Modeling soil depth from topographic and land cover attributes

Teklu K. Tesfa,[1] David G. Tarboton,[1] David G. Chandler,[2] and James P. McNamara[3]

Received 19 September 2008; revised 5 June 2009; accepted 30 June 2009; published 29 October 2009.

[1] Soil depth is an important input parameter in hydrological and ecological modeling. Presently, the soil depth data available in national soil databases (STATSGO and SSURGO) from the Natural Resources Conservation Service are provided as averages within generalized land units (map units). Spatial uncertainty within these units limits their applicability for distributed modeling in complex terrain. This work reports statistical models for prediction of soil depth in a semiarid mountainous watershed that are based upon the relationship between soil depth and topographic and land cover attributes. Soil depth was surveyed by driving a rod into the ground until refusal at locations selected to represent the topographic and land cover variation in the Dry Creek Experimental Watershed near Boise, Idaho. The soil depth survey consisted of a model calibration set, measured at 819 locations over 8 subwatersheds representing topographic and land cover variability and a model testing set, measured at 130 more broadly distributed locations in the watershed. Many model input variables were developed for regression to the field data. Topographic attributes were derived from a digital elevation model. Land cover attributes were derived from Landsat remote sensing images and high-resolution aerial photographs. Generalized additive and random forests models were developed to predict soil depth over the watershed. They were able to explain about 50% of the soil depth spatial variation, which is an important improvement over the soil depth extracted from the SSURGO national soil database.

Citation: Tesfa, T. K., D. G. Tarboton, D. G. Chandler, and J. P. McNamara (2009), Modeling soil depth from topographic and land cover attributes, *Water Resour. Res.*, *45*, W10438, doi:10.1029/2008WR007474.

1. Introduction

[2] Soil depth is one of the most important input parameters for hydroecological models. Spatial patterns in soil depth arise from complex interactions of many factors (topography, parent material, climate, biological, chemical and physical processes) [*Jenny*, 1941; *Hoover and Hursh*, 1943; *Summerfield*, 1997]. As a result, prediction of soil depth at a point is difficult. Spatial patterns in soil depth significantly affect soil moisture, runoff generation, and subsurface and groundwater flow [*Freer et al.*, 2002; *Stieglitz et al.*, 2003; *McNamara et al.*, 2005; *Seyfried et al.*, 2009; *Gribb et al.*, 2009]. Soil depth also provides an indication of the available water capacity, and exerts a major control on biological productivity [*Gessler et al.*, 1995], which in turn affects evapotranspiration. Consequently, accurate representation of soil depth at scales relevant to these processes is increasingly important for use in distributed simulation models of hydrology and ecology. Soil depth is highly variable spatially and laborious, time consuming and difficult to practically measure even for a modestly sized watershed [*Dietrich et al.*, 1995]. There is thus a need for models that can predict the spatial pattern of soil depth.

[3] In the United States, the Natural Resources Conservation Service (NRCS) national soil databases (SSURGO and STATSGO) have been the main sources of soil depth information used as input for hydroecological modeling [*Anderson et al.*, 2006]. In these databases, soils are spatially represented as discrete map units with sharp boundaries. A map unit may comprise multiple soil components but these components are not represented spatially within the map unit. As a result, soil attributes are spatially represented at map unit level as a mean or some other representative value of the components. Such a representation of soils is discrete, highly generalized and is incompatible with other landscape data derived from digital elevation models [*Moore et al.*, 1993; *Zhu*, 1997; *Zhu and Mackay*, 2001; *Schmidt et al.*, 2005]. This limits applicability for spatially distributed hydroecological modeling.

[4] Various approaches have been explored to improve the characterization of soil properties over landscapes to overcome the limitations of existing soil databases created using conventional soil survey methods [*Mark and Csillag*, 1989; *Goodchild*, 1992; *Moore et al.*, 1993; *Bierkens and Burrough*, 1993; *Zhu and Band*, 1994; *Dietrich et al.*, 1995; *Zhu et al.*, 1996, 1997; *McBratney and Odeh*, 1997]. These are part of a general effort to develop and refine the spatial data for use in hydroecological modeling at scales consistent with other spatially distributed model inputs. *Schulz et al.*

[1] Civil and Environmental Engineering Department, Utah State University, Logan, Utah, USA.
[2] Civil Engineering Department, Kansas State University, Manhattan, Kansas, USA.
[3] Department of Geosciences, Boise State University, Boise, Idaho, USA.

0043-1397/09/2008WR007474$09.00

[2006] reviewed the importance of spatial data representation to advance understanding of hydrological processes in general. As models improve in their ability to capture small-scale details, the structure of spatial patterns in the data becomes increasingly important [*Grayson and Blöschl*, 2000]. In the context of soils, the spatial pattern with respect to topography has long been used by soil mappers [*Mark and Csillag*, 1989; *Goodchild*, 1992; *McBratney et al.*, 2003; *Scull et al.*, 2003] and recently is the focus of efforts in hydropedology that examine synergies between soil and hydrological processes [*Lin et al.*, 2006].

[5] Fuzzy logic has been suggested as an approach to refine the scale of soil information [*McBratney and Odeh*, 1997]. In particular *Zhu and Band* [1994] and *Zhu et al.* [1996, 1997] have developed a model SoLIM that combines fuzzy logic with GIS and expert system development techniques that capture the opinions of experts in the fuzzy logic functions used to map soil attributes from spatial soil forming factors. *Zhu and Mackay* [2001] took this approach one step further and evaluated the effects of spatial detail of soil information, generated with this model on watershed hydrological response. They showed that detailed spatial soil information influenced simulated hydrographs and net photosynthesis, underscoring the importance of detailed spatial soil information for hydroecological modeling. A limitation of this work was that no observed hydrographs were available for the watershed simulated, so it was not possible to quantify the improvement in hydrologic simulations because of the more detailed soil information. There are also concerns regarding the subjectivity of expert opinions captured in a model such as SoLIM.

[6] *Moore et al.* [1993] and *Gessler et al.* [1995] applied statistical approaches to model the pattern of soil properties over landscape. Relationships between soil properties and landscape factors (e.g., slope, wetness index, and plan curvature) were first extracted from point measurements and then used to predict soil properties over the remaining area. Geostatistical approaches have also been used to interpolate soil properties [*Bierkens and Burrough*, 1993; *Odeh et al.*, 1994; *Zhang et al.*, 1995; *Zhu*, 1997], but their application is often limited by the large amount of data required to define the spatial autocorrelation.

[7] In contrast to the statistical approaches mentioned above, *Dietrich et al.* [1995] suggested a process-based approach for predicting the spatial variation of colluvial soil depth. By assuming (1) that soil production is a function of soil depth (2) and that soil transport is proportional to slope and (3) that soil production is in local dynamic equilibrium with the divergence of soil transport, topographic curvature becomes a surrogate for soil production. Observations of cosmogenic ^{10}Be and ^{26}Al concentrations from bedrock, reported by *Heimsath et al.* [1997; 1999], validated the relationship between curvature and soil production with an exponentially decreasing dependence of soil production on depth for their Tennessee Valley site in California. These ideas have been further pursued in other areas [*Heimsath et al.*, 2000, 2001]. *Heimsath et al.* [2005] showed that on steep slopes a depth-dependent transport model is more broadly applicable. *Saco et al.* [2006] incorporated these ideas into a landscape evolution model that was used to evaluate the dependence of soil production on simulated soil moisture. This provided a mechanism whereby soil depths could vary spatially even under conditions of dynamic equilibrium, where a soil production function dependent only on soil depth would predict constant soil depth.

[8] The various modeling approaches for predicting soil depth over landscapes, described above, showed only partial success. While the physically based model has shown reasonable prediction capability in unchanneled valleys [*Dietrich et al.*, 1995], the cause of the exponential soil production function has not been explained and the potential dependence on other factors, such as soil moisture has only had limited evaluation. The roles of chemical and physical breakdown of the underlying rock and its influence on soil production, and the effects of various topographic factors (aspect, slope, elevation etc.) are not explicitly considered in these models.

[9] In this paper, we develop statistical models for prediction of the spatial pattern of soil depth over complex terrain from topographic and land cover attributes. We introduce new topographic attributes, derived from a digital elevation model (DEM), intended to have explanatory capability for soil depth. Various land cover attributes were derived from Landsat remote sensing images. Generalized additive models (GAM) and random forests (RF) statistical modeling techniques were applied to predict soil depth from these topographic and land cover attributes using soil depth data measured at 819 points in 8 subwatersheds within the Dry Creek Experimental Watershed (DCEW). This calibration data set was randomly divided into a training subset consisting of 75% of the data and a validation subset consisting of the remaining 25% that was used to estimate the prediction error for variable and model complexity selection [see, e.g., *Hastie et al.*, 2001, chapter 7]. Soil depth data measured at an additional 130 more broadly distributed locations within DCEW was used as an out of sample data set to test the model results. Predicted and measured soil depth was also aggregated at the scale of SSURGO map units and compared to soil depth from the SSURGO soil database.

2. Study Area

[10] This study was carried out in the Dry Creek Experimental Watershed (DCEW), about 28 km^2 in area, located in the semiarid southwestern region of Idaho approximately 13 km northeast of the city of Boise, United States (Figure 1). The general area, known as the Boise Front, comprises mountainous and foothills topography. Elevations in the DCEW range from 1000 m at the outlet where Dry Creek crosses Bogus Basin Road to 2100 m at the highest headwaters [*Williams*, 2005; *McNamara et al.*, 2005; *Williams et al.*, 2008]. The average slope is about 25%, with steeper north facing slopes than south facing slopes.

[11] The climate of DCEW has been classified by *McNamara et al.* [2005] using the Koeppen climate classification system [*Henderson-Sellers and Robinson*, 1986] as a steppe summer dry climate (BSk) for the lower part and moist continental climate with dry summers (Dsa) for the upper part. Precipitation is highest in winter, as snow in the highlands and rain in the lowlands, and in spring in the form of rain. There are occasional summer thunderstorms. Autumns are generally dry [*Williams*, 2005; *Williams et al.*, 2008]. The average annual precipitation ranges from 37 cm

