An evaluation of nonlinear methods for estimating catchment-scale soil moisture patterns based on topographic attributes

Michael L. Coleman and Jeffrey D. Niemann

ABSTRACT

Physical processes that impact soil moisture are typically expressed as nonlinear functions, but most previous research on the estimation of soil moisture has relied on linear techniques. In the present work, two machine learning techniques, a spatial artificial neural network (SANN) and a mixture model (MM), that can infer nonlinear relationships are compared to multiple linear regression (MLR) for estimating soil moisture patterns using topographic attributes as predictor variables. The methods are applied to time-domain reflectometry (TDR) soil moisture data collected at three catchments with varying characteristics (Tarrawarra, Satellite Station and Cache la Poudre) under different wetness conditions. The methods' performances with respect to the number of predictor attributes, the quantity of training data and the attributes employed are compared using the Nash–Sutcliffe coefficient of efficiency (NSCE) as the performance measure. The performances of the methods are dependent on the site studied, the average soil moisture and the quantity of training data provided. Although the methods often perform similarly, the best performing method overall is the SANN, which incorporates additional predictor variables more effectively than the other methods.

Key words | cross-validation, mixture model, neural network, nonlinear, soil moisture

Michael L. Coleman
Jeffrey D. Niemann (corresponding author)
Department of Civil and Environmental Engineering,
Campus Delivery 1372,
Colorado State University,
Fort Collins,
CO 80523-1372,
USA
E-mail: *jniemann@engr.colostate.edu*

INTRODUCTION

Soil moisture is an important hydrologic state variable owing to its influence on a variety of surface hydrologic processes and land surface–atmospheric interactions. For example, soil moisture affects both the partitioning of radiation into sensible and latent heat (Entekhabi *et al.* 1996) and the partitioning of rainfall into infiltration and runoff (Dunne & Black 1970). Additionally, soil moisture influences vegetation patterns (Eagleson 1978), land surface erosion processes (Moore *et al.* 1988) and soil development (Hillel 1998). Spatial patterns of soil moisture and their characteristics, such as connectivity of wet areas, are also important for hydrologic considerations (Hewlett & Hibbert 1967; Dunne & Black 1970; Dunne *et al.* 1975; Western *et al.* 2001). Soil moisture is typically observed using remote-sensing techniques or manual techniques, such as time-domain reflectometry (TDR). Unfortunately, neither approach is practical for observing soil moisture patterns within catchments at suitable resolutions (Robinson *et al.* 2003; Vereecken *et al.* 2008). The desire for accurate characterization of moisture patterns coupled with the difficulties in observing the patterns has led to many efforts to estimate moisture patterns (Yates & Warrick 1987; Nyberg 1996; Bardossy & Lehmann 1998; Western *et al.* 1999a; Sulebak *et al.* 2000).

Significant research has been devoted to investigating the correlations between soil moisture and topographic attributes and the effectiveness of multiple linear regression (MLR) for estimating soil moisture using topographic attributes as predictive data. Topographic attributes have been used because organized patterns of soil moisture resemble

doi: 10.2166/hydro.2012.145

patterns of topography and because surface elevation data are readily available for nearly all parts of the world. Zaslavsky & Sinai (1981) explained 81% of soil water variation 2 weeks after rainfall by curvature in an agricultural field near Beer-Sheba, Israel. Moore *et al.* (1988) found that 33% of the soil moisture variation on a transect of a 7.5 ha catchment in Australia could be explained by the wetness index, which is defined as the ratio of the specific contributing area (SCA) and the local slope, and that 41% of the variation could be explained by using both the wetness index and the topographic aspect. Nyberg (1996) explained between 15 and 42% of soil moisture variation by correlations with elevation, slope, wetness index and the logarithm of the contributing area. Western *et al.* (1999a) were able to explain up to 61% of spatial soil moisture variation under relatively wet conditions but only 22% in drier conditions by a combination of potential solar radiation index (PSRI), which is the ratio of the potential insolation of a surface with a particular slope and aspect to a hypothetical horizontal surface at the same location, and either the wetness index or the logarithm of the contributing area. Sulebak *et al.* (2000) found the combination of slope, aspect and profile curvature could explain 70% of moisture variation at two locations in Sweden. Green & Erskine (2004) found the highest correlations between soil moisture and topographic attributes for agricultural fields in Colorado on the wettest date considered and the strongest correlation with slope even though that attribute only explains approximately 20% of the variance. Despite some instances where MLR is effective, a conceptual inconsistency exists in linearly regressing soil moisture on topographic attributes because those attributes are generally associated with physical processes that relate nonlinearly to soil moisture (Rodriguez-Iturbe 2000). Such nonlinearities might produce nonlinearity in the relationships between soil moisture and topographic attributes. For example, Western *et al.* (1999a) noted that scattergraphs indicate a possible nonlinear relationship between the wetness index and soil moisture data from organized patterns in their dataset.

