with the proposed approach. In the remaining three groups (#1, 10, and 20) (Table 3), no common great group taxonomy was found. Further investigation indicated that the change of taxonomy classification over time is the main reason for the approach inability in solving edge-matching problems in the 3 groups. According to the 10th Edition of Keys to Soil Taxonomy by USDA, Chromoxererts was classified as Haploxererts, and Xerochrepts as Haploxerepts (Culman et al., 2010; USDA, 2006). In addition, classification was changed from Haplaquolls to Endoaquolls according to changes in Taxonomy in 1992 (USDA, 2009). With above information, the edge-matching issues in the three groups of #1, 10 and 20 (Table 3) could also be handled based on the proposed approach. This suggested that the investigation of changes of taxonomy classification over time could significantly improve the proposed procedure in handling the edge-matching issues.

5.2. The san Joaquin Valley watershed

The San Joaquin Valley watershed is located in the middle of California' Central Valley. The total area of the watershed is

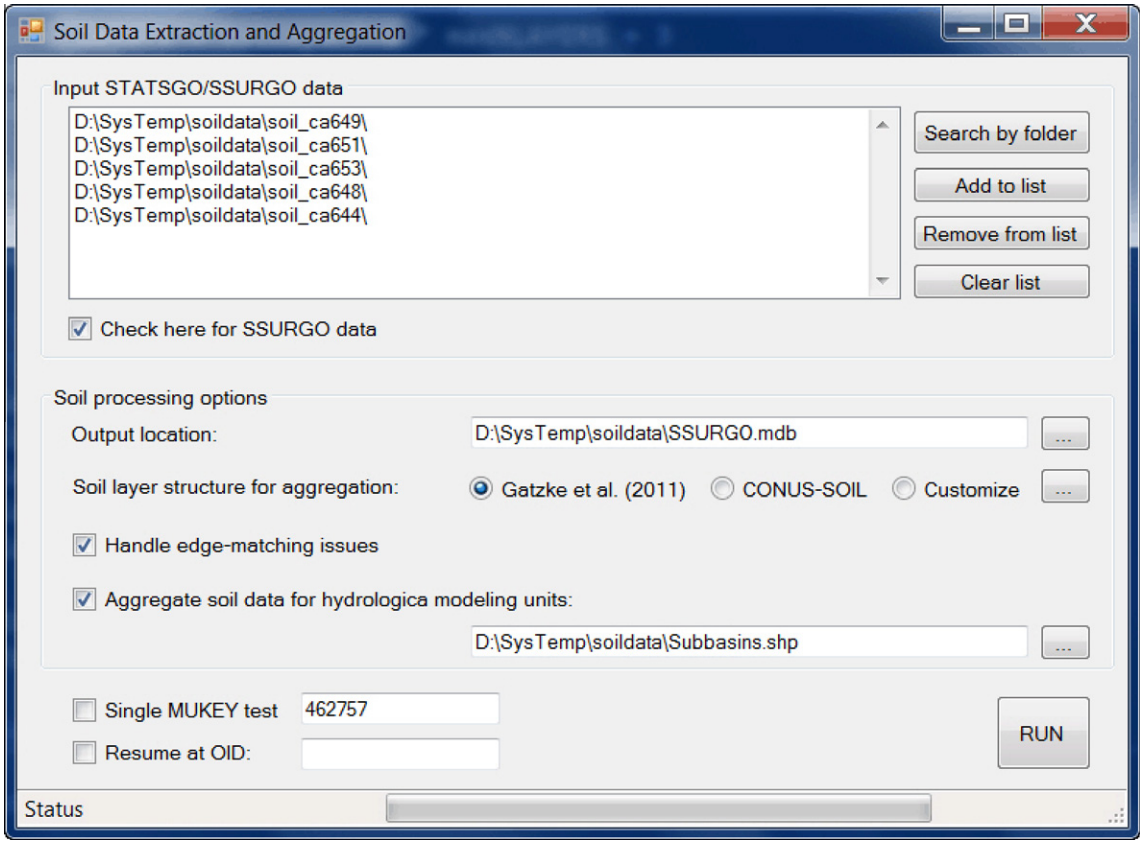

Fig. 4. Screenshot of the computer program implementing the soil aggregation approach developed in this study.