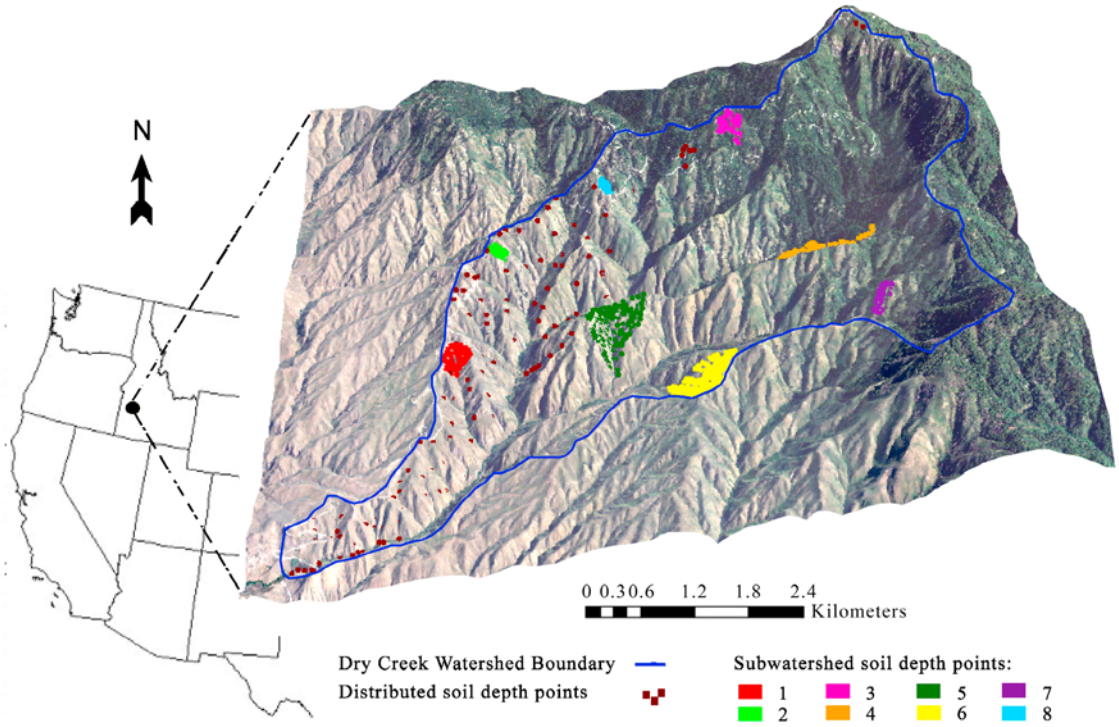

Figure 1. Dry Creek Experimental Watershed (DCEW) near Boise, Idaho, in the western United States. Points show locations where soil depth was sampled. Extent of DCEW is longitude 116.179–116.099°W and latitude 43.688–43.741°N.

at lower elevations to 57 cm at higher elevations [*Williams*, 2005]. Streamflow typically remains low in the early and midwinter and peaks in the early to midspring because of the annual snowmelt freshet [*McNamara et al.*, 2005].

[12] Vegetation in Dry Creek is dominated by grasses, forbs and sagebrush at lower elevations, transitioning into chaparral and then fir, spruce, and pines at higher elevations [*McNamara et al.*, 2005]. Soils are formed from weathering of the underlying Idaho Batholith, which is a granite intrusion ranging in age from 75 to 85 million years [*McNamara et al.*, 2005]. The soils range from loam to sandy loam in texture [*Williams*, 2005; *Gribb et al.*, 2009] and according to the SSURGO soil database the percentages of total sand, silt and clay range from 42% to 76%, 12% to 39% and 8% to 18% respectively. However, *Gribb et al.* [2009] reported that the gravel content can be up to 38%. The soils are generally well drained and have high surface erosion potential. Soils on the south facing slopes generally have coarser texture than soils covering the north facing slopes. South facing slopes have more rock outcrops than the north facing slopes.

3. Methodology

3.1. Field Data Collection

[13] Eight subwatersheds were selected to represent the elevation, slope, aspect and land cover variability present within DCEW. Soil depth, topographic curvature (field observed curvature), and vegetation were surveyed at a total of 819 points within the eight subwatersheds. Survey locations were chosen to represent the range of topographic and land cover variation in the subwatersheds. At each survey point the GPS location (with 3 to 6 m accuracy) was recorded. An Aerial photograph (with 1 m resolution) and field notes were used to refine the GPS positioning of the survey locations. At each location two or three soil depth replicates 2–3 m apart were collected by driving a 220 cm long 1.27 cm diameter sharpened copper coated steel rod graduated at 5 cm interval vertically into the ground using a fence post pounder until refusal. For the first set of surveys two replicate depth measurements were made and a third measurement was made if the difference between the first two was more than 20 cm. For the later surveys three depth measurement replicates were recorded at all points.

[14] The advantage of the depth to refusal method is that it is a direct and simple measurement of soil depth. It is inexpensive, albeit laborious and time consuming and limited to depths to which a rod can be pounded. A disadvantage is that the measurement is biased to underestimating actual depth to bedrock, since there is uncertainty as to what actually causes refusal. Rocks and gravel that occur as residual relicts from weathering or colluvium may limit the rod penetration resulting in underestimation of soil depth. Figure 2 illustrates the soil profile in Dry Creek at locations where pits have been dug. These illustrate some of the irregularity of soil depth and occurrence of rocks that may result in underestimation of soil depth from this depth to refusal approach. Soil depth recorded by pounding the rod to refusal at two road cuts, where bedrock was exposed, gave soil depth consistent with the visible depth to bedrock. While there is uncertainty in any one soil depth measurement due to these effects, taken in aggregate they seem to provide reliable information on soil depth.

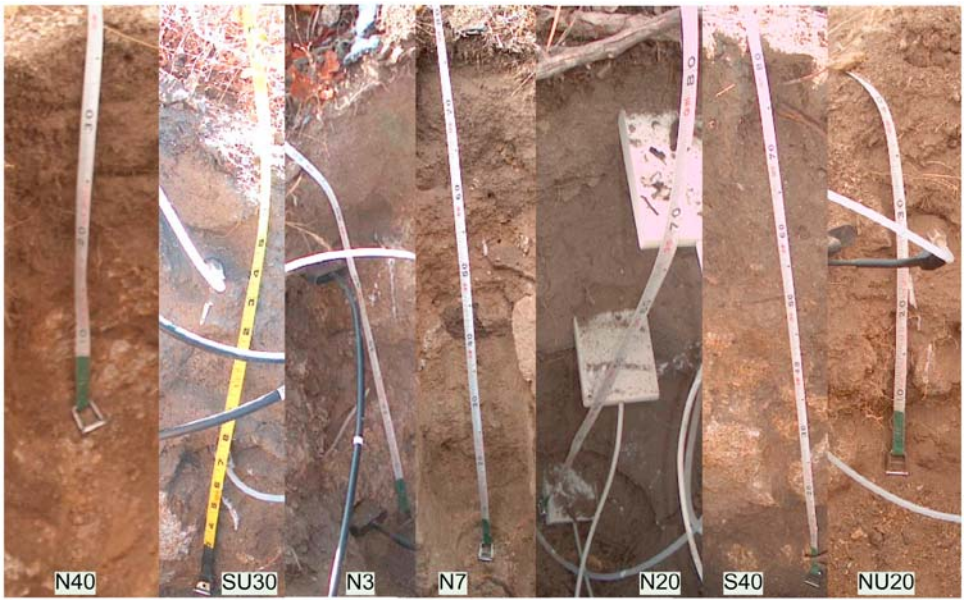

Figure 2. Illustrations of soil profiles in DCEW from soil pits.

[15] To quantify the uncertainty in our data we examined the variability of the range from replicate depth to refusal measurements at 641 of the sample points in six subwatersheds. (This information was not available for two subwatersheds where only the depth replicate average was recorded in the field). The mean depth replicate range was 9.2 cm with 95th percentile of 25 cm and maximum range of 75 cm. This indicates that although in the most extreme case the depth range was 75 cm, that for the vast majority of points the uncertainty was less that 25 cm with an average uncertainty around 10 cm.

[16] The soil depth survey was carried out in 2005 and 2006, during early spring when the soil was moist and more easily penetrated by the rod. Topographic curvature was recorded by visual assessment as concave (−1), convex (1) or intermediate (0) and the dominant land cover type was recorded as one of bare, grass, mixed grass and shrubs, shrubs, coniferous forest or deciduous forest. The first author carried out this survey for 761 of the points in seven subwatersheds, while soil depth data for 58 points in the eighth subwatershed, had been previously collected using the same methods [*Williams et al.*, 2008]. The data from these 819 points are designated as the calibration data set. A further 130 soil depth observations were collected using the same method at more broadly distributed locations, at least 50 m away from the selected subwatersheds, within the boundary of the watershed, and generally on the southwest side logistically accessible from the road. These are designated as the testing data set.

3.2. Geospatial Data

[17] Primary geospatial data used included a digital elevation model (DEM) (obtained from the USGS Web site http://seamless.usgs.gov/), Landsat TM imagery (path 41 row 30 obtained from the USGS), an aerial photograph (obtained from NRCS Idaho State Office), and the SSURGO soil database for survey area symbol ID903 (Boise Front) (obtained from NRCS Idaho State Office). A wide range of geospatial explanatory attributes were derived from the DEM and Landsat TM images.

3.2.1. Data Derived From the DEM

[18] The 1/3 arc sec DEM from the USGS seamless data server was projected to a 5 m grid for the derivation of the topographic attributes (Table 1) considered as potential explanatory variables for predicting soil depth over the landscape. Although the spatial footprint of the USGS DEM is likely 10 to 30 m, a 5 m grid resolution was chosen to limit degradation due to interpretation and projection from the geographic coordinate data provided by the USGS. Exploiting the general terrain-based flow analysis concepts for enriching the information content from digital elevation models [*Tarboton*, 1997; *Tarboton and Ames*, 2001; *Tarboton and Baker*, 2008], a number of new topographic attributes were derived from the DEM. First a flow field is derived by filling spurious sinks from the DEM then calculating flow directions, using either the D8 or D∞ flow model. The D8 model [*O'Callaghan and Mark*, 1984] assigns flow from each DEM grid cell to one downslope neighbor in the direction of steepest descent. The D∞ flow model [*Tarboton*, 1997] apportions flow between adjacent neighbors on the basis of the direction of steepest downward slope on the eight triangular facets constructed in a 3×3 grid cell window using the center cell and each two neighboring grid cells in turn. For the purposes of obtaining additional flow related derivative quantities from the DEM the important outcome from deriving the flow field is the set of proportions, P_{ij}, defining the proportion of grid cell i that drain to grid cell j. For the D8 method these are either 0, or 1, while for the D∞ model these are between 0 and 1, subject to the condition that $\sum_i P_{ij} = 1$. With the flow field defined using proportions, recursion, extending the recursive algorithms used for contributing area [*Mark*, 1988; *Tarboton*, 1997; *Tarboton and Baker*, 2008], can be used to define and compute an extensive set of derivative attributes that have potential explanatory capability for soil depth. A complete list of topographically derived explanatory