Geostatistical techniques (Journel & Huijbregts 1978; Kitanidis 1993; Chiles & Delfiner 1999) have also been applied to the tasks of soil moisture characterization and estimation. Yates & Warrick (1987) used cokriging with bare soil temperature and percent sand content as ancillary variables to estimate soil moisture in Arizona. They found that, if the ancillary variable is well correlated with soil moisture, then the cokriging estimates are better than estimates from ordinary kriging. Bardossy & Lehmann (1998) estimated soil moisture patterns from a 630 ha catchment in Germany with several geostatistical techniques. They found that Bayes–Markov updating (BMU), a simplified form of Bayes–Markov kriging (Zhu & Journel 1993), has the lowest errors of all the tested methods for the conditions analyzed. BMU can incorporate ancillary data in a nonlinear manner and its performance with either the wetness index or land use as ancillary data is better than both ordinary kriging and external drift kriging. The method performs slightly better with the wetness index than with land use for that dataset. A common assumption made in geostatistical analyses is that of a stationary random field, but previous research has indicated that soil moisture patterns are not random but exhibit spatial organization (Dunne *et al.* 1975; Rodriguez-Iturbe *et al.* 1995; Western *et al.* 1999a).

In addition to those standard methods, other methods have also been developed and/or employed for estimating soil moisture patterns. Western *et al.* (1999a) applied LOWESS regression (Hirsch *et al.* 1993) to estimate soil moisture based on topographic attributes and found that it does not substantially improve the amount of variance explained compared with linear regression. However, the extent of their LOWESS analyses in that work is not clear. Wilson *et al.* (2005) developed a linear estimation method with coefficients that vary nonlinearly with the spatial average soil moisture and applied the method to data from the Maharungi catchment in New Zealand. They found that the use of terrain attributes alone does not estimate realistic spatial moisture patterns but that the inclusion of a spatially stable residual pattern improves the estimates and concluded that factors other than topography are also important to soil moisture patterns. Finally, Perry & Niemann (2007, 2008) used empirical orthogonal functions (EOFs) to decompose a time series of spatial moisture patterns into patterns of covariation that are present to some extent on every date and interpolated those stable patterns with linear regressions against topographic attributes.

Our objectives are to evaluate the abilities of selected nonlinear estimation techniques for estimating soil

moisture from sparse soil moisture observations and to compare the results of those techniques to the results of MLR. We hypothesize that the nonlinear techniques will avoid some of the previously mentioned shortcomings of other methods and improve estimation of spatial soil moisture patterns. The nonlinear estimation methods used in this research are a spatial artificial neural network (SANN) (Shin & Salas 2000a) and mixture modeling (MM) with multivariate Gaussian distribution functions (McLachlan & Peel 2000). These methods were selected because they are unsupervised machine learning techniques that do not assume *a priori* any specific form of relationship between soil moisture and the predictor data. In addition, the SANN and MM are both kernel density estimation methods, which attempt to estimate the joint probability density function between soil moisture and the predictor variables. Finally, Green *et al.* (2007) applied the SANN to crop yields, which are closely related to soil moisture, and found good performance. Most previous applications of machine learning techniques in hydrology have focused on time-series data (Lin *et al.* 2006; Liu *et al.* 2008; Wang *et al.* 2009), including those focused specifically on soil moisture (Gill *et al.* 2006; Elshorbagy & El-Baroudy 2009; Ahmad *et al.* 2010). Here, the SANN and MM are applied to three study areas representing diverse climates and landscape characteristics to test their predictive abilities when different processes may dominate the moisture pattern formation.