approximately15,000 km^2 and mainly enclosed by the SSAs of CA642, CA644, CA647, CA648, CA649, CA651, and the northern portion of CA653 (Fig. 5). Minor areas (7%) of the study domain are covered by other SSAs which are not included in this study. Detailed information on the site description is provided in our previous studies (Gatzke et al., 2011; Luo et al., 2008).

In total there are 1207 MUs (with 1422 components and 44 great-group taxonomies) in the watershed and 156 of them are associated with the edge-matching issues, indicating a fraction area of 27%. Out of the identified MUs with the issue, 140 MUs (90%) were handled by the developed data processing and merged into 39 new MUs. During the processing, soil data in 121 components were aggregated into 81 new components with common classifications of taxonomy. Finally, the processed soil map and data have 1106 MUs and 1382 components. In summary, the developed method handled majority of the edge-matching issues while generally kept the spatial resolution and variability in the original soil data.

6. Discussion and conclusion

Procedures of soil data processing to prepare input data for hydrologic models were developed and implemented in this study. The aggregation method generates a single soil profile for each of the modeling units of the simulation domain based on soil data in all components of the respective geographic region. Compared to the simple approach using only the dominant component, the method with area-weighted averages from all involved components preserved the intrinsic spatial resolution and variability in the source data and provided more representative soil parameters for the hydrologic models. In this study, the cumulative soil depth in the resultant soil profile was determined by matching the total saturated water content in the soil profiles before and after the soil data aggregation. This was considered as an improvement to the conventional depth-slicing algorithm in soil data processing, in which the cumulative soil depth for aggregation was set as the maximum over all involved components and the total water storage may be artificially increased during the aggregation.

This study also investigated the nature and the solution of edge-matching issues in SSURGO soil data across multiple soil survey areas. For hydrologic simulations, the artificial splits of MUs and associated components by SSA boundaries may mislead the determinations of representative soil types in the simulation domain. The proposed solution for the issue is to identify and restore the split components based on their common great group taxonomy. Two case studies were conducted to demonstrate the edge-matching issue and the capability of the developed methods in handling the issue. In a small area enclosed by SSAs of CA642 and CA647 (4000 km^2) 20% of the included MUs are associated with the edge-matching issues, while the ratio is 27% for the watershed of San Joaquin Valley (15,000 km^2). The developed approach by taxonomy-based aggregation solved the edge-matching issues for about 90% of the identified MUs in both studies. The reductions of MUs and components numbers before and after the aggregation are less than 10%, indicating that the soil-data preparation procedure has the capability to effectively handle the edge-matching issues and maintain the spatial resolution in the SSURGO soil database. The study also indicated that further investigation on the classification changes of taxonomy over time will significantly improve the performance of this approach in handling edge-matching issues in SSURGO.

The effects of the edge-matching issue on hydrologic simulations are associated to the spatial scale of modeling units, rather

Table 3
List of the grouped MUs which are adjacent and across the boundary of CA642 and CA647.