Table 1. DEM-Based Explanatory Variables

Symbol[a]	Description
elv	Elevation above sea level
sca	Specific catchment area from the D∞ method [*Tarboton*, 1997]. This is contributing area divided by the grid cell size (from TauDEM specific catchment area function).[b]
plncurv	Plan curvature is the curvature of the surface perpendicular to the direction of the maximum slope (from ArcGIS spatial analysis tools curvature function) [*Moore et al.*, 1991,1993]. A positive value indicates convex up; a negative value indicates concave up; and zero indicates flat surface.
prfcurv	Profile curvature is the curvature of the surface in the direction of maximum slope (from ArcGIS spatial analyst tools curvature function) [*Moore et al.*, 1991; 1993]. A negative value indicates convex up surface; a positive value indicates concave up; and zero indicates flat surface.
gncurv	The second derivative of the surface computed by fitting a fourth-order polynomial equation to a 3 × 3 grid cell window (from ArcGIS spatial analyst tools curvature function) [*Moore et al.*, 1991; 1993].
aspg	The direction that a topographic slope faces expressed in terms of degrees from the north (from ArcGIS spatial analyst tools aspect function)
slpg	Magnitude of topographic slope computed using finite differences on a 3 × 3 grid cell window (from ArcGIS spatial analyst tools slope function)
*ang**	D∞ flow direction [*Tarboton*, 1997]: the direction of the steepest outward slope from the triangular facets centered on each grid cell and is reported as the angle in radians counter-clockwise from east (TauDEM Dinf flow directions function)
ad8	D8 contributing area: the number of grid cells draining through each grid cell using the single flow direction model (TauDEM D8 contributing area function)
sd8	D8 slope: the steepest outward slope from a grid cell to one of its eight neighbors reported as drop/distance, i.e., tan of the angle (TauDEM D8 flow directions function)
stdist	D8 distance to stream: horizontal distance from each grid cell to a stream grid cell traced along D8 flow directions by moving until a stream grid cell as defined by the Stream Raster grid is encountered (TauDEM flow distance to Streams function)
slpt	D∞ slope [*Tarboton*, 1997]: the steepest outward slope from the triangular facets centered on each grid cell reported as drop/distance, i.e., tan of the slope angle (TauDEM Dinf flow directions function)
*plen**	D8 longest upslope length: the length of the flow path from the furthest cell that drains to each cell along D8 flow directions (TauDEM grid network order and flow path lengths function)
*tlen**	D8 total upslope length: the total length of flow paths draining to each grid cell along D8 flow directions (TauDEM grid network order and flow path lengths function)
sd8a	Slope averaged over a 100 m path traced downslope along D8 flow directions (from GRAIP, D8 slope with downslope averaging function)[c]
p	The D8 flow direction grid representing the flow direction from each grid cell to one of its adjacent or diagonal neighbors, encoded as 1 to 8 counterclockwise starting at east (TauDEM D8 flow directions function)
sar	Wetness index inverse: an index calculated as slope/specific catchment area (TauDEM wetness index inverse function).
*sph8**	D8 horizontal slope position from equation (16)
*modcurv**	Curvature modeled based on field observed curvature from equation (29).
*lhr**	Longest D∞ horizontal distance to ridge from equation (1)
*shr**	Shortest D∞ horizontal distance to ridge from equation (2)
*ahr**	Average D∞ horizontal distance to ridge from equation (3)
*lhs**	Longest D∞ horizontal distance to stream from equation (4)
*shs**	Shortest D∞ horizontal distance to stream from equation (5)
*ahs**	Average D∞ horizontal distance to stream from equation (6)
*lvr**	Longest D∞ vertical rise to ridge from equation (1) with vdist, equation (7)
*svr**	Shortest D∞ vertical rise to ridge from equation (2) with vdist, equation (7)
*avr**	Average D∞ vertical rise to ridge from equation (3) with vdist, equation (7)
*lvs**	Longest D∞ vertical drop to stream from equation (4) with vdist, equation (8)
*svs**	Shortest D∞ vertical drop to stream from equation (5) with vdist, equation (8)
*avs**	Average D∞ vertical drop to stream from equation (6) with vdist, equation (8)
*lsr**	Longest surface distance to ridge from equation (1) with sdist, equation (9)
*ssr**	Shortest surface distance to ridge from equation (2) with sdist, equation (9)
*asr**	Average surface distance to ridge from equation (3) with sdist, equation (9)
*lss**	Longest surface distance to stream from equation (4) with sdist, equation (9)
*sss**	Shortest surface distance to stream from equation (5) with sdist, equation (9)
*ass**	Average surface distance to stream from equation (6) with sdist, equation (9)
*lps**	Longest Pythagoras distance to stream from equation (10)
*sps**	Shortest Pythagoras distance to stream from equation (11)
*aps**	Average Pythagoras distance to stream from equation (12)
*lpr**	Longest Pythagoras distance to ridge from equation (13)
*spr**	Shortest Pythagoras distance to ridge from equation (14)
*apr**	Average Pythagoras distance to ridge from equation (15)
*lsph**	D∞ Longest horizontal slope position from equation (17)

Table 1. (continued)

Symbol[a]	Description
*ssph**	D∞ Shortest horizontal slope position from equation (18)
*asph**	D∞ Average horizontal slope position from equation (19)
*lspv**	Longest vertical slope position from equation (20)
*sspv**	Shortest vertical slope position from equation (21)
*aspv**	Average vertical slope position from equation (22)
*lspp**	Longest Pythagoras slope position from equation (23)
*sspp**	Shortest Pythagoras slope position from equation (24)
*aspp**	Average Pythagoras slope position from equation (25)
*lspr**	Longest slope position ratio from equation (26)
*sspr**	Shortest slope position ratio from equation (27)
*aspr**	Average slope position ratio from equation (28)

[a]Asterisk indicates a new topographic variable.
[b]TauDEM is the terrain analysis using digital elevation models software (http://hydrology.usu.edu/taudem/).
[c]GRAIP is the Geomorphologic Road Analysis Inventory Package software (http://www.engineering.usu.edu/dtarb/graip).

variables is given in Table 1 with new attributes indicated by an asterisk. Figure 3 gives definitions for variations of the newly derived distance to ridge and distance to stream attributes. Algorithms for new topographic attributes are given in sections 3.2.1.1–3.2.1.11.

3.2.1.1. Horizontal Distance to Ridge (*hr)

[19] The horizontal distance to ridge is defined as the horizontal flow distance tracing upslope from a grid cell to a grid cell that does not receive flow from an upslope neighbor computed on the basis of the D∞ flow model. Because multiple flow paths may converge at any grid cell, there may be multiple upslope ridge grid cells. We therefore define three variants of the horizontal distance to ridge function. The longest horizontal distance to ridge (*lhr*) is the flow distance to the furthest upslope ridge grid cell. The shortest horizontal distance to ridge (*shr*) is the flow distance to the nearest upslope ridge grid cell. The average horizontal distance to the ridge (*avr*) is the mean horizontal flow distance calculated by weighting the distance on the basis of the proportions of incoming flow from upslope grid cells. Numerically, these are evaluated recursively as follows:

$$lhr(x_i) = \underset{\{k:P_{ki}>0\}}{\text{Max}} \left(hdist(x_i, x_k) + lhr(x_k)\right) \quad \text{if} \quad \sum P_{ki} > 0, 0 \quad \text{otherwise} \tag{1}$$

$$shr(x_i) = \underset{\{k:P_{ki}>0\}}{\text{Min}} \left(hdist(x_i, x_k) + shr(x_k)\right) \quad \text{if} \quad \sum P_{ki} > 0, 0 \quad \text{otherwise} \tag{2}$$

$$ahr(x_i) = \frac{\sum P_{ki}(hdist(x_i, x_k) + ahr(x_k))}{\sum P_{ki}} \quad \text{if} \quad \sum P_{ki} > 0, 0 \quad \text{otherwise} \tag{3}$$

where $hdist(x_i, x_k)$ gives the horizontal distance from grid cell x_i to upslope neighbor x_k, accounting for whether the cells are adjacent or diagonal neighbors. The notation $\{k{:}P_{ki} > 0\}$ indicates the set of neighbors, k, that have a proportion of their flow contributing to grid cell i. The minimization or maximization is over this set. These functions are recursive because they depend on the value at an upslope neighbor, x_k. The terminal condition for these recursions is that ridge grid cells that have no contribution from upslope (i.e., $\sum P_{ki} = 0$) are assigned a distance value of 0.

3.2.1.2. Horizontal Distance to Stream (*hs)

[20] The horizontal distance to stream is calculated tracing downslope from a grid cell to a stream on the basis of the D∞ flow model. There are again three variants for this: the longest (*lhs*), shortest (*shs*) and average (*ahs*) horizontal flow distance to the stream. Numerically, these are evaluated recursively as follows:

$$lhs(x_i) = \underset{\{k:P_{ik}>0\}}{\text{Max}} \left(hdist(x_i, x_k) + lhs(x_k)\right) \tag{4}$$

$$shs(x_i) = \underset{\{k:P_{ik}>0\}}{\text{Min}} \left(hdist(x_i, x_k) + shs(x_k)\right) \tag{5}$$

$$ahs(x_i) = \sum P_{ik}(hdist(x_i, x_k) + ahs(x_k)) / \sum_{\{k:ahs(x_k)\geq 0\}} P_{ik} \tag{6}$$

In this case the recursions are downslope, because they traverse grid cells downslope terminating at grid cells on a stream raster grid for which $lhs(x_k)$, $shs(x_k)$ and $ahs(x_k)$ are initialized to 0. In evaluating these distance to stream functions

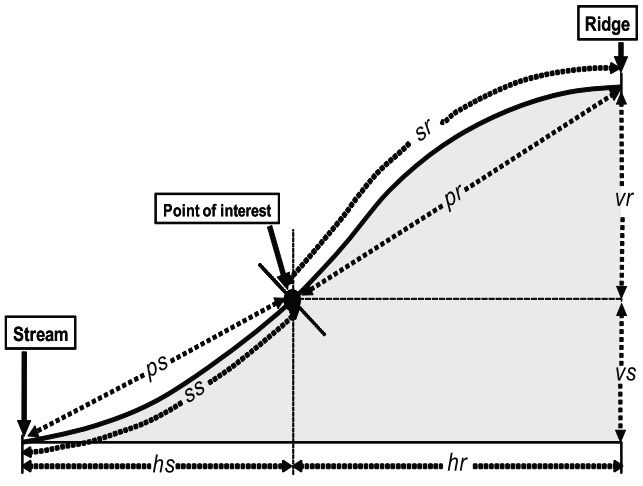

Figure 3. Definitions of some derived topographic attributes.

the stream raster grid was determined using TauDEM (http://hydrology.usu.edu/taudem/) with drainage area threshold of 100 5 × 5 m grid cells.

3.2.1.3. Vertical Rise to Ridge (*vr)

[21] The longest (*lvr*), shortest (*svr*) and average (*avr*) vertical rise to ridge from any grid cell x_i is defined by tracing upslope from a grid cell completely analogously to horizontal distance to ridge calculations on the basis of the D∞ flow model, but instead using elevation differences $z_k - z_i$ in place of the $hdist(x_i, x_k)$ function in (1 to 3). Specifically, $hdist(x_i, x_k)$ is replaced by

$$vdist(x_i, x_k) = z_k - z_i \tag{7}$$

3.2.1.4. Vertical Drop to Stream (*vs)

[22] Similarly, the longest (*lvs*), shortest (*svs*) and average (*avs*) vertical drop to stream from any grid cell x_i is calculated tracing downslope from a grid cell completely analogously to horizontal distance to stream calculations on the basis of the D∞ flow model, but instead using elevation differences $z_i - z_k$ in place of the $hdist(x_i, x_k)$ function in (4 to 6). Specifically, $hdist(x_i, x_k)$ is replaced by

$$vdist(x_i, x_k) = z_i - z_k \tag{8}$$

3.2.1.5. Surface Distance to Ridge (*sr)

[23] The surface distance is defined as the flow distance along the slope (Figure 3). The surface distance between grid cells is given by

$$sdist(x_i, x_k) = \frac{hdist(x_i, x_k)}{\cos(\text{atan}(slp\_t))} \tag{9}$$

where *slp_t* is the slope (recorded as drop/distance or tan) computed on the basis of the D∞ flow model. Longest (*lsr*), shortest (*ssr*) and average (*asr*) surface flow distances to ridge from any grid cell x_i are calculated by using $sdist(x_i, x_k)$ rather than $hdist(x_i, x_k)$ in equations (1)–(3).