METHODS

Multiple linear regression (MLR)

MLR is used as a baseline estimation method due to its simplicity and common use. The general model for linear regression can be written in matrix form as

$$\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\theta}$ is an $n \times 1$ vector of observed responses (soil moisture in this analysis), n is the number of observations, $\boldsymbol{X}$ is an $n \times d + 1$ matrix, d is the number of predictor variables (topographic attributes), $\boldsymbol{\beta}$ is a $d + 1 \times 1$ vector of unknown coefficients and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of residuals, or errors. Use of the ordinary least squares (OLS) criterion for estimating the coefficient vector leads to the following equation for the coefficient estimates:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\theta}. \tag{2}$$

OLS provides the minimum-variance unbiased coefficient estimates regardless of the distribution properties of the errors (Draper & Smith 1981). The main drawback of linear regression for this application is that it cannot account for possible nonlinear relationships between the topographic attributes and soil moisture.

Spatial artificial neural network (SANN)

The SANN method was developed by Shin & Salas (2000a) and can be viewed as a specific implementation of the Nadaraya–Watson model, or kernel regression (Nadaraya 1964; Watson 1964; Bishop 2006). The SANN has been used previously for regional drought analysis (Shin & Salas 2000b) and to estimate crop yields from topographic attributes (Green *et al.* 2007). Martinez *et al.* (2004) investigated the sensitivity of the SANN to its internal parameters using grain yield data and found optimal parameter values, which were subsequently used by Green *et al.* (2007).

The SANN is similar to kernel density estimation using multivariate Gaussian kernels. We can represent soil moisture, θ, at some location in space as a random variable in a d-dimensional domain by $\theta(\boldsymbol{x})$, where $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]$ is a vector of topographic attributes associated with the same spatial location. The optimal estimator of the soil moisture value is then the conditional expectation given by (Bishop 1995)

$$E[\theta(\boldsymbol{x})|\boldsymbol{x}] = \frac{\int_{-\infty}^{\infty} \theta(\boldsymbol{x}) p(\boldsymbol{x}, \theta) \mathrm{d}\theta}{\int_{-\infty}^{\infty} p(\boldsymbol{x}, \theta) \mathrm{d}\theta} \tag{3}$$

where $p(\boldsymbol{x}, \theta)$ is the joint probability density function of $\boldsymbol{x}$ and θ. The probability density function is estimated using multivariate Gaussian kernel density estimation (Specht 1991). If we observe the soil moisture $\theta(\boldsymbol{x})$ and a vector of topographic attributes $\boldsymbol{x}$ at N locations given by $[\boldsymbol{x}_n | n = 1, \ldots, N]$,

then the Gaussian kernel density estimator at any point $\boldsymbol{x}$ in the domain is given by

$$p(\boldsymbol{x},\theta) = \frac{1}{N}\sum_{n=1}^{N} G(\boldsymbol{x}|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)G(\theta|\mu_\theta,\sigma_\theta) \tag{4}$$

where

$$G(\boldsymbol{x}|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_n|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_n)^{\mathrm{T}}\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_n)\right] \tag{5}$$

with $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ the mean and covariance, respectively, of the Gaussian kernel associated with the nth observation. For simplicity, the variance of each kernel function is taken as equal in all dimensions of the predictor variable subspace so that all diagonal entries of $\boldsymbol{\Sigma}_n$ are equal. Also, the covariances (off-diagonal entries) are assumed to be zero. The diagonal entries of $\boldsymbol{\Sigma}_n$ are then denoted as σ_n^2. Also

$$G(\theta|\mu_\theta,\sigma_\theta) = \frac{1}{(2\pi)^{1/2}\sigma_\theta}\exp\left[-\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2}\right] \tag{6}$$

where μ_θ and σ_θ are the mean and standard deviation, respectively, of the soil moisture. After substituting Equation (4) into Equation (3) and simplifying, the result is

$$\hat{\theta}(\boldsymbol{x}) = \frac{\sum_{n=1}^{N}\theta(\boldsymbol{x}_n)G(\boldsymbol{x}|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)}{\sum_{n=1}^{N}G(\boldsymbol{x}|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)} \tag{7}$$

which may be used as a point estimator for θ. Note that in the simplification the numerator terms involving $G(\theta|\mu_\theta,\sigma_\theta)$ become the $\theta(\boldsymbol{x}_n)$terms inside the summation in Equation (7) while in the denominator each term involving $G(\theta|\mu_\theta,\sigma_\theta)$ integrates to a value of 1.