Map-unit key	Area (ha)	Component key	Great-group taxonomy	Coverage in MU (%)	Group ID
466932	860	466932:646092	Endoaquolls	45	1
466932	860	466932:646093	Endoaquolls	40	1
462825	1630	462825:636520	Haplaquolls	40	1
462825	1630	462825:636519	Haplaquolls	40	1
466941	945	466941:646126	Haploxeralfs	85	2
462937	770	462937:637160	Haploxeralfs	85	2
466942	2165	466942:646132	Xerofluvents	60	3
466942	2165	466942:646133	Xerorthents	30	3
462936	1980	462936:637154	Xerofluvents	85	3
466956	1795	466956:646205	Argixerolls	90	4
462813	4260	462813:636455	Argixerolls	85	4
466961	330	466961:646232	Haploxerolls	85	5
462819	7900	462819:636487	Haploxerolls	85	5
466962	365	466962:646237	Haploxerolls	85	6
462820	7500	462820:636494	Haploxerolls	85	6
466963	3695	466963:646242	Natrixeralfs	90	7
462888	11540	462888:636892	Natrixeralfs	85	7
466965	100	466965:646250	Haploxerolls	45	8
466965	100	466965:646251	Haploxerolls	25	8
462929	18990	462929:637114	Haploxerolls	85	8
466966	480	466966:646255	Haploxerolls	85	9
462800	1730	462800:636384	Haploxerolls	45	9
462800	1730	462800:636385	Haploxerolls	25	9
466969	815	466969:646270	Haploxererts	85	10
462778	1830	462778:636261	Chromoxererts	85	10
466970	1000	466970:646276	Haploxererts	50	11
466970	1000	466970:646277	Haploxerolls	35	11
462782	1670	462782:636287	Chromoxererts	50	11
462782	1670	462782:636288	Haploxerolls	35	11
466978	4940	466978:646324	Xerorthents	45	12
462923	9450	462923:637078	Xerorthents	35	12
462923	9450	462923:637080	Xerorthents	20	12
466979	22780	466979:646329	Xerorthents	45	13
462897	3510	462897:636941	Xerorthents	30	13
466980	40	466980:646334	Haploxerolls	75	14
462875	16190	462875:636821	Haploxerolls	85	14
466981	550	466981:646341	Haploxerolls	75	15
462877	1350	462877:636832	Haploxerolls	85	15
466989	260	466989:646386	Palexeralfs	45	16
466989	260	466989:646387	Haploxeralfs	40	16
462849	970	462849:636660	Palexeralfs	45	16
462849	970	462849:636661	Haploxeralfs	40	16
466990	2340	466990:646391	Argixerolls	80	17
462839	2850	462839:636599	Argixerolls	85	17
466991	10440	466991:646396	Argixerolls	40	18
466991	10440	466991:646397	Haploxeralfs	25	18
466991	10440	466991:646398	Palexeralfs	20	18
462840	10760	462840:636605	Argixerolls	40	18
462840	10760	462840:636606	Haploxeralfs	25	18
462840	10760	462840:636607	Palexeralfs	20	18
466992	3240	466992:646402	Haploxerepts	45	19
466992	3240	466992:646403	Palexeralfs	20	19
462869	4740	462869:636773	Xerochrepts	45	19
462869	4740	462869:636774	Palexeralfs	20	19
466994	4050	466994:646414	Haploxerepts	85	20
462867	5210	462867:636758	Xerochrepts	85	20
466995	1240	466995:646419	Haploxeralfs	35	21
466995	1240	466995:646420	Haploxerepts	30	21
462894	14290	462894:636917	Haploxeralfs	35	21
462894	14290	462894:636918	Xerochrepts	30	21
466996	4050	466996:646426	Haploxeralfs	50	22
462895	5800	462895:636925	Haploxeralfs	50	22
462904	1470	462904:636977	Chromoxererts	50	22
462904	1470	462904:636978	Haploxeralfs	35	22

than the scale of the simulation domain. If a simulation (such as grid-based models) is based on modeling units with areas smaller than typical MUs, the edge-matching will not be a problem by considering the small fraction of modeling units across SSAs to the total units. Most of the hydrologic models at watershed or catchment scales, however, have spatial resolutions significantly larger than the typical MU size (e.g. median size of about 1 km^2 in California). Examples are provided in this study for the effects