3.2.1.6. Surface Distance to Stream (*ss)

[24] Similarly surface distance to stream from any grid cell x_i is calculated on the basis of the D∞ flow model by using $sdist(x_i, x_k)$ rather than $hdist(x_i, x_k)$ in equations (4)–(6). Here *lss*, *sss* and *ass* are used to denote the longest, shortest and average surface distances to the stream.

3.2.1.7. Pythagoras Distances to Stream (*ps)

[25] The Pythagoras distance is defined by considering both vertical and horizontal flow distances along the full length of a hillslope (Figure 3), and combining them using Pythagoras' theorem. We define the following Pythagoras distances on the basis of the different variants of vertical and horizontal flow distances defined above.

Longest Pythagoras distance to stream

$$lps = \sqrt{lhs^2 + lvs^2} \tag{10}$$

Shortest Pythagoras distance to stream

$$sps = \sqrt{shs^2 + svs^2} \tag{11}$$

Average Pythagoras distance to stream

$$aps = \sqrt{ahs^2 + avs^2} \tag{12}$$

3.2.1.8. Pythagoras Distance to Ridge (*pr)

[26] Similarly three variants of Pythagoras distances to the ridge are defined as follows:

Longest Pythagoras distance to ridge

$$lpr = \sqrt{lhr^2 + lvr^2} \tag{13}$$

Shortest Pythagoras distance to ridge

$$spr = \sqrt{shr^2 + svr^2} \tag{14}$$

Average Pythagoras distance to ridge

$$apr = \sqrt{ahr^2 + avr^2} \tag{15}$$

3.2.1.9. Slope Position (*sp)

[27] The relative position of a point on a hillslope can be defined on the basis of the distance to the stream compared to the total length of the hillslope from the distance to the ridge plus the distance to the stream. Given the several variants on distances to the stream and ridge, both horizontal and vertical, we define a number of slope position variants as follows:

D8 horizontal slope position

$$sph8 = \frac{stdist}{stdist + plen} \tag{16}$$

where *plen* is the D8 longest upslope length (Table 1).

D∞ longest horizontal slope position

$$lsph = \frac{lhs}{lhr + lhs} \tag{17}$$

D∞ shortest horizontal slope position

$$ssph = \frac{shs}{shr + shs} \tag{18}$$

D∞ average horizontal slope position

$$asph = \frac{ahs}{ahr + ahs} \tag{19}$$

D∞ longest vertical slope position

$$lspv = \frac{lvs}{lvr + lvs} \tag{20}$$

D∞ shortest vertical slope position

$$sspv = \frac{svs}{svr + svs} \tag{21}$$

D∞ average vertical slope position

$$aspv = \frac{avs}{avr + avs} \tag{22}$$

Table 2. Explanatory Variables Derived From Landsat TM Image

Symbol	Description
lc	Land cover map derived using supervised classification in ERDAS IMAGINE. Land cover is represented as a numerical value encoded as follows: 1 road, rock outcrop and bare; 2 grass; 3 mixed grass and shrub; 4 shrub, riparian and deciduous forest; 5 coniferous forest
pc1	First principal component from ERDAS IMAGINE principal component analysis function
pc2	Second principal component from ERDAS IMAGINE principal component analysis function
pc3	Third principal component from ERDAS IMAGINE principal component analysis function
tc1	First tasseled cap component from ERDAS IMAGINE tasseled cap transformation function (represents brightness)
tc2	Second tasseled cap component from ERDAS IMAGINE tasseled cap transformation function (represents greenness)
tc3	Third tasseled cap component from ERDAS IMAGINE tasseled cap transformation function (represents wetness)
ndvi	Normalized difference vegetation index from ERDAS IMAGINE NDVI function
vi	Vegetation index from ERDAS IMAGINE vegetation index function
cc	Canopy cover index calculated following *Zhu and Band* [1994]

Longest Pythagoras slope position

$$lspp = \frac{lps}{lpr + lps} \tag{23}$$

Shortest Pythagoras slope position

$$sspp = \frac{sps}{spr + sps} \tag{24}$$

Average Pythagoras slope position

$$aspp = \frac{aps}{apr + aps} \tag{25}$$

Slope position varies from 0 at the stream to 1 at the ridge, providing a measure of how far up the slope a point is.

3.2.1.10. Slope Position Ratio (**spra*)

[28] Slope position ratio is the ratio of vertical slope position to horizontal slope position. It can be defined as longest, shortest and average depending on the type of the slope position used in its calculation.

Longest slope position ratio

$$lspr = \frac{lspv}{lsph + lspv} \tag{26}$$

Shortest slope position ratio

$$sspr = \frac{sspv}{ssph + sspv} \tag{27}$$

Average slope position ratio

$$aspr = \frac{aspv}{asph + aspv} \tag{28}$$

[29] Slope position ratio is bound between 0 and 1 and provides an indication of the curvature of the slope. Slope position ratio greater than 0.5 occurs when the vertical slope position is greater than the horizontal slope position, meaning that a point is further up the slope in a vertical sense than horizontal sense as occurs when the slope is convex. Points on a concave slope that fall below the straight line from ridge to stream will have slope position ratio less than 0.5.

3.2.1.11. Modeled Field Curvature (mod_curv)

[30] In our preliminary work, the field observed curvature, encoded as −1, 0, 1 for concave, intermediate and convex respectively, had some explanatory capability for soil depth. This is of limited practical use because field observed curvature is not available at unsampled locations where we want to predict soil depth. To obtain a quantity that captures similar information, but is based only on explanatory variables available for use in prediction we used stepwise regression to model field observed curvature as a function of other explanatory variables. The result from this process, designated as modeled field curvature (*modcurv*) is a continuous (as opposed to discrete −1, 0, 1) variable given by

$$modcurv = 0.055 + 0.115^{*}plncurv - 0.320^{*}sph8 + 8.150^{*}sar - 0.030^{*}gncurv \tag{29}$$

where, *plncurv*, *sph*8, *sar* and *gncurv* denote plan curvature, D8 horizontal slope position, wetness index inverse and general curvature respectively (Table 1). This continuous quantity is used as a surrogate for discrete field observed curvature and was taken as an explanatory variable alongside the other DEM derived variables in the statistical model development.

3.2.2. Data Derived From Remote Sensing Images

[31] After georeferencing and rectification, the Landsat TM image of June 2001 path 41 row 30 was used to derive various land cover attributes (Table 2) that are potentially important for modeling soil depth. Six Landsat TM bands (1, 2, 3, 4, 5, and 7) were used as input information.

[32] A thematic map of land cover (*lc*) was created through supervised classification of the Landsat image. The aerial photograph was used to select training sites where the field observed land cover types (road, rock outcrop and bare area; grasses; mixed grasses and shrubs; shrubs riparian and deciduous forests; and coniferous forests) were identified and used in the ERDAS IMAGINE supervised classification algorithm [*ERDAS*, 1997] to produce land cover classes.

[33] Principal component analysis [*Jensen*, 1996] was used to identify orthogonal components from the six Landsat input bands that explain significant variance. The first three components that explained 99% of the variance were retained as land cover attributes (*pc*1, *pc*2, *pc*3).

[34] The tasseled cap transformation [*Kauth and Thomas*, 1976; *Crist and Cicone*, 1984] was used to convert the six Landsat TM bands (1, 2, 3, 4, 5, and 7) into three components (*tc*1, *tc*2, *tc*3) designated as brightness, greenness and wetness. These are weighted linear combinations of the TM bands:

$$tci = a1{*}tm1 + a2{*}tm2 + a3{*}tm3 + a4{*}tm4 + a5{*}tm5 + a7{*}tm7 \quad (30)$$

where each tasseled cap component (*tci*, i = 1, 2, or 3) is evaluated using coefficients *aj* for each Landsat TM band, *j*, that were derived by *Crist and Cicone* [1984] from empirical observation.

[35] The normalized difference vegetation index (*ndvi*) [*Jensen*, 1996], vegetation index (*vi*) [*Jensen*, 1996] and canopy cover (*cc*) [*Zhu and Band*, 1994] were derived using the following equations:

$$ndvi = \frac{tm4 - tm3}{tm4 + tm3} \quad (31)$$

$$vi = tm4 - tm3 \quad (32)$$

$$cc = 100\left(1 - \frac{tm5 - tm5_{\min}}{tm5_{\max} - tm5_{\min}}\right) \quad (33)$$

In the last equation Landsat thematic mapper (TM) band *tm*5 is the middle infrared radiance TM band and subscripts min and max designate the lowest and highest values of this in the image.

3.3. Statistical Analysis

3.3.1. Normalization

[36] Box-Cox transformations [*Sakia*, 1992] were used to transform the measured soil depth (*sd*) and each explanatory variable so that their distribution was near normal:

$$t(x) = \frac{(x^{\lambda} - 1)}{\lambda} \quad (34)$$

Here, $t(x)$ denotes the transform of variable x with transformation parameter λ. λ was selected to maximize the Shapiro-Wilks normality test W statistic as implemented in R [*Shapiro and Wilk*, 1965; *R Development Core Team*, 2007]. Normalized variables were used in all the statistical modeling works in this paper.

3.3.2. Models

[37] We applied two types of prediction methods: Generalized Additive Models (GAM) [*Hastie and Tibshirani*, 1990] and Random Forests (RF) [*Breiman*, 2001] to predict soil depth using the explanatory variables (Tables 1 and 2).

[38] GAM [*Hastie and Tibshirani*, 1990] is a statistical approach that generalizes multiple regression by replacing linear combinations of the explanatory variables with combinations of nonparametric smoothing or fitting functions, estimated through a back-fitting algorithm. The GAM model is:

$$E(sd|x_1, x_2, \ldots, x_p) = \alpha + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) \quad (35)$$

where, $x_1, x_2, \ldots, x_p$ are explanatory variables (predictors), *sd* is soil depth (response variable) and f_i are nonparametric smoothing splines that relate *sd* to the $x_1, x_2, \ldots, x_p$. The model assumes that the mean of *sd* is an additive combination of nonlinear functions of the explanatory variables $x_1, x_2, \ldots, x_p$. We used the GAM package [*Hastie*, 2008] as implemented in R [*R Development Core Team*, 2007].

[39] Random Forests (RF) is a statistical classification and regression model that combines many classification and regression trees [*Breiman*, 2001]. Each tree is built from a bootstrap sample drawn from the training data set with replacement. We used the Random Forests package [*Liaw and Wiener*, 2002] in the R software [*R Development Core Team*, 2007] to develop RF prediction models. The three main aspects of RF are as follows: (1) from the training set containing s points, s points are sampled with replacement to build a regression tree; (2) among k explanatory variables, m $\ll$ k is specified so that at each node m variables are randomly sampled and the best split among them is identified; and (3) each tree is grown until the specified minimum terminal node size is reached. Steps 1, 2, and 3 are followed to construct n trees. Each tree provides a prediction for any new data point and the random forest predictor is formed by taking the average over the n trees.

[40] In applying this model we used m = *k*/3 and n = 500. The R default value of 5 was used for the minimum terminal node size.

3.3.3. Variable Selection and Model Complexity

[41] Questions in developing a predictive regression model include which potential explanatory variables to use and what to do about interdependent explanatory variables. Many of our explanatory variables are variants on similar quantities, so we are specifically concerned about the effect of this explanatory variable correlation on model prediction error. *Breiman* [2001] indicates that in the RF model correlated explanatory variables can contribute to high prediction error. A matrix giving the cross correlation between all 65 explanatory variables was computed using all 819 data points in the calibration data set to assess the interdependence between explanatory variables. The RF algorithm provides a measure of variable importance, that we used in conjunction with the correlation between explanatory variables to identify models with varying complexity.