In order to use Equation (7) to estimate soil moisture, the widths of the kernel functions need to be specified. The width of the kernel function centered on observation n is denoted σ_n and is calculated by

$$\sigma_n = \mathrm{RMSD}_n/F \tag{8}$$

where RMSD_n is the root-mean-squared Euclidean distance (measured in the attribute domain) between data point n and its nearest P neighbors. The number of neighbors P and the factor F are the two parameters of the SANN. Both parameters help determine the spatial scale of the kernel function and their values must be specified prior to the SANN training. The parameter P relates directly to the kernel widths, and the effects of P on the individual kernel widths depend on the data configuration and density. The F parameter is related inversely to the kernel widths and affects all kernels to the same degree. We tested the effects of adjusting each parameter and found that comparable results were achieved through manipulation of either parameter. Therefore, in the present implementation, the value of F was fixed at 2.5, which is the value recommended by Martinez *et al.* (2004) for large datasets, but P was considered a free parameter and various values were tested (see below). One potential drawback of the SANN is that it requires all the observations to be stored in order to make future estimates, which can make evaluation slow if the quantity of data is large.

Mixture model (MM)

The MM method, like the SANN, is capable of capturing nonlinear relationships between the topographic attributes and soil moisture. Operationally, it is also similar to the SANN in that estimates are made by conditioning a multivariate density function with the values of the predictor variables. However, the number of kernel functions employed by MM to form the model is less than the number of observations and the method employs the expectation–maximization (EM) algorithm (Dempster *et al.* 1977; McLachlan & Peel 2000) to identify optimal locations for the kernels. In the present implementation, Gaussian functions were used as the kernel functions in the MM, and we will refer to them as components, which is common terminology in the MM literature. The multivariate density function developed by MM has the form (Bishop 2006)

$$p(\boldsymbol{x},\theta) = \sum_{k=1}^{K}\pi_k G(\{\boldsymbol{x},\theta\}|\boldsymbol{\mu_k},\Sigma_k) \tag{9}$$

where K is the number of components in the model and is the only parameter of this method, π_k is the mixing coefficient associated with the kth component, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, are the mean and covariance, respectively, of the kth component. The mixing coefficient values lie between 0 and 1, and they sum to 1. The form of each Gaussian component is

$$G(\{\boldsymbol{x},\theta\}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{(d+1)/2}|\Sigma_k|^{1/2}} \times \exp\left[-\frac{1}{2}(\{\boldsymbol{x},\theta\}-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\{\boldsymbol{x},\theta\}-\boldsymbol{\mu}_k)\right] \quad (10)$$

where d is the number of predictor variables. The values of π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, are determined by the EM algorithm, which maximizes the likelihood of the model. Equation (3) is used again to estimate the soil moisture given the values of the ancillary data for any given location. The MATLAB statistics toolbox (Mathworks 2009) version of MM, *gmdistribution*, was used to implement the MM method. As with the SANN, the density function is then conditioned on the observed value of the topographic attribute to determine the soil moisture estimate.

Experimental design

A Monte Carlo cross-validation methodology, which provides an empirical measure of the methods' generalization abilities, was employed (McLachlan & Peel 2000; Bishop 2006) to assess the effectiveness of the different methods for estimating soil moisture. The cross-validation methodology divides each set of observations into training and testing sets. The training sets were created by randomly sampling the data available at each site 30 times for each of five sampling rates: 10, 25, 50, 75 and 95%. For each training set, the observations at unsampled locations comprise the testing set. Because all the methods are empirical, they require some amount of observations to develop a model. Each of the three estimation methods was used to develop a model of the relationship between topographic attributes and soil moisture values based on the training sets. After a model was developed from a specific training set, soil moisture was estimated with the model at all locations, but the Nash–Sutcliffe coefficient of efficiency (NSCE) (Nash & Sutcliffe 1970) was calculated based only on the testing set to measure the model performance. The NSCE is defined as