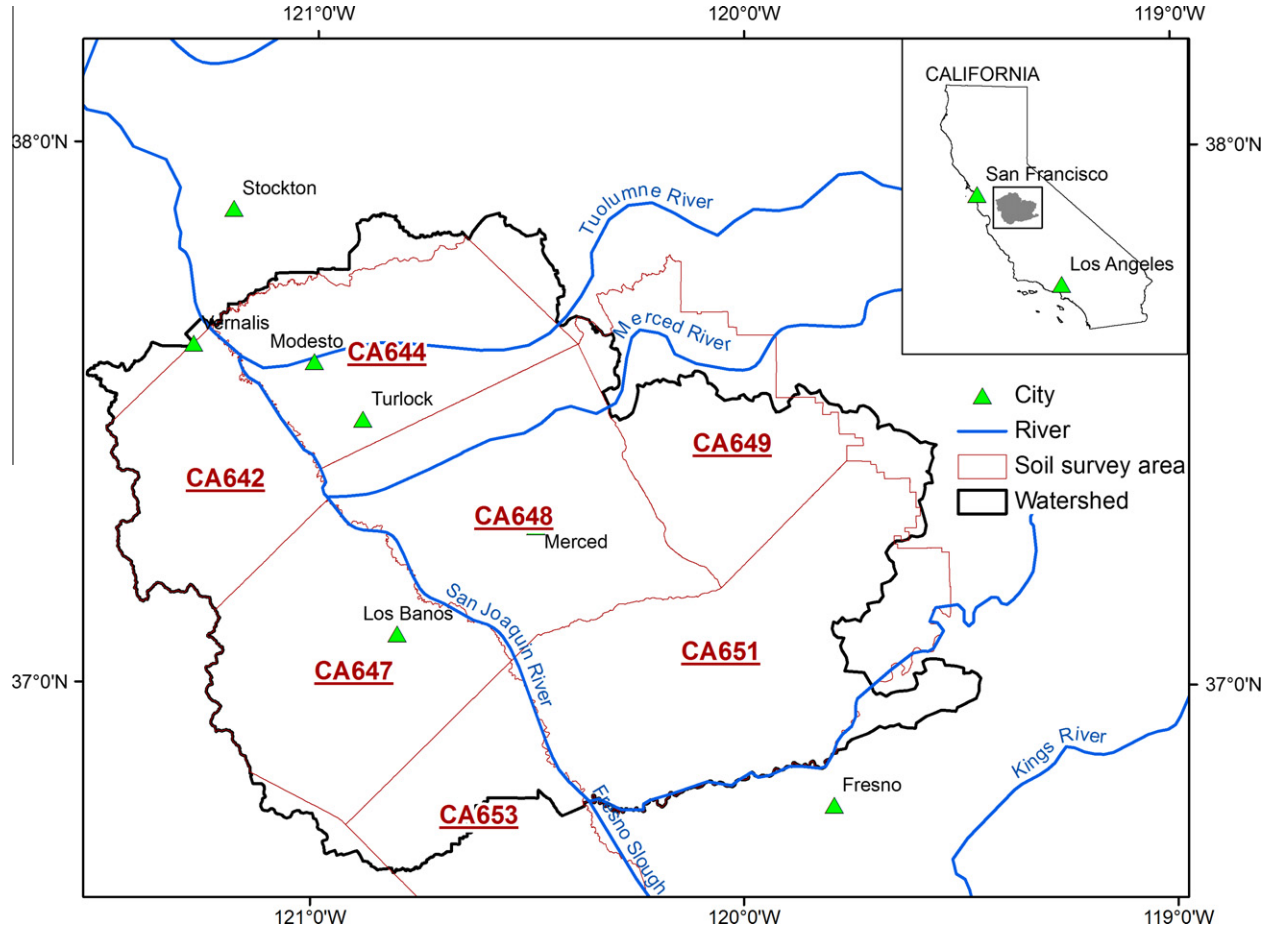

Fig. 5. The San Joaquin Valley watershed and enclosed soil survey areas of SSURGO.

of the edge-matching issue on the determination of representative soil types and the distribution of HRUs. Future studies are needed to determine the effects of the edge-matching issues and the proposed solution on hydrologic simulations at various spatial scales.

