An Enhanced and Automated Approach for Deriving *a priori* SAC-SMA Parameters from the Soil Survey Geographic Database

Yu Zhang*

Ziya Zhang

Seann Reed

Victor Koren

Office of Hydrologic Development,

National Weather Service

Silver Spring, MD, 20910

July 26, 2007

*yu.zhang@noaa.gov

Abstract

This paper presents an automated approach for deriving gridded *a priori* parameters for the National Weather Service (NWS) Sacramento Soil Moisture Accounting (SAC-SMA) model from the Soil Survey Geographic (SSURGO) Database and National Land Cover Database (NLCD). Our approach considerably extends methods previously used in the NWS and offers automated and geographically invariant ways of extracting soil information, interpreting soil texture, and aggregating SAC-SMA parameters. The methodology is comprised of four components, all of which are implemented in open-source software, notably R (a statistical package) and the Geographic Resources Analysis Support System (GRASS; a free Geographic Information System). The first and second components are SSURGO and land cover preprocessors, which are written in R and GRASS, respectively. The third component is the parameter generator based on both R and GRASS; it produces 11 SAC-SMA parameters for each soil survey area on an approximately 30-m resolution grid. The last component is a C++ based postprocessor that creates parameters on the Hydrologic Rainfall Analysis Project (HRAP) grid and for the area of interest. We describe the scientific basis and technical features of the four components, and demonstrate their efficacy through the creation of mosaicked parameter grids covering the geographic domains of six NWS River Forecast Centers (NWSRFCs).

1. Introduction

For the last several years the Office of Hydrologic Development (OHD) of the National Weather Service (NWS) has been investigating the use of distributed hydrologic models to provide enhanced river forecasts and other products and services to the Nation (Carter, 2002). In February 2007, the NWS delivered the first operational version of a distributed model. In comparison to the existing NWS lumped (spatially averaged) operational model, the distributed model offers enhanced capabilities to capture spatial variations in meteorological forcing and land surface hydrologic properties. In the Distributed Model Intercomparison Project (DMIP; Smith et al. (2004)), the NWS distributed model performed favorably against the NWS lumped model and other distributed models in several catchments (Reed et al., 2004). Interested readers are referred to Koren et al. (2003, 2004) for a complete description of the NWS distributed model and to Reed et al. (2007) for flash flood applications of this model. Hereafter, we refer to the NWS distributed model as the *NWS-DHM* or simply the *DHM*.

At present, the water balance component of DHM is the Sacramento Soil Moisture Accounting model (SAC-SMA; Burnash et al. (1973) and Burnash (1995)). A prerequisite to the application of the gridded SAC-SMA is a set of spatially distributed parameters whose values are reflective of the spatial distribution of corresponding physical characteristics. Koren et al. (2000) laid out an *a priori* estimation framework in which the values of 11 SAC-SMA parameters are related to observable soil properties (*a priori* refers to the fact that no model calibration is involved). This framework has been employed in deriving *a priori* parameters from a gridded State Soil Geographic Database (STATSGO; (Miller and White, 1998)), and the results have been used in the DMIP. More recently, Anderson et al. (2006) developed a set of methods utilizing this framework that allow one to derive the parameters from Soil Survey Geographic database (SSURGO) and National Land Cover Dataset (NLCD), two public domain data sources that offer by far the most detailed characterizations of soil and land cover on a national scale.

The approach described by Anderson et al. (2006) has been applied to a number of catchments and it has been shown to, in general, enhance the accuracy of predictions of SAC-SMA in the absence of calibration (see Zhang et al. (2006)). Yet, at least three major limitations exist for this approach which to a large extent hinder its wider application. Firstly, some of the underlying methods require *ad hoc* adjustments to account for geographic variations in the configuration of SSURGO data. Secondly, the approach is implemented in a way that extensive manual processing via graphical user interface (GUI) is needed. Thirdly, the application of this approach is predicated on the access to proprietary software (e.g., Microsoft Access and ESRI Arcview). These limitations, combined with the sheer volume and complexity of SSURGO and NLCD data sets, prompted the demand for an automated and geographically consistent approach that would allow modelers to derive the *a priori* parameters in a reproducible and timely manner. In response to this demand, we developed a highly automated approach that retains the basic procedures used by Anderson et al. (2006) but is based on systematic underlying methodologies that are applicable to most, if not all, geographic settings. Moreover, this approach is implemented exclusively via open source software packages on a Linux platform, and therefore the developed algorithms can be applied and tested without the restriction of software licensing requirements. While the approach was intended for the parameter derivation of SAC-SMA model, its methodologies and tools are applicable to similar tasks for other hydrologic models whose parameters can be based on SSURGO, NLCD, or a combination of both (such as the Soil and Water Assessment Tool, or SWAT; see Wang and Melesse (2006)). In this paper we present this approach and demonstrate its efficacy through the derivation of a set of mosaicked parameters for 25 states that encompass the geographic domains of six NWS River Forecast Centers (RFCs) which were designated as the initial experimental sites for evaluating the distributed model. In addition to facilitating the deployment of DHM for operational river and flash flood forecasting in these RFCs, our approach is also expected to help determine parameter values for the lumped model presently being used in these RFCs.

The remainder of the paper is organized as follows: Section 2 reviews the *a priori* parameter estimation framework. Section 3 describes the needed data sets and the data processing procedures. Section 4 examines the resulting parameter grids and Section 5 summarizes the work.

2. The *a priori* Parameter Estimation Framework

The formulation of the SAC-SMA model was originally presented in Burnash et al. (1973). In the model, soil horizons are generalized into a relatively thin *upper zone* (UZ) and a *lower zone* (LZ), with water stored in each zone further partitioned into *free water* that drains by gravity and *tension water* held by the suction of soil matrix. The free water storage of the lower zone is subdivided into *supplemental* and *primary storages*, which account for faster and slower draining groundwater flows, respectively. *Percolation* is allowed from the upper to the lower zone. During a rainfall event, the runoff rate is determined jointly by rainfall, UZ storages and the percolation rates (see Koren et al. (2003, 2004), Anderson et al. (2006) and the references therein).

The framework of Koren et al. (2000) offers a means for estimating 11 soil-related SAC-SMA parameters (Table 1) from soil and land cover data. The soil data set provides three sources of information 1) hydrologic soil group (HSG), 2) texture class and 3) vertical soil profile. In Anderson et al. (2006)'s approach, HSG is used jointly with land cover to determine the Soil Conservation Service (SCS, now Natural Resource Conservation Service, or NRCS) curve number (CN). Table 2 shows the CN values assigned to each combination of HSG and land cover class following such an approach (see methodology in Appendix A). Soil texture is used to estimate hydrologic properties that would form the basis for estimating SAC-SMA parameters. These properties include porosity (θ_s) field capacity (θ_{fld}), wilting points (θ_{wp}) and saturated hydraulic conductivity (K_s). The vertical soil profile is used in delineating the upper and lower zones.

The first step of parameter estimation entails estimating the thickness of the upper zone Z_{up}. The method of estimation developed by Koren et al. (2000) is based on the NRCS approach documented in McCuen (1982). Essentially, the curve number determines the initial rain abstraction I_a for a soil column via the formula $I_a = 5.08(\frac{1000}{CN} - 10)$ (Chap. 10 of NRCS-ARS (2004)). I_a is assumed to be 20% of additional water needed to saturate a soil column initially at its field capacity. Therefore, Z_{up} is uniquely determined by I_a, soil porosity θ_s and field capacity θ_{fld}. For a multi-horizon soil column, Z_{up} needs to satisfy the following condition:

$$I_a = \int_0^{Z_{up}} \theta_s(z) - \theta_{fld}(z)dz \qquad (1)$$

where z is the depth from surface. The schematic of the estimation is shown in Figure 1. In practice, one adds up the free water storage of each horizon iteratively until the sum equals or exceeds the initial abstraction I_a (Fig. 1). In the latter case, the horizon where I_a is exceeded is split in such a way that allows the free water storage of the upper zone to equal I_a. The lower zone extends from this depth to the depth of bedrock, or to the upper edge of an impermeable layer when such a layer exists above the bedrock (Fig. 1). The 11 SAC-SMA parameters can be computed given the soil properties for the upper and lower zones (see Appendix A, and Koren et al. (2000) for additional details). Among these, three parameters, namely, UZTWM, UZFWM, LZTWM can be computed directly by adding corresponding quantities of each soil horizon, whereas estimating the rest requires vertically averaged soil properties (Appendix B).

3. Data, Methodology and Implementation

3a. *Soil Survey Geographic (SSURGO) Database*

The SSURGO project was undertaken by the NRCS to provide digitized soil maps for the entire United States at resolutions significantly higher than those for STATSGO (mapping scale from 1:12,000 to 1:63,360 for SSURGO compared to approximately 1:250,000 for STATSGO; http://www.ncgc.nrcs.usda.gov/products/datasets/ssurgo/description.html). The project is set to complete in 2008, and at present SSURGO maps are available for the majority of counties for most of the states (its status can be found in http://soildatamart.nrcs.usda.gov/StatusMaps/SoilDataAvailabilityMap.pdf). SSURGO data can be downloaded or purchased on a state basis from the NRCS data gateway (http://datagateway.nrcs.usda.gov/).

Figure. 3 depicts the hierarchical structure of the SSURGO database. The data sets are organized by *survey area*s, where a survey area is usually equivalent to a county. Within each survey area, soil patches which share similar characteristics are lumped into one *map unit* (Fig. 2). Within each map unit are multiple *soil components* with varying percentage areal coverage (Fig. 3). Each soil component encompasses one or multiple *horizons*, and at each horizon one or multiple soil texture classes may be present (Fig. 3). At a given horizon, the soil element corresponding to a unique texture class is hereafter referred to as a *subcomponent*.

Each horizon is associated with a unique vertical extent. Our approach adopts the strategy devised by Anderson et al. (2006) that simplifies soil structure and filters out the variables to be used in parameter derivation. In this strategy, for a map unit, only the primary soil component, i.e., the one with the highest percentage areal coverage, is selected while the rest are ignored. For the selected component, at a given horizon all the embedded subcomponents with identifiable texture are used to compute horizon-averaged soil properties (The method of identifying texture is presented in Section 3c). The simplification strategy further extends to determining the value of variables. For certain variables such as depth, three estimates are provided by SSURGO, namely, the lower, the representative and the upper estimates. Only the representative values are used here. Another element of the strategy is that the depth to a restrictive layer, when given in SSURGO, would be used as the depth to the bottom of the SAC-SMA lower zone (designated by Z_{max}) in lieu of the depth to bedrock (also provided in SSURGO), and the soil below Z_{max} is ignored in estimating the parameters.