[42] The RF measure of importance is determined as follows [*Liaw and Wiener*, 2002; *R Development Core Team*, 2007]. For each tree, the mean square error (MSE) on the out-of-bag portion of the data is recorded. Then the same is done after permuting each explanatory variable. The difference between the two accuracies are then averaged over all trees, and normalized by the standard error. The RF model was run using all 819 data points in the calibration data set with all 65 potential explanatory variables, and soil depth as the response variable. Because of randomness in the RF method the importance varies slightly each time it is run. We therefore ran the RF model 50 times and averaged variable importance across these runs. Explanatory variables were then ordered on the basis of their importance measures.

[43] The number of explanatory variables in a model is a measure of model complexity. We used the correlation matrix, together with the RF importance values to develop

Table 3. Illustrative Correlation Values

	X1	X2	X3
X2	0.7		
X3	0.5	0.6	
X4	0.3	0.2	0.1

sets of explanatory variables representing models of differing complexity by eliminating the variable of lesser importance from pairs of variables with correlation above a designated threshold. Variables were filtered out working sequentially from high to low correlation until no pairs with correlation greater than the threshold remained. So if for example the correlation between 4 explanatory variables is as shown in Table 3 and the threshold in effect is 0.4, first the variable pair (X1, X2) with cross correlation of 0.7 would be identified and the variable with lesser importance from this pair eliminated. Let's say this is variable X2, so that X1, X3 and X4 remain. The next highest correlation pair is (X1, X3) with correlation of 0.5. Note that the higher correlation of 0.6 between variables X2 and X3 is not next because variable X2 was already eliminated. Suppose that of the pair (X1, X3) that X1 has lesser importance. It is eliminated leaving behind variables X3 and X4. The correlation between these is less than the threshold, so both variables are retained and the model for this correlation threshold comprises two explanatory variables, X3 and X4. Lower thresholds result in fewer variables, so a range of models with differing complexity were developed using thresholds ranging from 0.15 to 0.9 in increments of 0.05. This approach reduced the correlation between variables selected for inclusion in a model. Models of differing complexity were also constructed using explanatory variables directly from the variable list ordered by importance.

[44] To evaluate appropriate model complexity, we randomly split our calibration sample of 819 data points into two parts, designated as the training and validation sets as illustrated in Figure 4. The separate testing data set of 130 points more broadly distributed in the watershed was withheld from this process, so that it could be used for evaluation of the final models.

[45] Both GAM and RF models were applied, using the training data set of 614 data points to fit the models. Prediction error was computed for both the training and validation data set. The validation data set prediction error provided an out of sample estimate appropriate for trading off variance due to complexity with bias due to too few explanatory variables [see, e.g., *Hastie et al.*, 2001]. The results from this analysis allowed us to select the explanatory variables and degree of model complexity. This was done without and with the new topographic variables derived in this research to evaluate the contribution of the new variables in predicting soil depth. The new variables are indicated with an asterisk in Table 1.

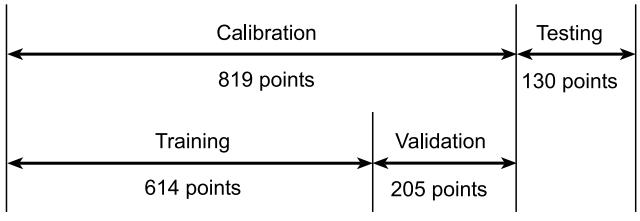

Figure 4. Division of data into training, validation, and testing sets.

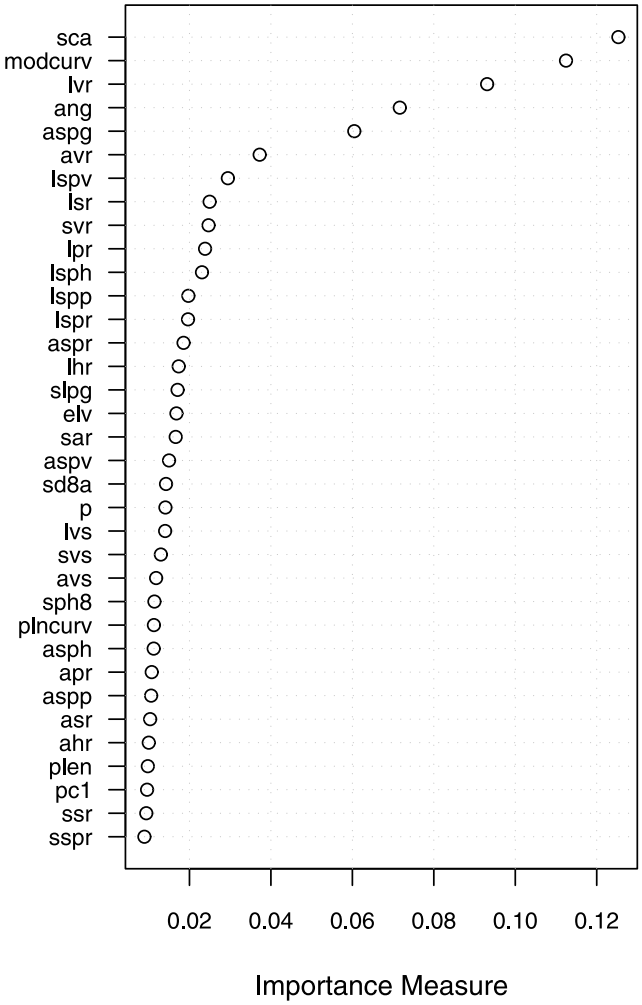

Figure 5. Variable importance measure of the Box-Cox transformed explanatory variables averaged from 50 RF model runs.

3.3.4. Testing

[46] Once the explanatory variables and models with appropriate complexity had been selected, they were applied using the full calibration data set as input. Both RF and GAM models were used to predict soil depth for the entire watershed with and without the new topographic variables. We then compared the testing data set with the model soil depth values at testing locations using the Nash-Sutcliffe efficiency coefficient (*NSE*) [*Nash and Sutcliffe*, 1970]:

$$NSE = 1 - \frac{\sum \left(SD_o - SD_p\right)^2}{\sum \left(SD_o - SD_m\right)^2} \tag{36}$$

where SD_o, SD_p, and SD_m are observed (measured), predicted, and mean of observed (measured) soil depths respectively. *NSE* is a normalized model performance measure that compares the mean square error generated by a particular

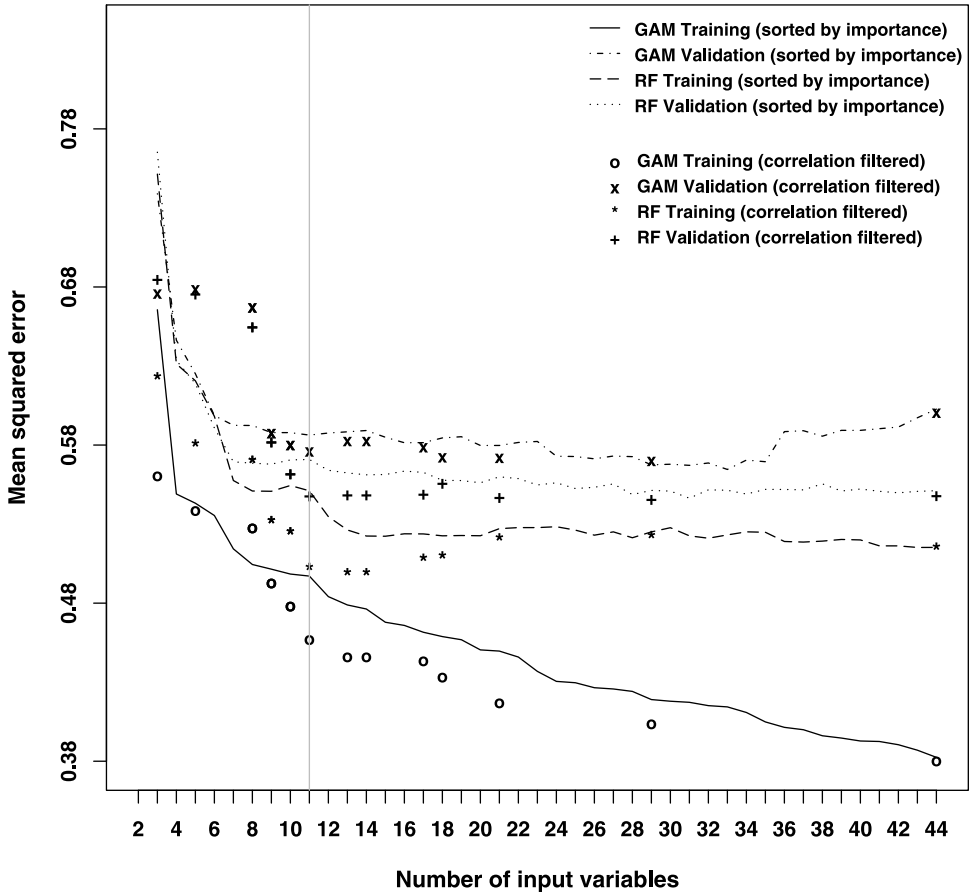

Figure 6. Number of input variables (model complexity) versus mean square error with explanatory variables selected directly using importance (solid curve) and filtered by correlation (symbols) from all candidate explanatory variables in Tables 1 and 2 (new variables included).

model to the variance of the observations [*Schaefli and Gupta*, 2007].

3.3.5. Comparison of Predicted and SSURGO Soil Depths

[47] A shape file of soil depth was developed from the SSURGO soil database as an average of the soil components within a soil mapping unit (the spatial resolution of SSURGO soil database). The generalization present in SSURGO soil maps limits their applicability at a point scale. Therefore we also aggregated the observed and predicted soil depth values within each SSURGO soil mapping unit and compared average observed, model predicted, and SSURGO soil depth values at this spatial scale using *NSE*.

4. Results

4.1. Variable Selection and Model Complexity

[48] Figure 5 shows explanatory variables with importance values greater than or equal to 0.009, ordered on the basis of their average importance values from 50 RF runs with all 819 calibration data points and all 65 explanatory variables. Figure 5 suggests that the variables *sca*, *modcurv*, *lvr* and *ang* are the four most important explanatory variables in predicting soil depth.