Acknowledgement

Authors would like to thank the University of California Kearney Foundation for the financial support for this study (Kearney Foundation 2008.002).

References

Anderson, R.M., Koren, V.I., Reed, S.M., 2006. Using SSURGO data to improve Sacramento Model a priori parameter estimates. J. Hydrol. 320 (1–2), 103–116.

Beaudette, D.E., O'Geen, A.T., 2011. Algorithms for Quantitative Pedology, a Toolkit for Soil Scientists <http://aqp.r-forge.r-project.org/>. Department of Land, Air, and Water Resources, University of California. Davis, CA.

Buell, G.R., Markewich, H.W., 2004. Data Compilation, Synthesis, and Calculations Used for Organic-Carbon Storage and Inventory Estimates for Mineral Soils of the Mississippi River Basin. USGS Professional Paper 1686-A. U.S. Geological Survey. Madison, WI.

Burns, I.S., Scott, S., Levick, L., Hernandez, M., Goodrich, D.C., Semmens, D.J., Kepner, W.G., Miller, S.N., 2004. Automated Geospatial Watershed Assessment (AGWA) – A GIS-Based Hydrologic Modeling Tool: Documentation and User Manual (http://www.epa.gov/esd/land-sci/agwa/pdf/user_manual.pdf). U.S. Department of Agriculture, Agriculture Research Service. Tucson, AZ.

Culman, S.W., Young-Mathews, A., Hollander, A.D., Ferris, H., Sánchez-Moreno, S., Geen, A.T.O., Jackson, L.E., 2010. Biodiversity is associated with indicators of soil ecosystem functions over a landscape gradient of agricultural intensification. Landscape Ecol. 25 (9), 1333–1348.

Davidson, E., Lefebvre, P., 1993. Estimating regional carbon stocks and spatially covarying edaphic factors using soil maps at three scales. Biogeochemistry 22 (2), 107–131.

FGDC, 1997. Soil Geographic Data Standard, FGDC-STD-006 <http://www.fgdc.gov/standards/projects/FGDC-standards-projects/soils/>. Soil Data Subcommittee, Federal Geographic Data Committee. Reston, VA.

Gatzke, S.E., Beaudette, D.E., Ficklin, D.L., Luo, Y., O'Geen, A.T., Zhang, M., 2011. Aggregation strategies for SSURGO data: effects on SWAT soil inputs and hydrologic outputs. Soil Sci. Soc. Am. J. 75 (5), 1908–1921.

Geza, M., McCray, J.E., 2007. Effects of soil data resolution on SWAT model stream flow and water. J. Environ. Manage. 88 (3), 394–406.

Geza, M., Poeter, E.P., McCray, J.E., 2009. Quantifying predictive uncertainty for a mountain-watershed model. J. Hydrol. 376 (1–2), 170–181.

Johnson, M.S., Coon, W.F., Mehta, V.K., Steenhuis, T.S., Brooks, E.S., Boll, J., 2003. Application of two hydrologic models with different runoff mechanisms to a hillslope dominated watershed in the northeastern US: a comparison of HSPF and SMR. J. Hydrol. 284 (1–4), 57–76.

Luo, Y., 2006. Geo-Referenced Multimedia Environmental Modeling of Chemical Fate and Transport, Ph.D. Dissertation <http://gradworks.umi.com/32/34/3234320.html>. Department of Civil and Environmental Engineering. University of Connecticut.

Luo, Y., Zhang, X., Liu, X., Ficklin, D., Zhang, M., 2008. Dynamic modeling of organophosphate pesticide load in surface water in the northern San Joaquin Valley watershed of California. Environ. Pollut. 156 (3), 1171–1181.

Miller, D.A., White, R.A., 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth Interact. 2 (2), 1–26.