The SSURGO data set for a survey area consists of spatial and tabular files. Available in three formats, namely, ESRI Arcview shapefile, ESRI Arc Coverage and ESRI Arc Interchange, the spatial files encapsulate the location of each map unit (represented as polygons). In our approach the Arcview shapefile format is used. The tabular data are provided in multiple pipe-delimited text files that contain information on soil texture and associated properties at various depths. These sources of information can be mapped to the spatial polygons via a key "MUSYM". The NRCS provides an Microsoft Access interface for performing such mappings and for extracting soil information. This interface was an integral component of the approach of Anderson et al. (2006). In order to fully automate the SSURGO processing, the present approach bypasses this interface and instead relies on a set of scripts for information extraction. Due to the fact that the column titles are absent from the tabular files, the scripts parse metadata provided by NRCS to determine the column titles (the metadata can be found at http://soildatamart.nrcs.usda.gov/documents/SSURGOMetadataTableColumns.pdf). This yields a text file that lists all the column titles (columnHeading.txt) and this file is provided as a part of the electronic supplement (see Appendix C).

3b. *National Land Cover Database (NLCD)*

The NLCD data set was created from Land Remote-Sensing Satellite (Landsat) images by the Multi-Resolution Land Characteristics (MRLC) Consortium. Two versions are currently available: the NLCD 1992 and NLCD 2001. Methodologies for creating the data sets can be found in Vogelmann et al. (2001) and Homer et al. (2004), respectively. In both versions the land cover is represented on approximately 30-m grids in the coordinates of Albers Equal Area (AEA). The NLCD 1992 files are partitioned along state boundaries (California and Texas are sub-divided into multiple sub-regions). The NLCD 2001, by contrast, is organized by zones. There are 14 super-zones for the conterminous United States (CONUS). The data for each zone can be downloaded at the MRLC website (http://www.mrlc.gov/).

NLCD employs a modified version of the Anderson land-use and land-cover classification system (Anderson et al., 1972) with nine broad categories (see details at http://landcover.usgs.gov/classes.php#similar). The sub-categories differ slightly for NLCD 1992 and 2001, with the latter containing refined categories under Shrubland, Herbaceous Upland Natural/Semi-Natural Vegetation, and Wetlands (Table 2). To account for these differences, methods have been developed separately for using the two versions of NLCD, but only the latter is described here, for it provides more recent land cover data which is deemed of closer relevance to the forecasting missions of NWS.

3c. *Processing Strategy and Software Implementation*

Given the large volume of SSURGO and NLCD data, it is difficult to derive the parameters on a national grid in one pass. Instead, the derivation is done incrementally. In the beginning, SAC-SMA parameters are obtained for each SSURGO soil survey area on approximately 30-m grids. These grids for a given parameter are then merged to yield a state-wide data set on the grid of Hydrologic Rainfall Analysis Project (HRAP; Reed and Maidmant (1999); actual resolutions include 1km (1/4 HRAP), 2km (1/2 HRAP), and 4km (full HRAP)), with each HRAP grid cell being a basic spatial unit of radar rainfall input and DHM (Reed, 2003). The HRAP-based parameters from multiple states are subsequently mosaicked to produce the final data set.

From an implementation standpoint, the parameter derivation procedures are combined into three phases. Figure 4 provides a schematic of the derivation process. The first phase entails preprocessing SSURGO and NLCD to correct the errors in the raw data, to extract information relevant to parameter derivation, and to save the results in formats usable in later phases. This phase is implemented in two software components, i.e, the SSURGO and NLCD *preprocessors* (Fig. 4). The second phase entails generating parameters for each soil survey area, and the correspondent software component is referred to as *parameter generator* (Fig. 4). The third and final phase concerns postprocessing the 30-m parameter grids to yield mosaicked, HRAP-based gridded data sets, and the software implementation is termed *postprocessor* (Fig. 4).

The first three components, i.e., the two preprocessors and the parameter generator were implemented primarily in open source packages *R* (http://www.r-project.org) and *Geographic Resources Analysis Support System* (GRASS; http://grass.itc.it/), whereas the various codes that constitute the postprocessor are written in C++. Supplemental scripts were developed in Unix shell and Perl (http://www.perl.com) to automate the processing. R is a multi-platform statistical package with its origin traced back to the S language (see Becker et al. (1988)). GRASS is a Geographic Information System (GIS) package originally conceived by the U.S. Army Construction Engineering Research Laboratories (USA-CERL). Running primarily on Unix and Linux environment, GRASS allows users to add extensions via its C/C++ interface and permits batch processing via its shell interface. In our effort, R 2.4 and GRASS 6.1 were used. In comparison to the methods of Anderson et al. (2006), the current approach offers more systematic and automated ways of extracting information from SSURGO and NLCD data (SSURGO and NLCD preprocessors), for interpreting soil texture (SSURGO preprocessor), for computing parameters (parameter generator), and for aggregating parameters (postprocessor). All the scripts and programs are provided in the electronic supplement (see Appendix C). The notable differences between the previous and the present approaches are highlighted in Table 3, and the structural details of each of the four components in the present approach are provided below.

SSURGO Preprocessor

The data flow diagram for the SSURGO preprocessor is shown in Figure 5. The preprocessor takes the raw SSURGO tabular data and the raw attribute table from the shapefile as input, extracts and organizes the needed information, and generates a table that contains horizon-averaged soil properties for each map unit (mu_table.dbf), and an augmented attribute table with HSG information (Fig. 5). The entire preprocessing can be done using a single shell script (preprocessor.sh) that first creates directories and then calls six R scripts (see Table 4 for functionality). The six R scripts can be further broken into four groups, and the details follow below.

The first group consists of one R script (std.tname.R; Table 4 and Fig. 5) that is responsible for standardizing the file names. This is needed since the actual names of tabular files were found to sometimes differ from the standard ones given in the metadata. As an example, according to the metadata the tabular file that provides information on the restrictive layer is named "corestrictions.txt". However, for some survey areas the actual file name is "crstrcts.txt" instead. The script eliminates such discrepancies by creating a symbolic link with a corrected file name taken from the metadata and pointing to the actual tabular file.

The second group is comprised of three R scripts, namely, hydrologic.R, physical.R and zmax.R (Table 4 and Fig. 5). These scripts extract the information relevant to parameter derivation, namely the hydrologic soil groups, soil horizons and texture, and the maximum depth (i.e., Z_{max}), respectively (Table 4 and Fig. 5). Each script first identifies the dominant texture components in terms of percentage areal coverage. When there exists a unique dominant soil component with valid texture specification, the script would proceed to eliminate the rest of the components. In cases where there is a tie in areal coverage among multiple components, the script would select one. When either texture or horizon information is missing for the dominant component, the script would select the next component where such information is available. Each script generates a dBase file for storing the respective information (Fig. 5).

The third group is comprised of two scripts. The first one, aug.soil.attr.R, augments the attribute table of the soil data by adding the hydraulic soil groups. The second script, phy_lay_ave.R, takes two tables generated in the preceding component, i.e., physical.dbf and zmax.dbf, and computes averaged soil properties for each soil horizon (Fig. 5). This is accomplished in three steps. In the initial step, the script uses Z_{max} given in zmax.dbf to determine the lower boundary for the SAC-SMA lower soil zone. In the second step, the script follows Anderson et al. (2006) in ignoring the soil layers below Z_{max}. Subsequently, phy_lay_ave.R maps the soil texture of each subcomponent onto 12 simplified classes. phy_lay_ave.R then looks up the soil properties corresponding to each texture class from experimental measurements (see Table 5 for values; see Anderson et al. (2006) and sources cited therein). Previously, Anderson et al. (2006) relied on a manually prepared mapping table for determining simplified texture. The mapping table requires frequent updates due to the wide variations in the localized names in the "TEXTURE" field. This method is now replaced by an automated mapping algorithm that is applicable to any setting. Figure 6 shows the schematic of this algorithm. In a nutshell, the algorithm first examines the SSURGO field "TEXTURE". If this field contains a string that points unambiguously to a known texture class, then the corresponding simplified texture class is assigned accordingly (Scenario A in Fig. 6). However, in some situations the TEXTURE field contains only a generic descriptor without providing concrete information on the actual texture, and meanwhile the description field "MUNAME" contains an identifiable texture name. In these situations the latter would be used in lieu of the former to determine the simplified texture (Scenario B in Fig. 6). If neither field provides the needed information, a symbol of "O", which represents "Other", is assigned (Scenario C in Fig. 6). Finally, for each property, phy_lay_ave.R derives a unique value for each horizon by averaging the corresponding values for all embedded subcomponents whose texture class is valid (i.e., other than "O").

NLCD Preprocessor

Preprocessing of NLCD is done in three steps, each involving a GRASS/SHELL script (Table 6 and Fig. 7). It was found that the current NLCD data contains an erroneous value 127 (valid range of land cover is 1-99). In the first step the raw NLCD 2001 data sets are imported into GRASS via the script named import_2001.sh, wherein any cell with a value 127 is set to null (Table 6 and Fig. 7). Then the script named zone_to_state.sh is used to mosaic the zonal NLCD 2001 data sets and then divide the results along state boundaries to expedite the parameter derivation (Table 6 and Fig. 7). The NLCD data for each state is subsequently reprojected into geographic coordinates to match the SSURGO data via the script reproj.sh (Table 6 and Fig. 7) .

Parameter Generator

The parameter generator consists of a collection of GRASS functions and scripts that are wrapped in a shell script named param_gen.2001.sh (Table 7). Figure 8 provides a dissection of the generator. The parameter generator takes three sources of input, i.e., a) the NLCD 2001 in GRASS grid format (from NLCD preprocessor; Fig. 7), b) the augmented attribute table associated with the SSURGO shapefile (from SSURGO preprocessor; Fig. 5), and c) the horizon-averaged soil properties in mu_prop.dbf (from SSURGO preprocessor; Fig. 5). The parameter generator first computes a curve number grid on the basis of NLCD and soil hydraulic group data provided by b). To reduce computation, it then follows Anderson et al. (2006) in determining the parameters for each polygon, and in the end converts the polygon-based parameters onto 30-m GRASS grids (Fig. 8). The details of the generator follow below.