$$\mathrm{NSCE} = 1 - \frac{\sum_{n=1}^{N}(\theta_n - \hat{\theta}_n)^2}{\sum_{n=1}^{N}(\theta_n - \bar{\theta})^2} \quad (11)$$

where N is the number of data points in the testing set, θ_n is the nth observed soil moisture, $\hat{\theta}_n$ is the model estimate of the nth observation and $\bar{\theta}$ is the average of the observations. Note that the NSCE has a maximum value of 1, for which all observed variance is explained, but there is no minimum value. Because 30 realizations of the training data are produced for every sampling rate, the typical performance of a particular estimation method is characterized by the median NSCE calculated from the 30 realizations.

A large number of estimation scenarios were performed to compare the methods. For each of the three methods, all 63 possible subsets of the six topographic attributes were tested as predictor variable sets. For each of those subsets of predictor variables, 30 sets of training data were supplied to the method for each of the five sampling rates given above. This collection of estimation problems was repeated for all three wetness conditions at all three study sites. The SANN and MM methods additionally have parameters that were adjusted. The P parameter in the SANN was tested with values of 6, 9, 15, 20, 30 and 45. Larger values of P lead to larger kernel widths and a smoother estimated density function. For the MM method, the adjustable parameter is K, the number of Gaussian components, which was varied from 1 to 4. The number of components controls the potential complexity of the resulting density function and the inferred relationship between the topographic attributes and the soil moisture.

APPLICATION SITES AND DATA

Soil moisture data were compiled for three application sites: the Tarrawarra catchment in southeastern Australia, the Satellite Station site in the Maharungi catchment on the North Island of New Zealand and the Cache la Poudre site in north-central Colorado in the United States. Soil moisture patterns from three dates were used for each site. The

dates were chosen to represent dry, moderate and wet conditions as defined by the range of spatial mean soil moisture values Θ observed at each study site. All the soil moisture data were measured with TDR probes.

Tarrawarra

The first soil moisture dataset is from the Tarrawarra catchment located near Melbourne, Australia and was originally described by Western & Grayson (1998). The catchment has a temperate climate with an average annual rainfall of 820 millimeters (mm), average annual potential evapotranspiration of 830 mm and a rainfall deficit in summer and excess in winter. The catchment is covered by pasture for cattle grazing. Soils generally have a silty loam A horizon and a B horizon with higher clay content and soil depths vary from 40 centimeters (cm) in the upper catchment to over 2 meters (m) in the low areas.

Soil moisture data were collected in the top 30 cm of the soil with TDR probes on 13 dates between September 27, 1995 and November 29, 1996 on a 10×20 m sampling grid (Western *et al.* 1999b). We filtered the dataset to remove locations with missing values on any date, so the remaining dataset includes 454 points. February 14, 1996, September 27, 1995 and July 3, 1996 were selected as the dry, medium and wet dates, respectively. The spatial average soil moisture (Θ) values for those dates are 26.4, 38.0 and 45.3% volume of water/total volume (V/V), respectively. A digital elevation model (DEM) with 5 m resolution is available for this site. The topography consists of undulating hills (Figure 1) with no incised drainage channels and the catchment area is 10.5 ha. Total relief for the Tarrawarra catchment is 29 m. The dry moisture pattern does not visually exhibit any organization while the moderate and wet dates show strong and moderate degrees of organization, respectively, with wetter sites tending to occur in valley bottoms (Figure 1).