[49] Figure 6 shows the variation of mean square prediction error for training and validation data sets versus model complexity in terms of the number of input variables including the new topographic variables. The continuous lines in Figure 6 are from models developed using explanatory variables selected on the basis of RF importance only. There is a new model for each additional input variable. Both GAM and RF models were evaluated and Figure 6

Table 4. Groups of Explanatory Variables Created on the Basis of the Variable Importance and the Correlation Between Explanatory Variables

Group	Number of Variables	Correlation Coefficient Threshold
1	3	0.15
2	5	0.2
3	8	0.25
4	8	0.3
5	9	0.35
6	9	0.4
7	10	0.45
8	10	0.5
9	12	0.55
10	13	0.6
11	14	0.65
12	17	0.7
13	18	0.75
14	21	0.8
15	29	0.85
16	44	0.9

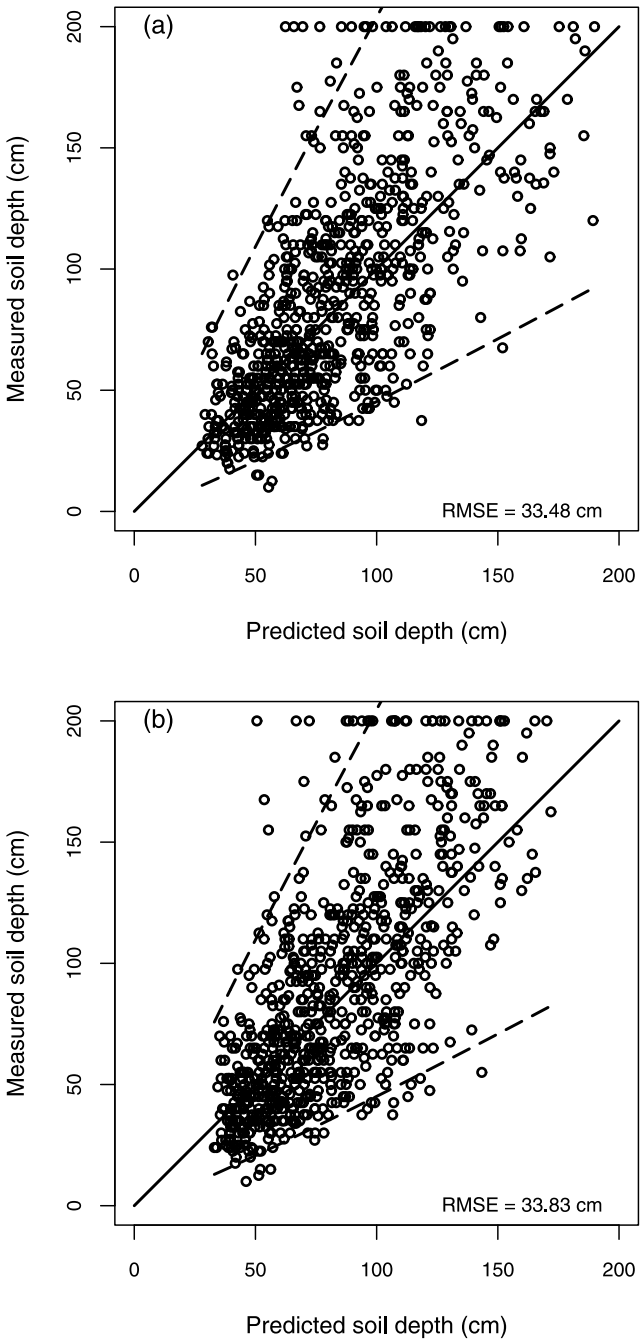

Figure 7. Predicted soil depth versus measured soil depth with ±2 standard error for (a) GAM and (b) RF calibration.

reports training and validation errors separately. The symbols in Figure 6 are from models developed using cross correlation as a filter to reduce interdependence among explanatory variables. Table 4 gives the number of variables in each group selected in this way.

[50] In Figure 6 for both the importance selected and correlation filtered models the training error of GAM decreases progressively for each additional input variable, while the validation error decreases initially and as the model complexity continues to increase further it starts to increase. For RF, both the training and validation errors decrease initially and as model complexity continues to increase they become essentially constant. This is consistent with RF being robust against overfitting. For both GAM and RF models the use of correlation filtered explanatory variables resulted in lower error. While training errors are smaller for GAM, the out of sample validation error from the RF model is less than from the GAM model. The least validation error for the RF model occurred with 11 correlation filtered input variables. Similarly, the validation error for GAM increases for complexity greater than 11 correlation filtered variables, although validation mean square error (MSE) at 18 and 21 input variables fluctuates slightly below the 11 input variable MSE. Nevertheless, in our

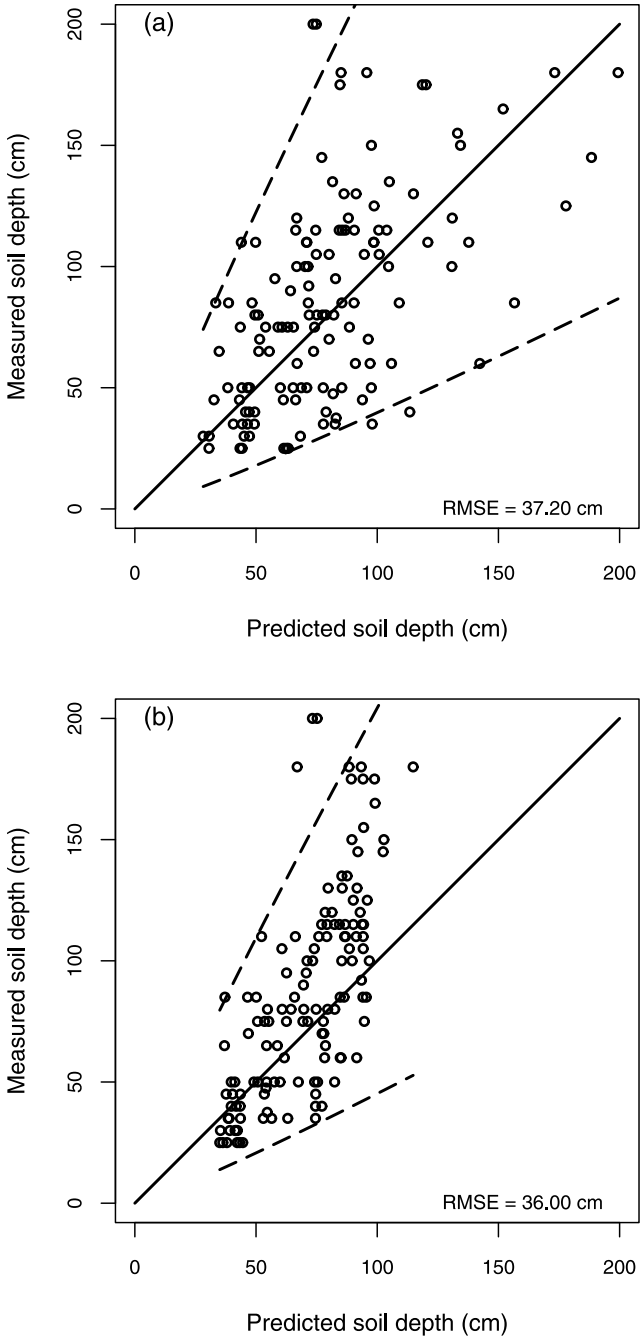

Figure 8. Predicted soil depth versus measured soil depth with ±2 standard error for (a) GAM and (b) RF testing.

Table 5. Comparison of NSE Values of RF and GAM Predicted Soil Depths for out of Sample Testing Data Set

Model Testing	NSE
GAM results without the new topographic variables	0.26
GAM results with the new topographic variables	0.47
RF results without the new topographic variables	0.31
RF results with the new topographic variables	0.52

judgment the point of diminishing returns has been reached at 11 input variables for both the RF and GAM models. Consequently we selected 11 correlation filtered explanatory variables as representing the optimum complexity for this data set: *sca*, *modcurv*, *ang*, *avr*, *lspv*, *slpg*, *elv*, *sd*8*a*, *lvs*, and *plncurv* from Table 1 and *pc*1 from Table 2. Except for *pc*1, which is a land cover attribute these are all topographic attributes. The similar analysis without the new topographic variables resulted in lowest validation error with 7 correlation filtered input variables: *sca*, *aspg*, *slpg*, *sd*8*a*, *elv*, *p* from Table 1 and *pc*2 Table 2.

4.2. Model Evaluation

[51] On the basis of the selection of 11 correlation filtered explanatory variables above, including new topographic variables, RF and GAM models were developed using these variables with the full calibration set of 819 data points. Figure 7 shows the scatterplots of GAM (Figure 7a) and RF (Figure 7b) predicted versus the measured soil depth respectively, for the calibration data. Here the results have been transformed back into regular soil depth quantities. In Figure 7 the diagonal (central) lines represent the 1:1 line (predicted = observed). The two diverging dash lines, above and below the 1:1 line, show the predicted soil depth ±2 standard errors representing 95 percent confidence intervals. These lines diverge as a result of the Box-Cox back transformation. Figure 8 shows similar scatterplots for the testing data that was not used in model development.

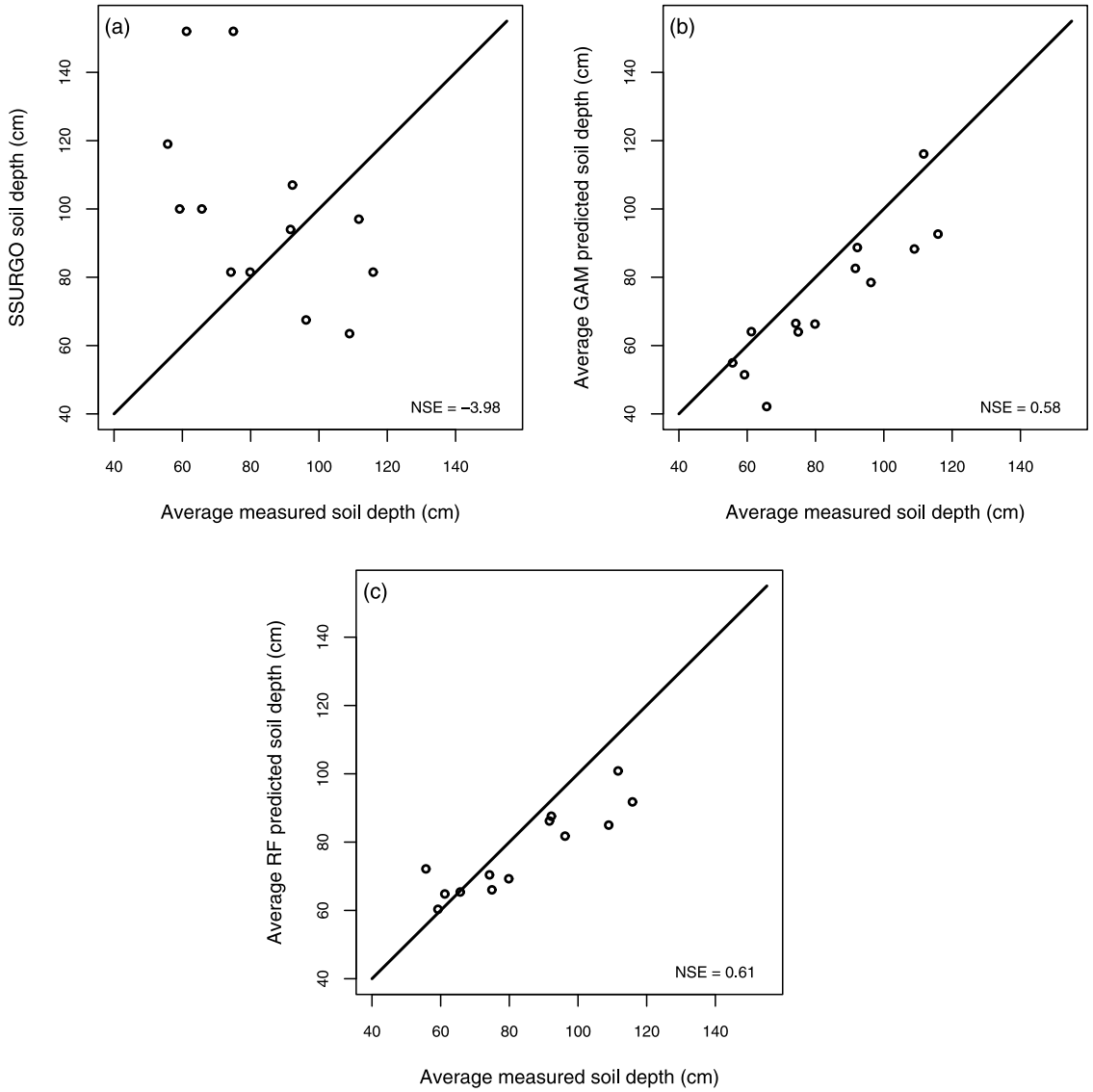

Figure 9. Plots showing (a) SSURGO map unit soil depths versus measured soil depth averaged in each map unit, (b) GAM predicted soil depth versus measured soil depth averaged in each map unit, and (c) RF predicted soil depth versus measured soil depth averaged in each map unit.