The parameter generation begins with importing the shapefile attribute table into GRASS via a script import_ssurgo.sh (Table 7; Fig. 8). The resultant GRASS attribute table is subsequently rasterized to yield two 30-m grids, namely the grid of soil hydraulic groups and that of polygon identification numbers (Fig. 8). The former is then coupled with the NLCD grid to compute curve numbers on the 30-m grid via an external GRASS function named r.cn.2001 (Table 7; Fig. 8). This curve number grid is then coupled with the polygon ID grid via another GRASS function r.cn.ave.poly to derive polygon-averaged curve number (Table 7; Fig. 8). Subsequently, an R script sac_sma.each.R is used to join the curve number with the attribute table and to yield 11 SAC-SMA parameters along with upper and lower zone soil properties for each polygon (Table 7; Fig. 8). Upon completion, the parameter generator invokes a GRASS rasterization routine to convert polygon-based parameter and soil properties to respective gridded products whose spatial coordinates are identical to the 30-m land cover grid (Fig. 8). The parameter generator creates a secondary product, i.e., a grid of HRAP coordinates, via an external GRASS/C function named r.ll.hrap (Table 7; Fig. 8). To elaborate, the function computes and stores the coordinates of the 1/4 HRAP pixel for each 30-m cell of the parameter grid on the basis of the latitude and longitude of the latter. This resultant grid is used in conjunction with the parameter grids for computing averaged parameter values for each 1/4 HRAP pixel in the postprocessor.

Postprocessor

The postprocessor combines 30-m parameter grids for all survey areas needed and produces SAC-SMA parameter values on 1/4, 1/2 and full HRAP grids. As illustrated in Figure 9, the postprocessing takes place in four steps. In the first step, a C++ program (grass2xmrg; Table 8 and Fig. 9) reads in parameter and the associated HRAP ID grids for all soil survey areas within a state in a sequential manner, and then computes the number of embedded 30-m cells and the averaged parameter values for each 1/4 HRAP pixel (Fig. 9). The reason for keeping track of the former is to provide a consistent way of computing pixel-average values across county and state boundaries. To elaborate, a pixel along survey boundaries likely overlaps with multiple survey areas. For such a pixel, survey-area level parameter estimation results in multiple pixel-mean values, each based solely on the values within one survey area and stored in an individual file. Combining the average value for this pixel needs to take into account the variation in the overlapping between the pixel and survey areas as well as the number of missing values in the sub-pixel cells. Hence the number of embedded 30-m cells with valid values (not missing) is used here as the weight in computing a weighted average of the parameters for each pixel.

In the second step, state-wise, HRAP-based parameter grids are mosaicked through a C++ program (mergeXMRG; Table 8 and Fig. 9). This program reads in the parameter values and associated number of embedded cells with valid values. It then uses the latter variable as the weight to compute a weighted average of parameters for each 1/4 HRAP pixel. In the last step, the 1/4 HRAP grids are aggregated via a C++ program (aggrgXMRG; Table 8 and Fig. 9) onto 1/2 and full HRAP grids. Once again the number of embedded cells are used as the weight to account for the variations in its value across pixels (the number of sub-pixel cells varies depending on the number of missing values).

4. Derivation of a Multi-State *a priori* Parameter Set

The automated approach was employed in deriving a multi-state parameter product for a region that covers the geographic domains of six RFCs, i.e., CNRFC (California-Nevada), CBRFC (Colorado Basin), WGRFC (West Gulf), ABRFC (Arkansas-Red Basin), LMRFC (Lower Mississippi) and SERFC (Southeast). This region encompasses 25 states and a total of 1713 survey areas.

For this effort, NLCD 2001 data for 14 super-zones were obtained and processed. The parameter derivation was performed on three Linux workstations at the OHD. Each workstation is equipped with two 32-bit Intel processors and approximately two gigabytes of memory. Approximately 140 gigabytes of disk space were allocated for storing the intermediate and final products. The absence of license-related restrictions permitted simultaneous parameter derivation for multiple states on multiple workstations. When utilizing all three workstations, the entire process of parameter derivation took about a week to complete.

Since the previous approach developed by Anderson et al. (2006) required manual processing and ran only on Windows and HP-UX rather on Linux, a precise comparison of the performance of the two approaches is not feasible. Nevertheless, past experience, along with the results from testing the previous approach against several counties shows that, excluding the time for downloading SSURGO and preprocessing NLCD data, on the average about six hours are required for obtaining the HRAP-based product for a single survey area (In this approach the SSURGO and NLCD were downloaded from sources that differ from the previous one, and NLCD 2001 data was processed for most parts of the country and the processing time is negligible when divided by the number of survey areas). By contrast, applying the present approach for deriving the HRAP-based product for 92 survey areas took only about 22 hours to complete using a single processor. This means that about 0.24 hour is needed per survey area. A break-down of the approximate time needed in each component is shown in Table 9. The resulting grids of UZTWM and LZFSM are shown in Figure 11. These parameter grids are undergoing evaluation against similar parameter sets derived using STATSGO data to determine whether and how the use of the former would lead to improvements to the accuracy in streamflow predictions.

5. Summary

This paper presents an enhanced and automated approach for deriving *a priori* parameters for the NWS distributed hydrologic model from Soil Survey Geographic database and National Land Cover Dataset. The approach is implemented entirely in open source software packages, notably R and GRASS. It consists of four elements: i) SSURGO preprocessor; ii) NLCD preprocessor; iii) parameter generator and iv) parameter post-processor. These elements offer systematic and reproducible ways of acquiring, processing and computing parameters. The approach was demonstrated in the derivation of a set of multi-state a priori parameter grids that cover the drainage of six River Forecast Centers, and was shown to significantly reduce the time of parameter estimation. The methodologies and the associated software, in particular those for deriving curve number, identifying soil texture, and computing horizon- and vertically averaged soil properties, can be adopted to facilitate the implementation of other hydrologic models that can utilize SSURGO and NLCD data (such as SWAT).

Acknowledgment

We are indebted to Thomas Adams at Ohio River Forecast Center (OHRFC) who contributed sample GRASS scripts for this project and provided many insightful comments to the manuscript. We would also like to acknowledge James Wickham at MRLC, Charles Larson at USGS, and Sam Brown at USDA for their generous support.

Appendix A

The curve number is assigned on the basis of hydrologic soil group (HSG) and land cover type in accordance to the empirical relations published in Chapter 9 of NRCS-ARS (2004) assuming dry antecedent condition. There are four primary HSGs and each corresponds to an estimated runoff potential, i.e., A (low), B (moderate), C (moderately high) and D (high) (see Chapter 7 of NRCS-ARS (2004)). Besides the four classes, there are also dual classifications (e.g., "A/D", "B/D" and "C/D"). It is assumed here that a) it is the variation in soil moisture level that results in the pairwise difference in drainage characteristics (say, between "**A**" and "**A**/D"), and b) such a difference is automatically accounted for by soil moisture states generated by a model such as DHM. Therefore, no additional distinction needs to be made in the curve number between two HSGs sharing the same primary designator, e.g., "**A**" and "**A**/D". This is reflected in the curve number assignment in Table 2.

Appendix B

Zonal-averages of soil property χ for upper and lower zones (denoted by χ^u and χ^l respectively) are defined as follows:

$$\chi^u = \int_0^{Z_{up}} \chi(z)dz \tag{B.1}$$

and

$$\chi^l = \int_{Z_{up}}^{Z_{max}} \chi(z)dz \tag{B.2}$$

where Z_{up} and Z_{max} are the depths to the bottom of the upper and lower zones, respectively. In these equations,χ can be porosity θ_s, field capacity θ_{fld}, wilting point θ_{wp}, and saturated hydraulic conductivity K_{sat}. The SAC parameters for the upper and lower zones can be defined correspondingly (the definitions of these parameters can be found in Koren et al. (2000)).

$$UZTWM = \int_0^{Z_{up}} \theta_{fld}(z) - \theta_{wp}(z)dz \tag{B.3}$$

$$UZTWM = \int_0^{Z_{up}} \theta_s(z) - \theta_{fld}(z)dz \tag{B.4}$$

$$UZK = 1 - \left(\frac{\theta^u_{wp}}{\theta^u_s}\right)^n \tag{B.5}$$

$$LZTWM = \int_{Z_{up}}^{Z_{max}} \theta_{fld}(z) - \theta_{wp}(z)dz \tag{B.6}$$

$$LZFSM = \left(\frac{\theta^l_{wp}}{\theta^l_s}\right)^n \int_{Z_{up}}^{Z_{max}} \theta_s(z) - \theta_{fld}(z)dz \tag{B.7}$$

$$LZFPM = \left[1 - \left(\frac{\theta^l_{wp}}{\theta^l_s}\right)^n\right] \int_{Z_{up}}^{Z_{max}} \theta_s(z) - \theta_{fld}(z) dz \tag{B.8}$$

$$LZSK = \frac{UZK}{1 + 2(1 - \theta^l_{wp})} \tag{B.9}$$

$$LZPK = 1 - e^{-\frac{1}{\mu}(1+\beta)\pi^2 K_s D_s^2 (Z_{max} - Z_{up})\delta t} \tag{B.10}$$

$$PFREE = \left(\frac{\theta^l_{wp}}{\theta^l_s}\right)^n \tag{B.11}$$

$$REXP = \left(\frac{\theta^l_{wp}}{\theta_{wp,sand} - 0.001)}\right)^{1/2} \tag{B.12}$$

where μ is defined as follows:

$$\mu = 3.5(\theta^l_s - \theta^l_{fld})^{1.66} \tag{B.13}$$

$$ZPERC = \frac{LZTWM + LZFSM(1 - LZSK) + LZFPM(1 - LZPK)}{LZFSM\ LZSK + LZFPM\ LZPK} \tag{B.14}$$

Appendix C

The electronic supplement is made up of an instruction file and five archives: nlcd_proc.tgz, param_ssurgo.tgz, grass_prog.tgz, and util_prog.tgz, and grass_dir.tgz. The first archive contains the scripts needed for processing NLCD data. The second archive contains the column title table and the scripts for SSURGO preprocessor, parameter generator and postprocessor. The third and fourth archives provide the GRASS extensions and utility programs for postprocessing, respectively. The fifth archive provides an example of GRASS directory structure and the GRASS projection files for NLCD, SSURGO, and parameter products.