Peschel, J.M., Haan, P.K., Lacey, R.E., 2003. A SSURGO Pre-Processing Extension for the ArcView Soil and Water Assessment Tool, ASAE paper number 0321223. 2003 ASAE Annual International Meeting. Las Vegas, NV.

Peschel, J.M., Haan, P.K., Lacey, R.E., 2006. Influences of soil dataset resolution on hydrologic modeling. J. Am. Water Res. Assoc. 42 (5), 1371–1389.

Polyakov, V., Fares, A., Kubo, D., Jacobi, J., Smith, C., 2007. Evaluation of a non-point source pollution model, AnnAGNPS, in a tropical watershed. Environ. Modell. Software 22 (11), 1617–1627.

Sahoo, G.B., Ray, C., De Carlo, E.H., 2006. Calibration and validation of a physically distributed hydrological model, MIKE SHE, to predict streamflow at high frequency in a flashy mountainous Hawaii stream. J. Hydrol. 327 (1–2), 94–109.

Santhi, C., Srinivasan, R., Arnold, J.G., Williams, J.R., 2006. A modeling approach to evaluate the impacts of water quality management plans implemented in a watershed in Texas. Environ. Modell. Software 21 (8), 1141–1157.

Sheshukov, A., Daggupati, P., Lee, M.-C., Douglas-Mankin, K., 2009. ArcMap tool for pre-processing SSURGO soil database for ArcSWAT. In: Proceedings of the 5th International SWAT Conference, Boulder, CO.

Singh, H.V., Kalin, L., Srivastava, P., 2011. Effect of soil data resolution on identification of critical source areas of sediment. J. Hydraulic Eng. 16 (3), 253–262.

Smemoe, C.M., Nelson, E.J., Zhao, B., 2004. Spatial averaging of land use and soil properties to develop the physically-based green and ampt parameters for HEC-1. Environ. Modell. Software 19 (6), 525–535.

Smith, S.V., Sleezer, R.O., Renwick, W.H., Buddemeier, E.W., 2005. Fates of eroded soil organic carbon: mississippi basin case study. Ecol. Appl. 15 (6), 1929–1940.

USDA, 1999. Soil Taxonomy, A Basic System of Soil Classification for Making and Interpreting Soil Surveys. U.S. Department of Agriculture, Natural Resources Conservation Service. Washington, DC.

USDA, 2006. Keys to Soil Taxonomy, 10th Ed. U.S. Department of Agriculture, Natural Resources Conservation Service. Washington, DC.

USDA, 2009. Official Soil Series Descriptions (OSD) for GUS series <https://soilseries.sc.egov.usda.gov/OSD_Docs/G/GUS.html>. U.S. Department of Agriculture, Natural Resources Conservation Service. Washington, DC.

USDA, 2011. Soil Survey Geographic (SSURGO) Database, <http://soildatamart.nrcs.usda.gov>. United States Department of Agriculture, Natural Resources Conservation Service. Washington, DC.

USGS, 1995. Soils data for the Conterminous United States Derived from the NRCS State Soil Geographic (STATSGO) Data Base <http://water.usgs.gov/GIS/metadata/usgswrd/XML/ussoils.xml>. U.S. Geological Survey. Reston, VA.

Winchell, M., Srinivasan, R., Di Luzio, M., Arnold, J., 2011. ArcSWAT interface for SWAT2009, user's guide <http://swatmodel.tamu.edu/software/arcswat/>. Blackland Research Center, Texas Agricultural Experiment Station, Temple, TX; Grassland, Soil and Water Research Laboratory, USDA Agriculutral Research Service, Temple, TX.

Wu, K., Johnston, C.A., 2007. Hydrologic response to climatic variability in a great lakes watershed: a case study with the SWAT model. J. Hydrol. 337 (1–2), 187–199.

Zhang, Y., Zhang, Z., Reed, S., Koren, V., 2011. An enhanced and automated approach for deriving a priori SAC-SMA parameters from the soil survey geographic database. Comput. Geosci. 37 (2), 219–231.