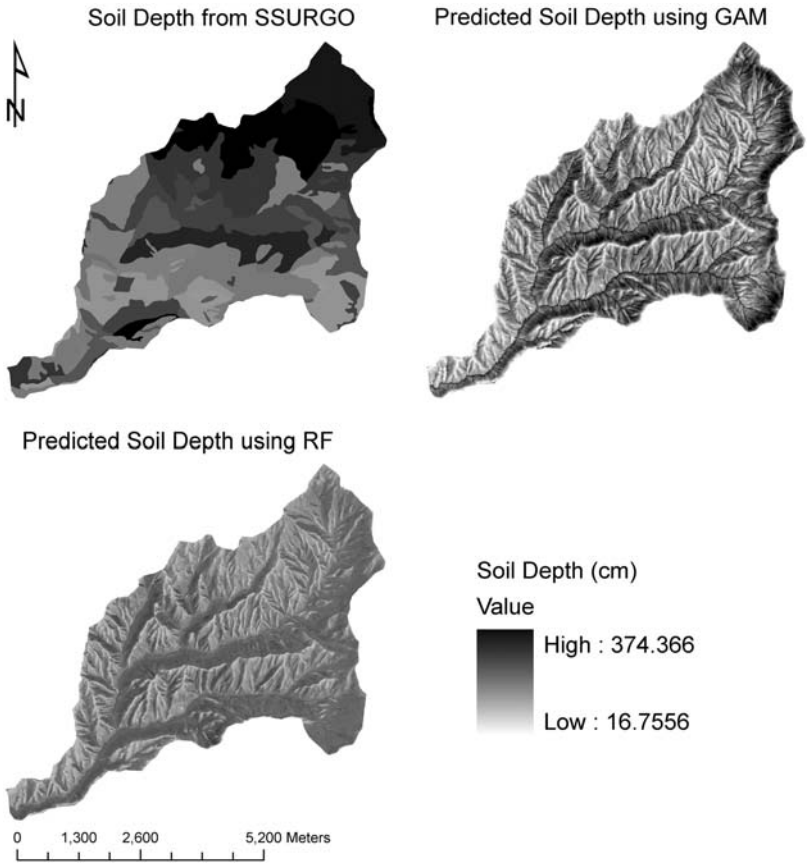

Figure 10. Comparison of maps of soil depths (top left) from SSURGO soil database, (top right) predicted with GAM, and (bottom) predicted with RF.

[52] RF and GAM models were also developed with the full calibration set of 819 data points using the 7 correlation filtered explanatory variables identified above that did not include new topographic variables. Table 5 shows testing data *NSE* values for GAM and RF models with (11 explanatory variables) and without (7 explanatory variables) the new topographic variables.

[53] Examining the differences between models with and without new topographic variables, the *NSE* values in Table 5 indicate that soil depth prediction using both models showed significant improvement due to the new topographic variables. The fraction of variability explained increases from 26% to 47% for GAM and 31% to 52% for RF when the new topographic variables are included. This represents close to a doubling in explained variability.

4.3. SSURGO Map Unit Scale Comparisons

[54] We aggregated the GAM and RF predicted and observed soil depths to a scale of SSURGO map units to compare the GAM and RF model predicted soil depths with SSURGO and observed soil depth at a consistent scale. Figure 9 shows the scatterplots of the SSURGO (Figure 9a), GAM (Figure 9b) and RF (Figure 9c) predicted soil depths versus measured soil depths aggregated over the SSURGO soil map units. Figure 9 also indicates *NSE* values. The SSURGO soil depth (Figure 9a) appears unrelated to soil depth measurements, with $NSE = -3.98$, even when data is aggregated at the SSURGO map unit scale. By contrast the GAM (Figure 9b) and RF (Figure 9c) models predict the aggregated observed soil depths with $NSE = 0.58$ and $NSE = 0.61$ respectively.

[55] Figure 10 compares the soil depth maps from SSURGO, GAM and RF. The SSURGO soil depth map divides the watershed into map units with abrupt boundaries. The spatial variation of soil depth with the topography is not expressed. The GAM and RF models provide soil depth maps at 5 m grid scale, which predict the variation of the soil depth with the landscape. The soil depth maps from GAM and RF predict that the ridges (convex areas) and south facing slopes have shallower soils as compared to the valleys (concave areas) and the north facing slopes respectively. This generally agrees with observations in this area and existing literature [*Heimsath et al.*, 2002; *Hoover and Hursh*, 1943].

5. Discussion and Conclusions

[56] Statistical models have been developed that predict soil depth over a landscape using topographic and land cover attributes. The variables identified as predictors included ten topographic variables: specific catchment area (*sca*), modeled curvature (*modcurv*), D∞ flow direction (*ang*), average rise to ridge (*avr*), longest vertical slope position (*lspv*), longest vertical drop to stream (*lvs*), slope (*slpg*), D8 slope averaged over 100 m downslope distance (*sd8a*), elevation (*elv*), and plan curvature (*plncurv*), and one land cover variable, the first component of principal component transformation of Landsat TM imagery. Thus,

topographic variables were found to be generally more important than the land cover variables in predicting soil depth for this data set.

[57] The topographic variables used included new digital elevation model derived variables such as rise to ridge, drop to stream, distances to ridge and stream, vertical and horizontal slope positions, and slope position ratios. The fraction of variability explained by both GAM and RF predictions was increased by about 20% because of the combined effect of the following new topographic variables that were selected as explanatory variables: modeled curvature (*modcurv*), D∞ flow direction (*ang*), average rise to ridge (*avr*), longest vertical slope position (*lspv*), and longest vertical drop to stream (*lvs*).

[58] These new topographic variables represent an important contribution to the science of modeling soil depth based on topographic information. In considering the physical basis of these variables as predictors of soil depth, the selection of specific catchment area (*sca*), modeled (*modcurv*) and plan curvature (*plncurv*) are consistent with literature that suggests that deeper soils occur in areas that are concave [*Heimsath et al.*, 2002; *Hoover and Hursh*, 1943]. Vertical slope position (*lspv*), vertical drop to stream (*lvs*) and average rise to ridge (*avr*), quantify position on a hillslope as predictive of soil depth. Slope variables (*slpg* and *sd*8*a*) quantify relationships between soil depth and slope. The appearance of absolute quantities, elevation (*elv*) and downslope angle (*ang*) as predictors, is a bit difficult to justify physically, where we would prefer more transferable relative variables such as the slope position quantities. We suspect that elevation is representing some slope position effects perhaps combined with climate, while downslope angle, which is measured counter clockwise from east discriminates between north and south facing slopes, related to local microclimate and consistent with observations of deeper soils on north facing slopes. The land cover principal component variable (*pc*1) quantifies the role played by land cover.

[59] Both GAM and RF modeled soil depths were able to explain about 50% of the measured soil depth variability in an out of sample test. Considering the uncontrolled uncertainties due to the complex local variation of soil depth, DEM errors and GPS reading errors, this is considered an important improvement toward solving the need for distributed soil depth information in distributed hydroecological modeling. A strength of this work is that the models were developed and validated against a comprehensive data set of measured soil depths. These models, which draw upon new topographic information and are based on comprehensive data, contribute to the scientific quantification of soil depth at a refined spatial scale. This is important for spatially distributed hydroecological models. While the soil depth models developed have specific application to the Dry Creek Experimental Watershed, the physical processes of soil development in Dry Creek are representative of a broad region with similar climate and parent material. Such data-based approaches, while relying on statistical relationships contribute to hydrological science involving soil depth by bringing a measure of objectivity to the approach not present, for example, in cited prior work that relied on expert opinions.

[60] The root-mean-square errors (RMSE) reported in Figure 8 and the NSE in Table 5 indicate that the RF model is slightly better than the GAM model for predicting soil depth at point scale in terms of these out of sample statistical measures. However, there is some indication in Figure 8 that the RF model underestimates the soil depth for deep soils. For the testing data the RF model never predicts soil deeper than about 120 cm, while soil depths up to 200 cm were observed and are predicted by the GAM model. This discrepancy may be due to the discrete nature of regression tree predictors that underlie the RF approach. In choosing between whether to use a GAM or RF model, the inability of the RF model to predict deep soils in the out of sample test leads us to favor the GAM model for spatial predictions of soil depth in Dry Creek Experimental Watershed. Both models, in our judgment provide a significant improvement over using SSURGO data, because we found that the soil depth extracted from the SSURGO soil database was not correlated at all with the observed soil depth when aggregated to the scale of SSURGO mapping units.

[61] The generality and transferability of this work to other areas still remains to be tested. The fact that calibrations using data from within 8 subwatersheds yielded reasonably good predictions for the 130 testing points more broadly distributed in the watershed gives some confidence that this model should hold for the Boise Front. Additional work is needed to test the approach and new explanatory variables in other areas. Additional work is also needed to assess the contribution from using this modeled soil depth information in hydroecological models.

[62] **Acknowledgments.** The data collection part of this research was funded by the Utah Drought Management project, a project of USDA-CSREES special research grant 2005-34552-15828 from the USDA Cooperative State Research, Education, and Extension Service. The modeling part was funded by the Inland Northwest Research Alliance (INRA). Field assistance was provided by Pam Aishlin and other students at Boise State University with funding from the National Science Foundation EPSCoR RII program (grant EPS-0447689).

References

Anderson, R. M., V. I. Koren, and S. M. Reed (2006), Using SSURGO data to improve Sacramento model a priori parameter estimates, *J. Hydrol.*, *320*, 103–116, doi:10.1016/j.jhydrol.2005.07.020.

Bierkens, M. F. P., and P. A. Burrough (1993), The indicator approach to categorical soil data. 2. Application to mapping and land-use suitability analysis, *J. Soil Sci.*, *44*, 369–381.

Breiman, L. (2001), Random forests, *Mach. Learn.*, *45*, 5–32, doi:10.1023/A:1010933404324.

Crist, E. P., and R. C. Cicone (1984), A physically based transformation of thematic mapper data—The TM tasseled cap, *IEEE Trans. Geosci. Remote Sens.*, *22*, 256–263, doi:10.1109/TGRS.1984.350619.

Dietrich, W. E., R. Reiss, M.-L. Hsu, and D. R. Montgomery (1995), A process-based model for colluvial soil depth and shallow landsliding using digital elevation data, *Hydrol. Processes*, *9*, 383–400, doi:10.1002/hyp.3360090311.