References

Anderson, J. R., E. E. Hardy, and J. T. Roach, 1972: A land-use classification system for use with remote-sensor data. *U. S. Geological Survey Circular*, **671**, 16 pp.

Anderson, R. M., V. Koren, and S. Reed, 2006: Using SSURGO data to improve Sacramento model a priori parameter estimates. *J. Hydrology*, **320**, 103–116.

Becker, R. A., J. M. Chambers, and A. R. Wilks, 1988: *The New S Language*. Chapman & Hall, London.

Burnash, R. J. C., 1995: The NWS river forecast system – catchment modeling. *Computer Models of Watershed Hydrology*, V. P. Singh, ed., Water Resources Publications, Littleton, Colorado, 311–366.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire, 1973: A generalized streamflow simulation system - conceptual modeling for digital computers. Technical report, United States Department of Commerce, National Weather Service and State of California, Department of Water Resources.

Carter, G. C., 2002: Infusing new science into the national weather service river forecast system. 2nd Federal Interagency Hydrologic Modeling Conference, Las Vegas, NV.

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan, 2004: Development of a 2001 national land-cover database for the United States. *Photogrammetric Engineering & Remote Sensing*, **70**, 829–840.

Koren, V., S. M. Reed, M. Smith, Z. Zhang, and D.-J. Seo, 2004: Hydrology Laboratory Research Modeling System (HL-RMS) of the US National Weather Service. *J. Hydrology*, **291**, 297–318.

Koren, V., M. Smith, and Q. Duan, 2003: Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall-runoff models. *Calibration of Watershed Models, Water Science and Applications*, Q. Duan, S. Sorooshian, H. Gupta, H. Rosseau, and R. Turcotte, eds., AGU, volume 6, 239–254.

Koren, V. I., M. Smith, D. Wang, and Z. Zhang, 2000: Use of soil property data in the derivation of conceptual rainfall-runoff model parameters. Proceedings of the 15th Conference on Hydrology, AMS,, Long Beach, CA, 103–106.

McCuen, R. H., 1982: *A guide to hydrologic analysis using SCS methods*. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

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

NRCS-ARS, 2004: *Hydraulics and Hydrology - Technical References NRCS National Engineering Handbook*. U.S. Department of Agriculture, Natural Resources Conservation Service, available at http://www.wcc.nrcs.usda.gov/hydro/hydro-techref-neh-630.html.

Reed, S., 2003: Deriving flow directions for coarse-resolution (1–4 km) gridded hydrologic modeling. *Water Resources Research*, **39**, 1231.

Reed, S., J. Schaake, and Z. Zhang, 2007: A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrol.*, **337**, 402–420.

Reed, S. M., V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, and D. Participants, 2004: Overall distributed model intercomparison project results. *J. Hydrology*, **298**, 27–60.

Reed, S. M. and D. R. Maidmant, 1999: Coordinate transformations for using NEXRAD data in GIS-based hydrologic modeling. *J. Hydrol. Engrg.*, **4**, 174–182.

Smith, M. B., D.-J. Seo, V. I. Koren, S. M. Reed, Z. Zhang, Q. Duan, F. Moreda, and S. Cong, 2004: The distributed model intercomparison project (DMIP): motivation and experiment design. *J. Hydrology*, **298**, 4–26.

Vogelmann, J., S. Howard, L. Yang, C. Larson, B. Wylie, and J. Van Driel, 2001: Completion of the 1990s national land cover data set for the conterminous United States. *Photogrammetric Engineering & Remote Sensing*, **67**, 650–662.

Wang, X. and A. M. Melesse, 2006: Effects of statsgo and ssurgo as inputs on swat model's snowmelt simulation. *J. American Resources Association*, **42**, 1217–1236.

Zhang, Z., V. Koren, S. M. Reed, M. Smith, and F. Moreda, 2006: Comparison of simulation results using SSURGO-based and STATSGO-based parameters in distributed hydrologic modeling. 3rd Federal Interagency Hydrologic Modeling Conference, Reno, NV.

List of Figures

List of Tables

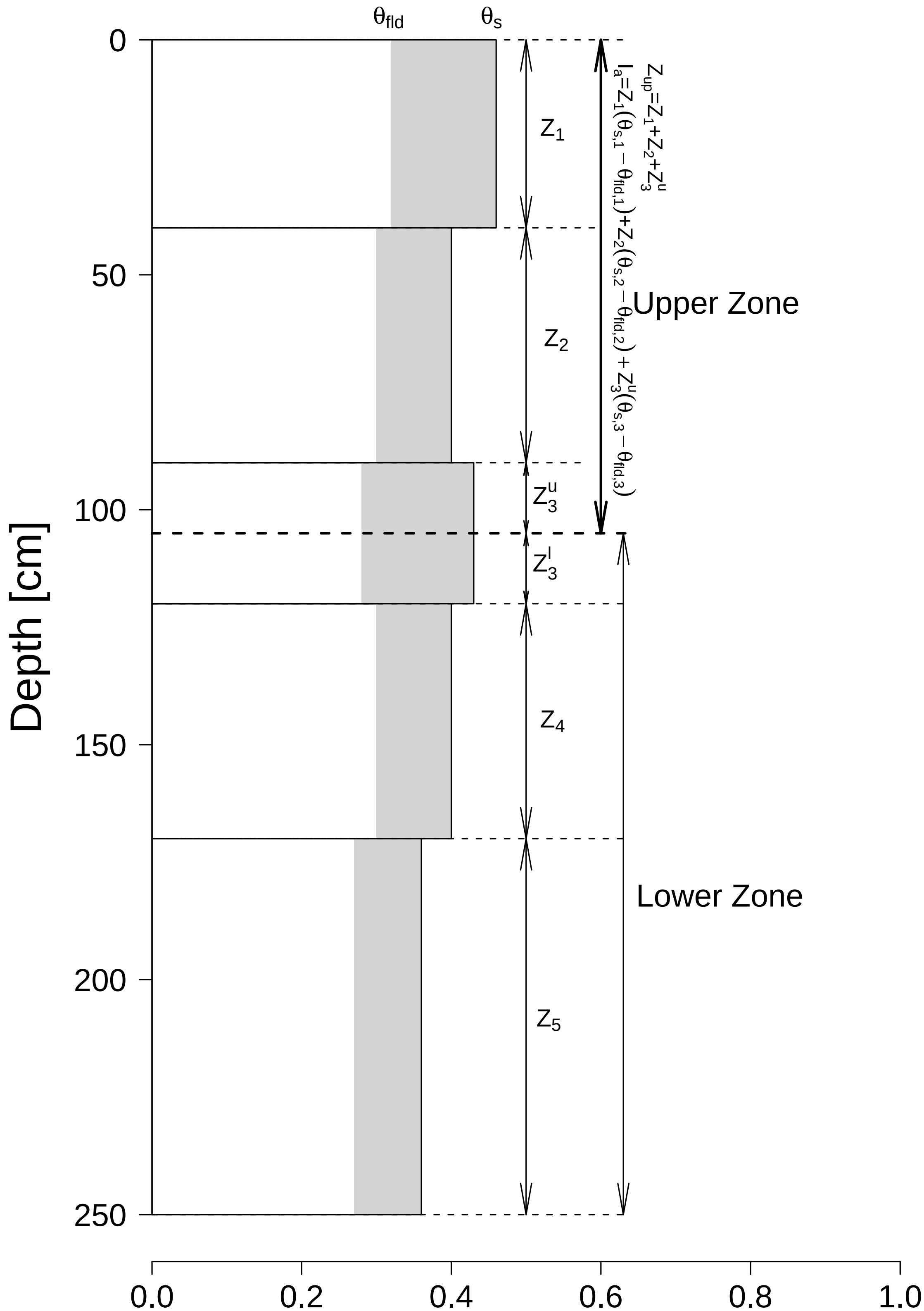

Figure 1: Schematic for the method of estimating Z_{up}.

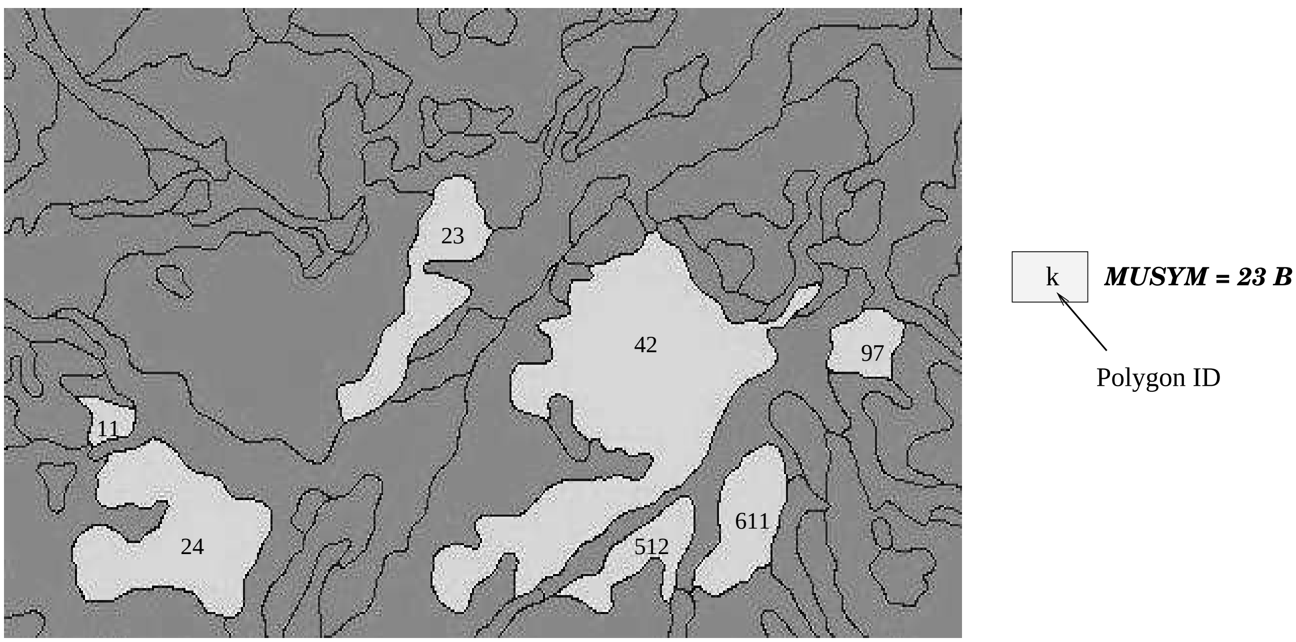

Figure 2: Spatial pattern of a map unit labeled "23B" (given by MUSYM). This map unit contains seven non-contiguous polygons.

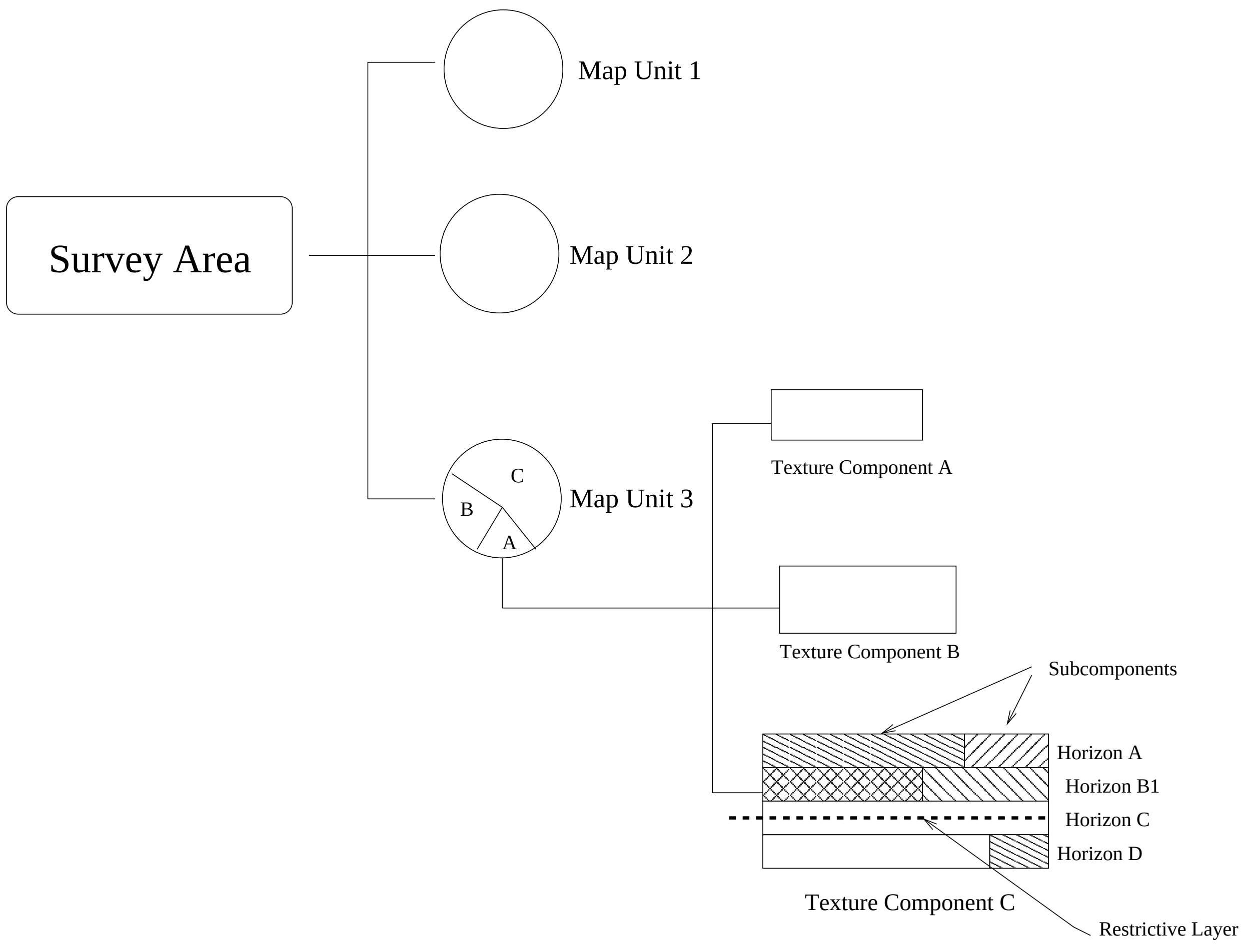

Figure 3: Hierarchy of SSURGO data (uniquely defined by MUSYM).

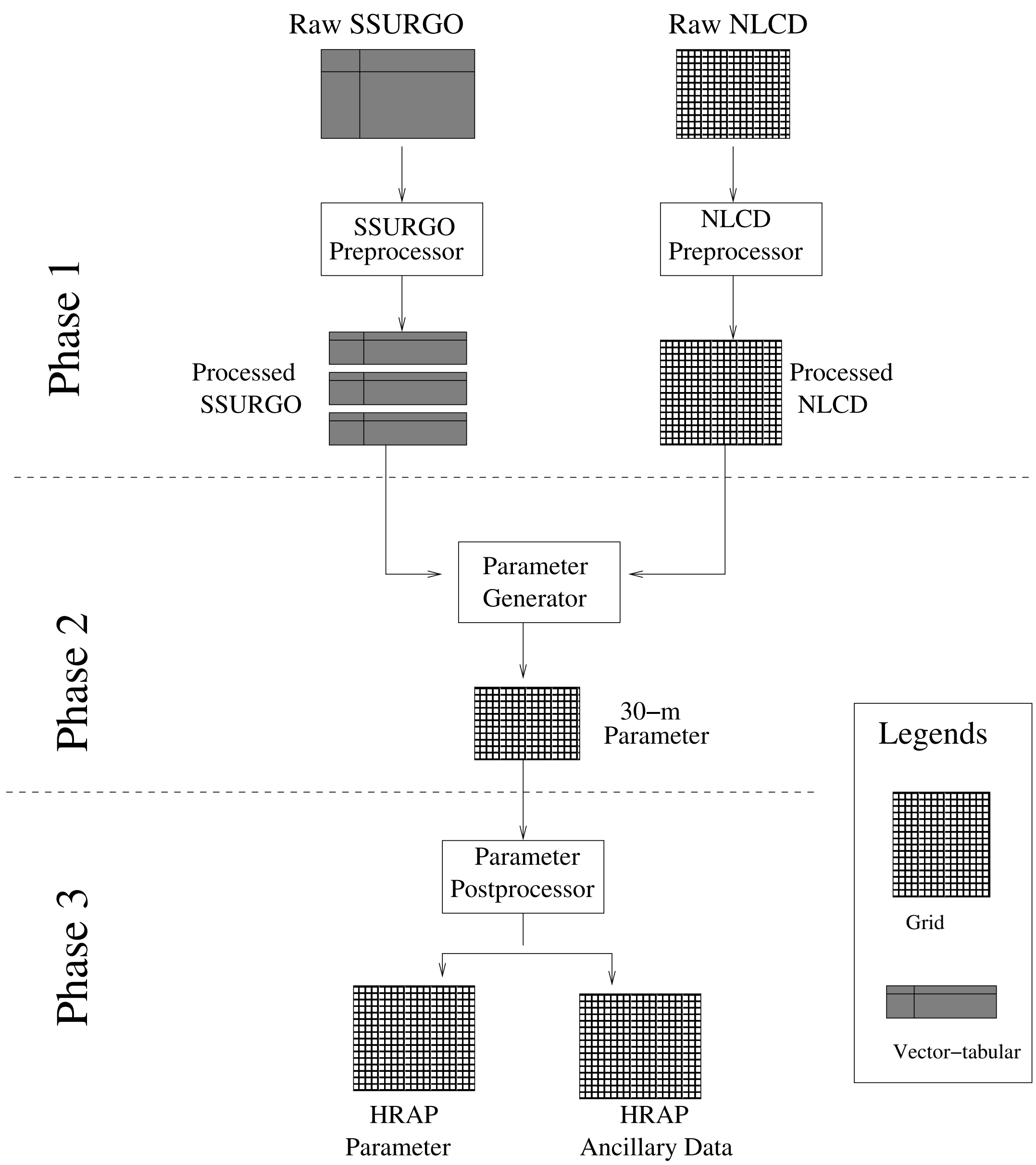

Figure 4: Schematic of processing procedures.