ERDAS (1997), *ERDAS Imagine Tour Guide*, Atlanta, Ga.

Freer, J., J. J. McDonnell, K. J. Beven, N. E. Peters, D. A. Burns, R. P. Hooper, B. Aulenbach, and C. Kendall (2002), The role of bedrock topography on subsurface storm flow, *Water Resour. Res.*, *38*(12), 1269, doi:10.1029/2001WR000872.

Gessler, P. E., I. D. Moore, N. J. McKenzie, and P. J. Ryan (1995), Soil-landscape modeling and spatial prediction of soil attributes, *Int. J. Geogr. Inf. Syst.*, *9*, 421–432, doi:10.1080/02693799508902047.

Goodchild, M. F. (1992), Geographical data modeling, *Comput. Geosci.*, *18*, 401–408, doi:10.1016/0098-3004(92)90069-4.

Grayson, R., and G. Blöschl (Eds.) (2000), *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, 432 pp., Cambridge Univ. Press, Cambridge, U. K.

Gribb, M., I. Forkutsa, A. J. Hansen, D. Chandler, and J. McNamara (2009), The effect of various soil hydraulic property estimates on soil

moisture simulations, *Vadose Zone J.*, *8*, 321–331, doi:10.2136/vzj2008.0088.
Hastie, T. (2008), GAM: Generalized additive models, R package version 1.0, R Found. for Stat. Comput. Vienna. (Available at http://www.R-project.org)
Hastie, T., and R. Tibshirani (1990), *Generalized Additive Models*, Chapman and Hall, London.
Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 533 pp., Springer, New York.
Heimsath, A. M., W. E. Dietrich, K. Nishiizumi, and R. C. Finkel (1997), The soil production function and landscape equilibrium, *Nature*, *388*, 358–361, doi:10.1038/41056.
Heimsath, A. M., W. E. Dietrich, K. Nishiizumi, and R. C. Finkel (1999), Cosmogenic nuclides, topography, and the spatial variation of soil depth, *Geomorphology*, *27*, 151–172, doi:10.1016/S0169-555X(98)00095-6.
Heimsath, A. M., J. Chappell, W. E. Dietrich, K. Nishiizumi, and R. C. Finkel (2000), Soil production on a retreating escarpment in southeastern Australia, *Geology*, *28*, 787–790, doi:10.1130/0091-7613(2000)28<787:SPOARE>2.0.CO;2.
Heimsath, A. M., W. E. Dietrich, K. Nishiizumi, and R. C. Finkel (2001), Stochastic processes of soil production and transport: Erosion rates, topographic variation and cosmogenic nuclides in the Oregon Coast Range, *Earth Surf. Processes Landforms*, *26*, 531–552, doi:10.1002/esp.209.
Heimsath, A. M., J. Chappell, N. A. Spooner, and D. G. Questiaux (2002), Creeping soil, *Geology*, *30*, 111–114, doi:10.1130/0091-7613(2002)030<0111:CS>2.0.CO;2.
Heimsath, A. M., D. J. Furbish, and W. E. Dietrich (2005), The illusion of diffusion: Field evidence for depth-dependent sediment transport, *Geology*, *33*, 949–952, doi:10.1130/G21868.1.
Henderson-Sellers, A., and P. J. Robinson (1986), *Contemporary Climatology*, John Wiley, New York.
Hoover, M. D., and C. R. Hursh (1943), Influence of topography and soil depth on runoff from forest land, *Eos Trans. AGU*, *24*, 6.
Jenny, H. (1941), *Factors of Soil Formation: A Quantitative System in Pedology*, 281 pp., McGraw-Hill, New York.
Jensen, J. R. (1996), *Introductory to Digital Image Processing: A Remote Sensing Perspective*, 2nd ed., Prentice-Hall, Englewood Cliffs, N. J.
Kauth, R. J., and G. S. Thomas (1976), The tasseled cap—A graphic description of the temporal development of agricultural crops as seen in Landsat, paper presented at Symposium on Machine Processing of Remotely Sensed Data, Purdue Univ., West Lafayette, Indiana.
Liaw, A., and M. Wiener (2002), Classification and regression by randomForest, *R News*, *2*, 18–22.
Lin, H., J. Bouma, Y. Pachepsky, A. Western, J. Thompson, R. van Genuchten, H.-J. Vogel, and A. Lilly (2006), Hydropedology: Synergistic integration of pedology and hydrology, *Water Resour. Res.*, *42*, W05301, doi:10.1029/2005WR004085.
Mark, D. M. (1988), Network models in geomorphology, in *Modelling in Geomorphological Systems*, edited by M. G. Anderson, pp. 73–97, John Wiley, New York.
Mark, D. M., and F. Csillag (1989), The nature of boundaries on 'area-class' maps, *Cartographica*, *27*, 56–78.
McBratney, A. B., and I. O. A. Odeh (1997), Application of fuzzy sets in soil science: Fuzzy logic, fuzzy measurements and fuzzy decisions, *Geoderma*, *77*, 85–113, doi:10.1016/S0016-7061(97)00017-7.
McBratney, A. B., M. L. M. Santos, and B. Minasny (2003), On digital soil mapping, *Geoderma*, *117*, 3–52, doi:10.1016/S0016-7061(03)00223-4.
McNamara, J. P., D. Chandler, M. Seyfried, and S. Achet (2005), Soil moisture states, lateral flow, and streamflow generation in a semi-arid, snowmelt-driven catchment, *Hydrol. Processes*, *19*, 4023–4038, doi:10.1002/hyp.5869.
Moore, I. D., R. B. Grayson, and A. R. Ladson (1991), Digital terrain modelling: A review of hydrological, geomorphological, and biological applications, *Hydrol. Processes*, *5*, 3–30, doi:10.1002/hyp.3360050103.
Moore, I. D., P. E. Gessler, G. A. Nielsen, and G. A. Peterson (1993), Soil attribute prediction using terrain analysis, *Soil Sci. Soc. Am. J.*, *57*, 443–452.
Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models, *J. Hydrol.*, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.
O'Callaghan, J. F., and D. M. Mark (1984), The extraction of drainage networks from digital elevation data, *Comput. Vision Graphics Image Process.*, *28*, 328–344.
Odeh, I. O. A., A. B. McBratney, and D. J. Chittleborough (1994), Spatial prediction of soil properties from landform attributes derived from a digital elevation model, *Geoderma*, *63*, 197–214, doi:10.1016/0016-7061(94)90063-9.
R Development Core Team (2007), R: A language and environment for statistical computing, 2.5.0, *Rep. 3-900051-07-0*, R Found. for Stat. Comput., Vienna. (Available at http://www.R-project.org)
Saco, P. M., G. R. Willgoose, and G. R. Hancock (2006), Spatial organization of soil depths using a landform evolution model, *J. Geophys. Res.*, *111*, F02016, doi:10.1029/2005JF000351.
Sakia, R. M. (1992), The Box-Cox transformation technique: A review, *Statistician*, *41*, 169–178, doi:10.2307/2348250.
Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrol. Processes*, *21*, 2075–2080, doi:10.1002/hyp.6825.
Schmidt, J., P. Tonkin, and A. Hewitt (2005), Quantitative soil-landscape models for the Haldon and Hurunui soil sets, New Zealand, *Aust. J. Soil Res.*, *43*, 127–137, doi:10.1071/SR04074.
Schulz, K., R. Seppelt, E. Zehe, H. J. Vogel, and S. Attinger (2006), Importance of spatial structures in advancing hydrological sciences, *Water Resour. Res.*, *42*, W03S03, doi:10.1029/2005WR004301.
Scull, P., J. Franklin, O. A. Chadwick, and D. McArthur (2003), Predictive soil mapping: A review, *Prog. Phys. Geogr.*, *27*, 171–197, doi:10.1191/0309133303pp366ra.
Seyfried, M. S., L. E. Grant, D. Marks, A. Winstral, and J. McNamara (2009), Simulated soil water storage effects on streamflow generation in a mountainous snowmelt environment, Idaho, USA, *Hydrol. Processes*, *23*, 858–873, doi:10.1002/hyp.7211.
Shapiro, S. S., and M. B. Wilk (1965), An analysis of variance test for normality (complete samples), *Biometrika*, *52*, 591–611.
Stieglitz, M., J. Shaman, J. McNamara, V. Engel, J. Shanley, and G. W. Kling (2003), An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport, *Global Biogeochem. Cycles*, *17*(4), 1105, doi:10.1029/2003GB002041.
Summerfield, M. A. (1997), *Global Geomorphology*, 537 pp. Longman, New York.
Tarboton, D. G. (1997), A new method for the determination of flow directions and contributing areas in grid digital elevation models, *Water Resour. Res.*, *33*, 309–319, doi:10.1029/96WR03137.
Tarboton, D. G., and D. P. Ames (2001), Advances in the mapping of flow networks from digital elevation data, paper presented at World Water and Environmental Resources Congress, Am. Soc. of Civ. Eng., Orlando, Fla., 20–24 May.
Tarboton, D. G., and M. E. Baker (2008), Towards an algebra for terrain-based flow analysis, in *Representing, Modeling and Visualizing the Natural Environment: Innovations in GIS 13*, edited by N. J. Mount et al., pp. 167–194, CRC Press, Boca Raton, Fla.
Williams, C. J. (2005), Characterization of the spatial and temporal controls on soil moisture and streamflow generation in a semi-arid headwater catchment, M.S. thesis, Boise State Univ., Boise, Idaho.
Williams, C. J., J. P. McNamara, and D. G. Chandler (2008), Controls on the temporal and spatial variability of soil moisture in a mountainous landscape: The signatures of snow and complex terrain, *Hydrol. Earth Syst. Sci. Discuss.*, *5*, 1927–1966.
Zhang, R., S. Rahman, G. F. Vance, and L. C. Mann (1995), Geostatistical analyses of trace elements in soils and plants, *Soil Sci.*, *159*, 383–390, doi:10.1097/00010694-199506000-00003.
Zhu, A. X. (1997), A similarity model for representing soil spatial information, *Geoderma*, *77*, 217–242, doi:10.1016/S0016-7061(97)00023-2.
Zhu, A. X., and L. E. Band (1994), A knowledge-based approach to data integration for soil mapping, *Can. J. Remote Sens.*, *20*, 408–418.
Zhu, A. X., and D. S. Mackay (2001), Effects of spatial detail of soil information on watershed modeling, *J. Hydrol.*, *248*, 54–77, doi:10.1016/S0022-1694(01)00390-0.
Zhu, A. X., L. E. Band, B. Dutton, and T. J. Nimlos (1996), Automated soil inference under fuzzy logic, *Ecol. Modell.*, *90*, 123–145, doi:10.1016/0304-3800(95)00161-1.
Zhu, A. X., L. Band, R. Vertessy, and B. Dutton (1997), Derivation of soil properties using a soil land inference model (SoLIM), *Soil Sci. Soc. Am. J.*, *61*, 523–533.

D. G. Chandler, Civil Engineering Department, Kansas State University, 2109 Fiedler Hall, Manhattan, KS 66506, USA.
J. P. McNamara, Department of Geosciences, Boise State University, 1910 University Drive, MG 225208, Boise, ID 83729, USA.
D. G. Tarboton and T. K. Tesfa, Civil and Environmental Engineering Department, Utah State University, 4110 Old Main Hill, Logan, UT 84322, USA. (david.tarboton@usu.edu)