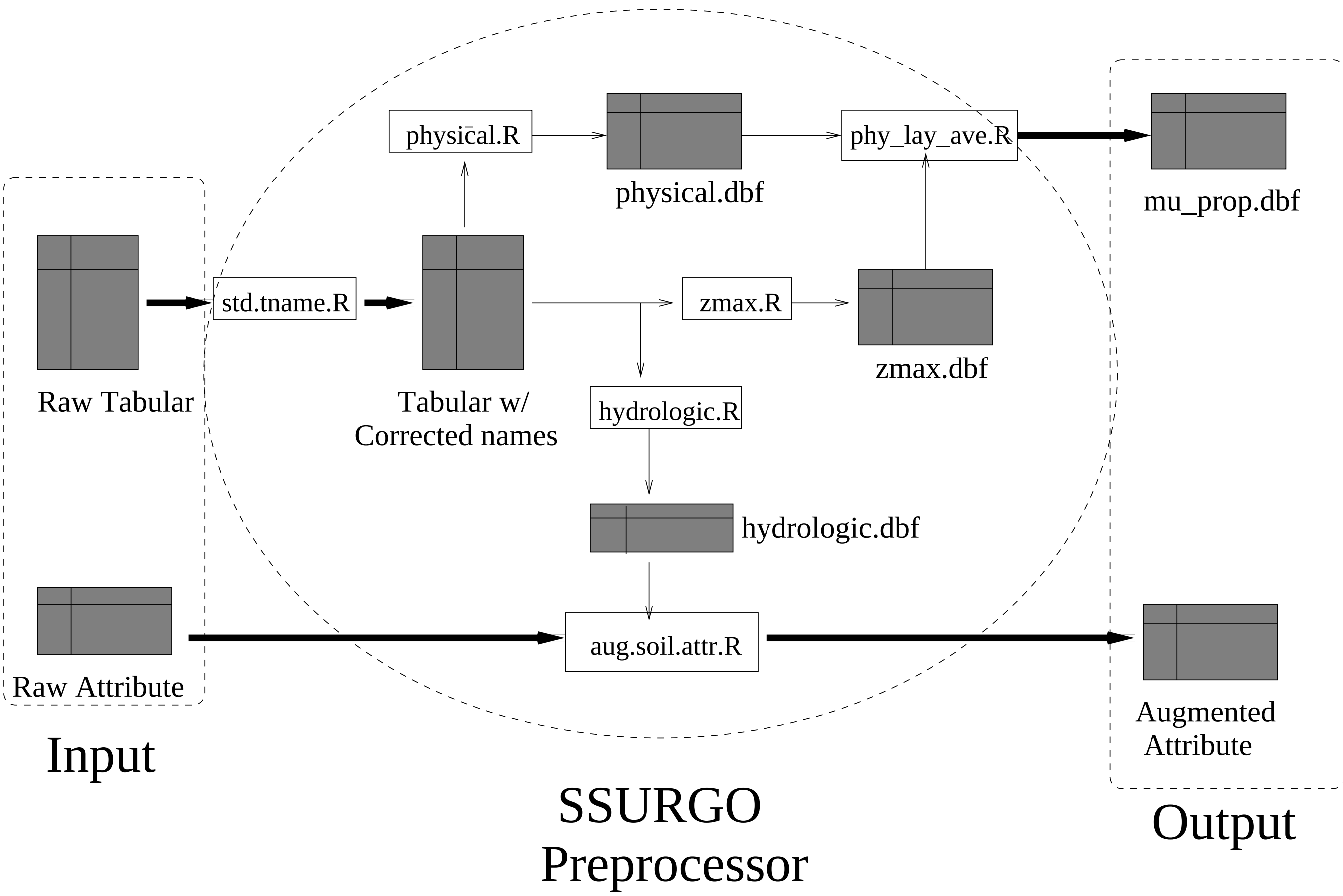

Figure 5: Data flow diagram for SSURGO Preprocessor.

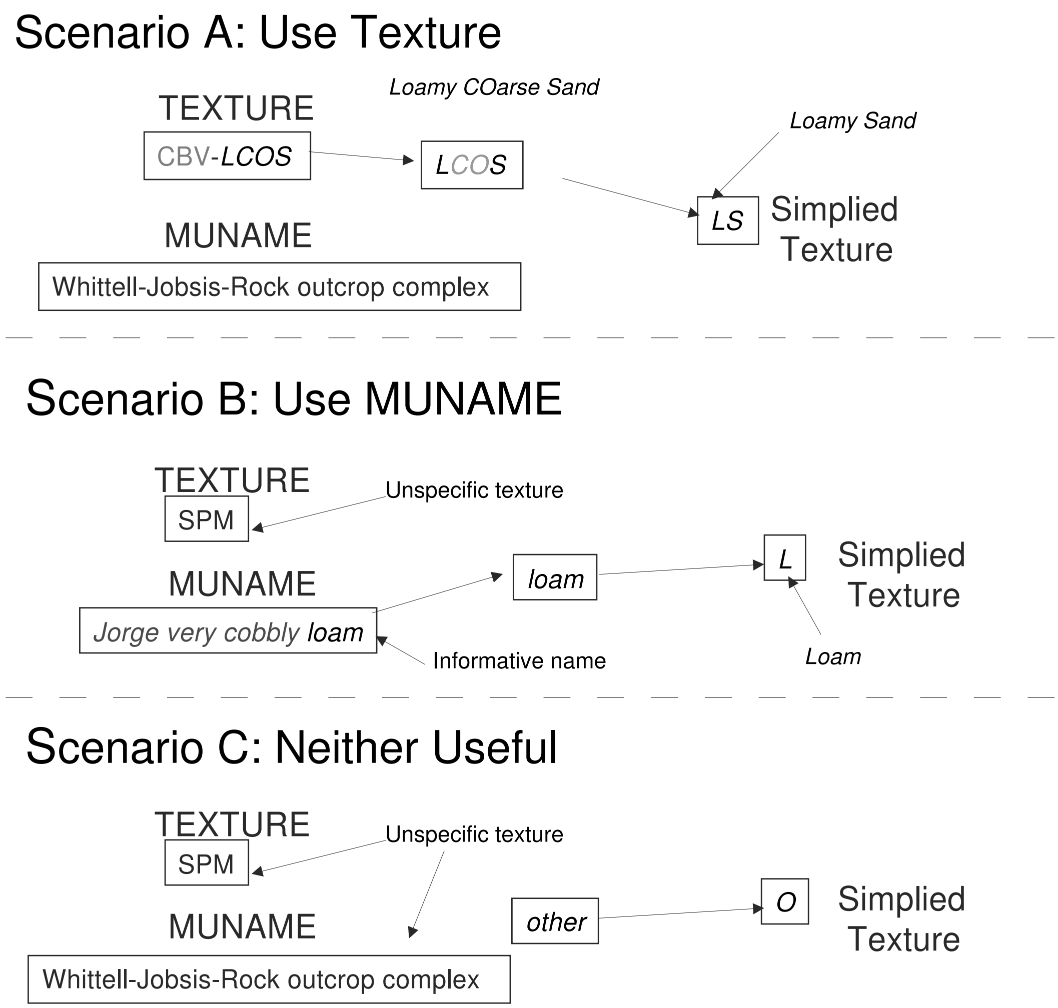

Figure 6: Illustration of texture mapper.

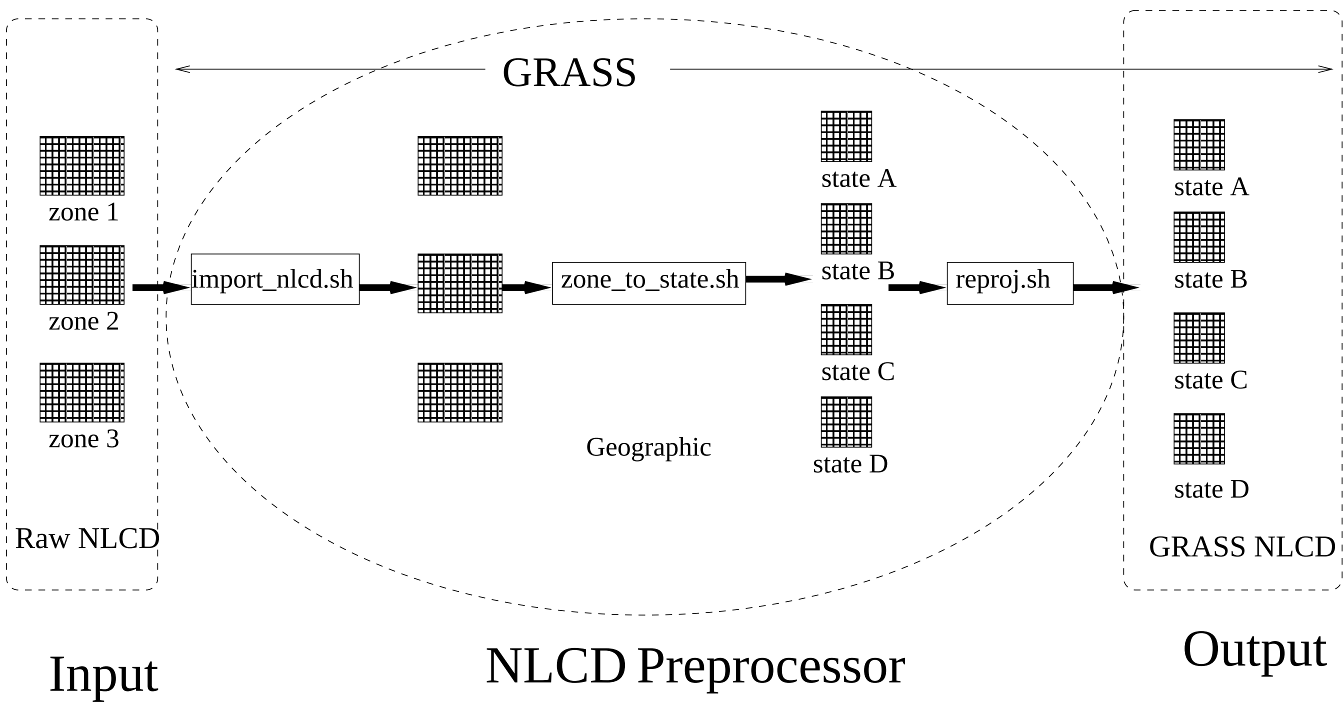

Figure 7: Data flow diagram for NLCD 2001 Preprocessor.

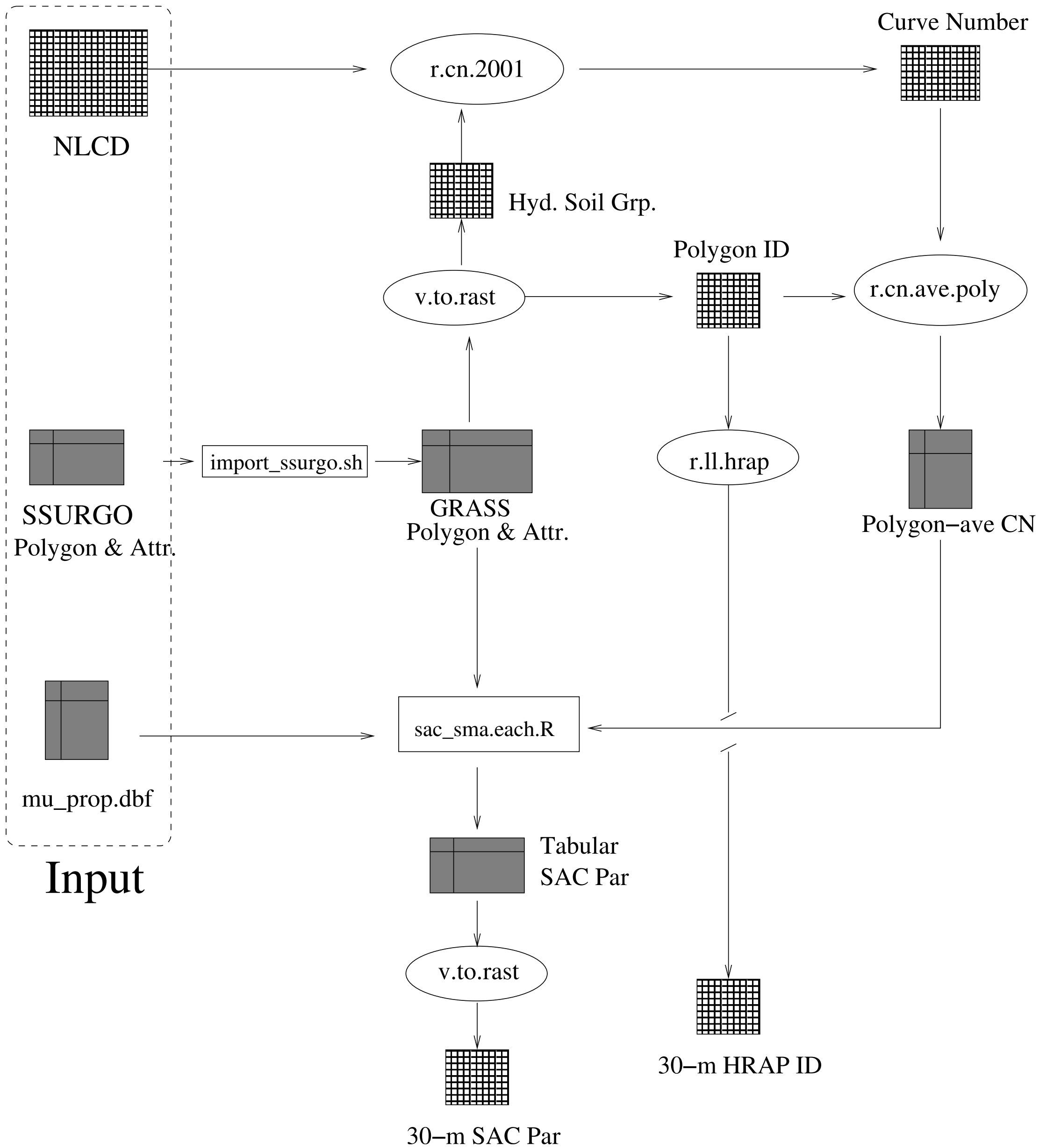

Figure 8: Data flow diagram for Parameter Generator. Among the inputs, NLCD data is the output from the NLCD 2001 preprocessor (Fig. 7), whereas the augmented attribute and mu_prop.dbf are outputs from the SSURGO preprocessor (Fig. 5).

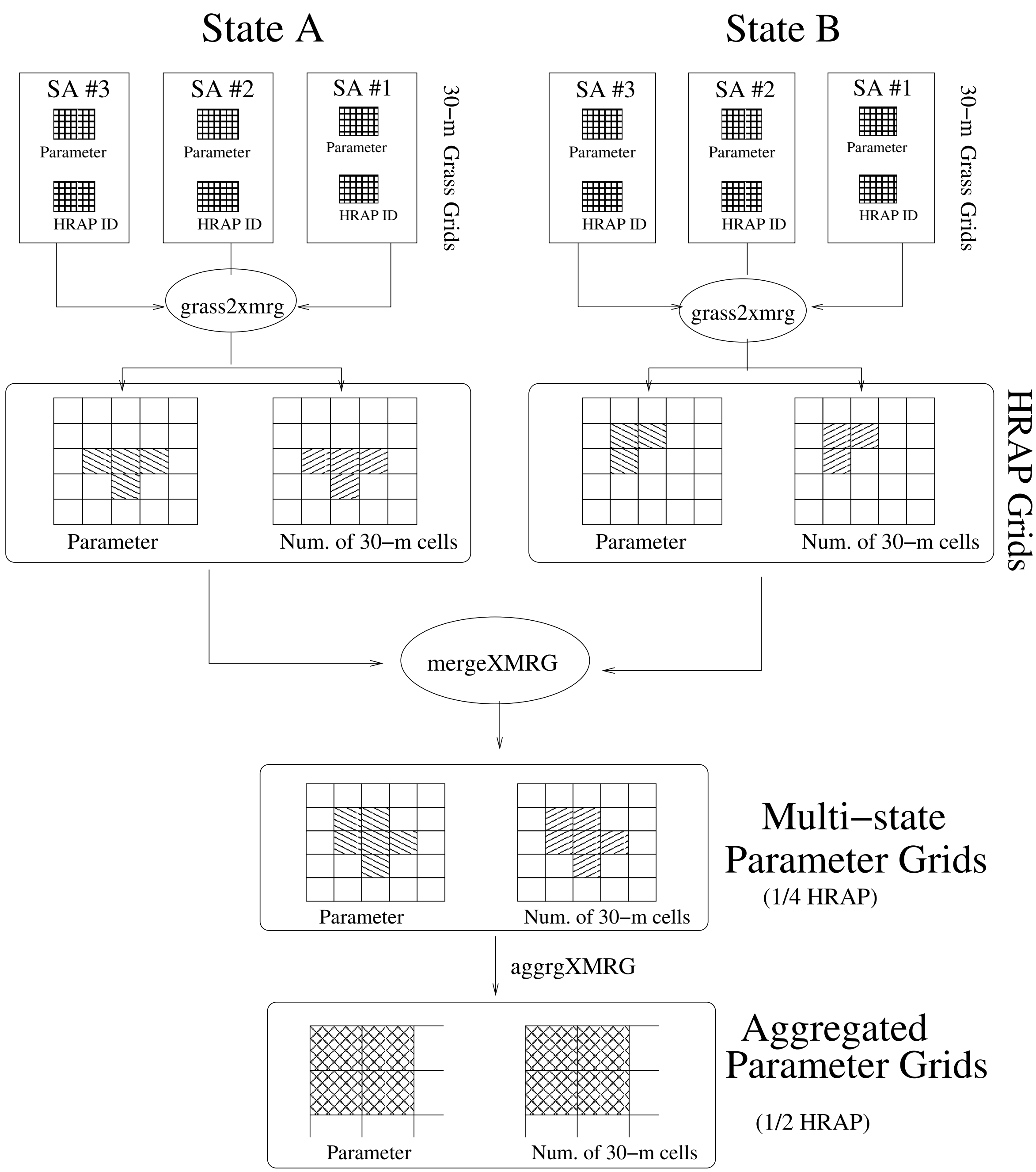

Figure 9: Data flow diagram for Parameter Postprocessor.

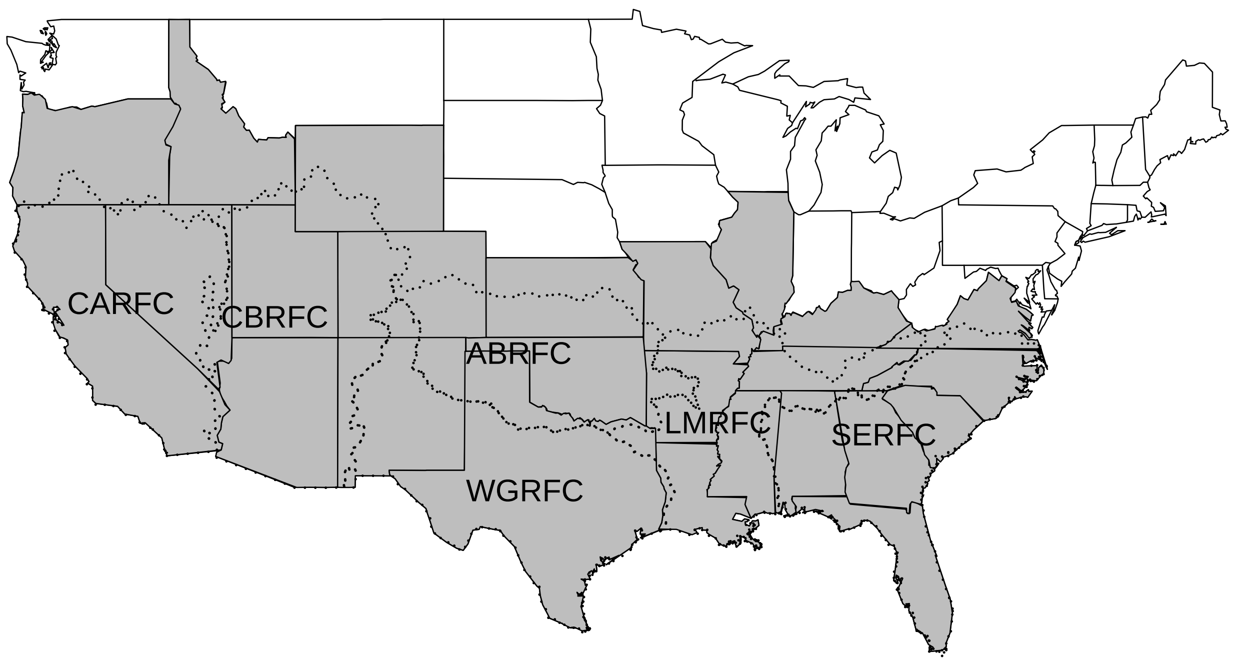

Figure 10: The states for which the parameters were derived. Superimposed are the geographic domains of six RFCs, namely, California-Nevada (CNRFC), Colorado Basin (CBRFC), Arkansas Red River Basin(ABRFC), West Gulf (WGRFC), Lower Missisippi (LMRFC) and Southeast (SERFC).

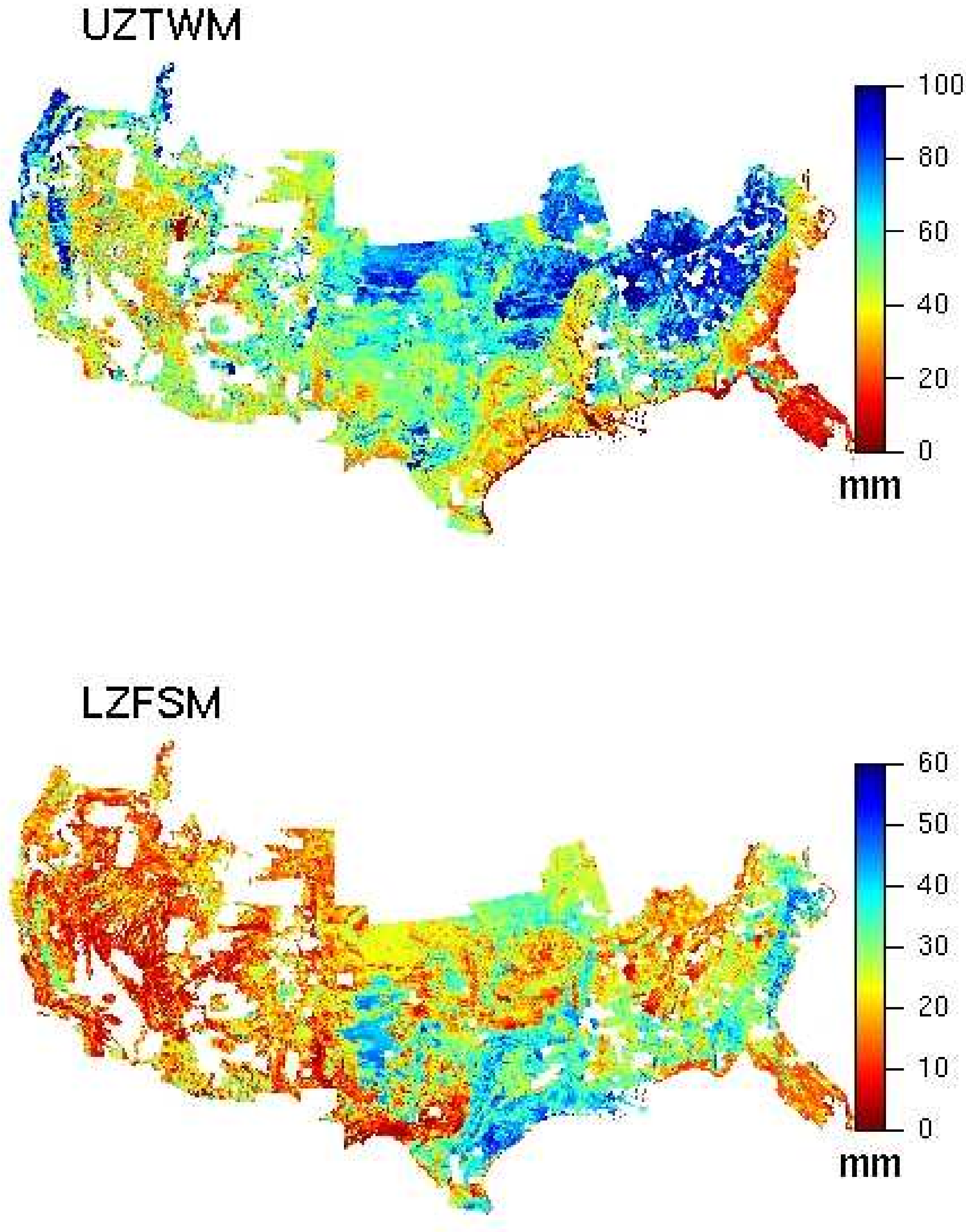

Figure 11: SSURGO-based, 25-state composite a priori grids (HRAP resolution) of UZTWM (top) and LZFSM (bottom). Left blank are areas where SSURGO data is currently unavailable, or no meaningful soil texture can be derived from the SSURGO data, or curve number reaches 100 (impervious surface).

Table 1: SAC-SMA Parameters

Symbol	Name	Typical Range[a]
UZTWM	Upper zone tension water capacity, mm	10-300
UZFWM	Upper zone free water capacity, mm	5-150
UZK	Interflow depletion rate, day^{-1}	0.1-0.75
ZPERC	Ratio of maximum and minimum percolation rates	5-350
REXP	Shape parameter of the percolation curve	1-5
LZTWM	The lower zone tension water capacity, mm	10-500
LZFSM	The lower zone supplemental free water capacity,mm	5-400
LZFPM	The lower zone primary free water capacity, mm	10-1000
LZSK	Depletion rate of lower zone supplemental free water storage, day^{-1}	0.01-0.35
LZPK	Depletion rate of lower zone primary free water storage, day^{-1}	0.001-0.05
PFREE	Percolation fraction that goes directly to the lower zone free water	0.0-0.8

[a]Ranges are based on lumped model calibration and do not necessarily constrain gridded values.

Table 2: NLCD 2001 Classes and Curve Number

Classes	ID		CN by Hydraulic Group			
		N/A	A,A/D	B,B/D	C,C/D	D
Water						
Water	11	-9999	100	100	100	100
Ice/Snow	12	-9999	95	95	95	95
Developed Areas						
Open Space	21	-9999	29	48	61	69
Low Intensity	22	-9999	40	56	67	74
Medium Intensity	23	-9999	58	70	79	83
High Intensity	24	-9999	70	79	84	87
Barren						
Bare Rock/Sand/Clay	31	-9999	95	95	95	95
Unconsolidated Shore	32	-9999	58	72	81	87
Forested Upland						
Deciduous Forest	41	-9999	19	39	53	61
Evergreen Forest	42	-9999	19	39	53	61
Mixed Forest	43	-9999	19	39	53	61
Shrubland						
Dwarf Scrub - Alaska	51	-9999	34	51	64	77
Shrub/Scrub Areas dominated by shrubs	52	-9999	34	52	64	72
Non-Natural Woody						
Orchards/Vineyards/Other	61	-9999	24	44	57	66
Herbaceous Upland						
Grasslands/Herbaceous	71	-9999	29	48	61	69
Sedge/Herbaceous - Alaska only	72	-9999	28	46	58	67
Lichens - Alaska only	73	-9999	47	61	72	77
Moss- Alaska only	74	-9999	47	61	72	77
Planted/Cultivated						
Pasture/Hay	81	-9999	29	48	61	69
Cultivated Crops	82	-9999	45	57	66	70
Wetland						
Woody Wetlands	90	-9999	100	100	100	100
Palustrine Forested Wetland	91	-9999	100	100	100	100
Palustrine Scrub/Shrub Wetland	92	-9999	100	100	100	100
Estuarine Forested Wetland	93	-9999	100	100	100	100
Estuarine Scrub/Shrub Wetland	94	-9999	100	100	100	100
Emergent Herbaceous Wetlands	95	-9999	100	100	100	100
Palustrine Emergent Wetland (Persistent)	96	-9999	100	100	100	100
Estuarine Emergent Wetland	97	-9999	100	100	100	100
Palustrine Aquatic Bed	98	-9999	100	100	100	100
Estuarine Aquatic Bed	99	-9999	100	100	100	100

Table 3: Highlights of Differences between the Previous and Current Approaches

Component	Task	Previous		Current	
		Software	Features	Software	Features
SSURGO Preprocessor	Table Extraction	MS-Access	GUI manual	R	offline, automated
SSURGO Preprocessor	Texture Mapping	MS-Excel Arcview	manual region-specific	R	automated region-independent
NLCD Preprocessor	NLCD Processing	Arcview	GUI manual	GRASS	offline automated
Parameter Generator	Parameter Generation	Arcview	GUI manual	GRASS/R GRASS/R	offline automated
Postprocessor	Postprocess	C/ArcInfo	requires ArcInfo Lib.	C++/GRASS	requires GRASS Lib.

Table 4: SSURGO Preprocessor

Script	Written In	Function
preprocessor.sh	BASH	Create directories and run the following R scripts
std.tname.R	R	Standardizes names of tabular files
hydrologic.R	R	Extracts hydraulic soil groups
physical.R	R	Extracts soil horizons and texture
zmax.R	R	Computes maximum depth of the soil layers
phy_lay_ave.R	R	Computes horizon-averaged soil properties
aug.soil.attr.R	R	Adds drainage group to SSURGO attribute table

Table 5: Soil properties

ID	Symbol	Texture	θ_s	θ_{fld}	θ_{wp}	K_s [mm h^{-1}]	μ
1	S	Sand	0.37	0.15	0.04	634.6	0.29
2	LS	Loamy Sand	0.39	0.19	0.05	562.6	0.23
3	SL	Sandy loam	0.42	0.27	0.09	124.8	0.15
4	SIL	Silt loam	0.47	0.35	0.15	25.9	0.10
5	SI	Silt	0.48	0.34	0.11	20.0	0.12
6	L	Loam	0.44	0.30	0.14	25.0	0.13
7	SCL	Sandy Clay Loam	0.42	0.29	0.16	22.7	0.12
8	SICL	Silty Clay Loam	0.48	0.41	0.24	6.1	0.04
9	CL	Clay Loam	0.45	0.36	0.21	8.8	0.07
10	SC	Sandy Clay	0.42	0.33	0.21	7.8	0.07
11	SIC	Silty Clay	0.48	0.43	0.28	3.7	0.02
12	C	Clay	0.46	0.40	0.28	4.6	0.03
13	O	Other	0.60	0.60	0.53	0.1	0.01

Table 6: NLCD Preprocessor

Script	Written In	Function
import_2001.sh	GRASS/BASH	Imports NLCD data into GRASS
zone_to_state.sh	GRASS/BASH	Derives state-wise NLCD data
reproj.sh	GRASS/BASH	Reproject NLCD to Geographic

Table 7: Parameter Generator

Script	Written In	Function
param.gen.2001.sh	BASH	Shell wrapper that runs the scripts below
import_ssurgo.sh	GRASS/BASH	Imports SSURGO shapefile into GRASS
r.cn.2001	GRASS/C	Computes Curve Number for NLCD 2001
r.cn.ave.poly	GRASS/C	Computes polygon-mean Curve Number
r.ll.hrap	GRASS/C	Computes HRAP ID from lat/lon location
sac_sma.each.R	R	Computes SAC-SMA parameters for each polygon

Table 8: Parameter Postprocessor

Script	Written In	Function
grass2xmrg	GRASS/C++	Merges 30-m parameter grids to 1/4 HRAP resolution
mergeXMRG	C++	Merges multiple 1/4 HRAP-based parameters grids
aggrgXMRG	C++	Aggregates 1/4 HRAP parameters to coarser resolution grids

Table 9: Processing Time for Previous and Present Approach (per survey area)

Task	Previous [h]	Present [h]
SSURGO Preprocessing	1	0.01
Parameter Generation	3	0.22
Parameter Aggregation	2	0.01
Total	6	0.24