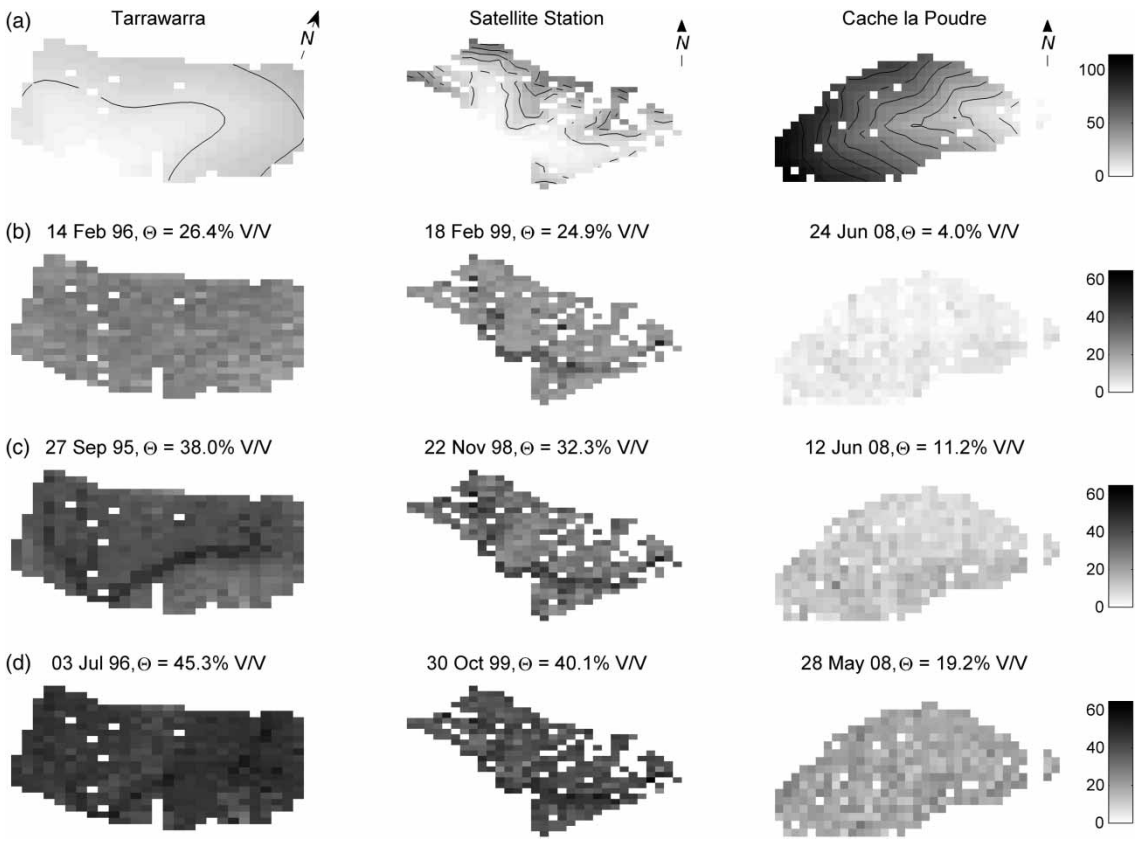

Figure 1 | (a) Elevations relative to the low point at each study site and spatial patterns of soil moisture for (b) dry, (c) medium and (d) wet dates used.

Satellite Station

The second dataset is from the Satellite Station field site of the Maharungi River Variability Experiment (MARVEX) conducted in New Zealand (Woods *et al.* 2001; Wilson *et al.* 2003). The Satellite Station site has a humid climate with annual rainfall of approximately 1,200 mm and an annual pan evapotranspiration of approximately 1,300 mm (Wilson *et al.* 2003). Hillslope soils are generally silty clay loam, while lowland valley soils are relatively deep alluvial fill with high clay content. The dataset used here is comprised of TDR measurements collected at 0–30 cm depth on a 40 m grid for six dates between April 1998 and November 1999. The dataset was again filtered so that it includes only locations with measurements on all dates (322 locations). The dates chosen for this study are February 18, 1998, March 25, 1998 and October 30, 1999, with spatial average soil moisture values of 24.9, 32.3 and 40.1% V/V, respectively. The topographic data used was from a 10 m resolution DEM. The Satellite Station site has undulating terrain with a total area of approximately 60 ha and it contains two subcatchments within it (Figure 1). The relief of the site is approximately 80 m. Visually, the moisture patterns at Satellite Station are more stable through time than those at Tarrawarra with the valley bottoms remaining wetter than the upland areas as the overall wetness condition changes (Figure 1).

Cache la Poudre

The final dataset is from a catchment near Rustic, Colorado that is part of the Cache la Poudre River basin (Lehman & Niemann 2008). The catchment has a semi-arid climate with an annual precipitation of about 400 mm and annual potential evapotranspiration of about 930 mm. The vegetation on the north-facing slope of the catchment is predominantly coniferous forest with scattered shrubs in open areas and near ridges. The south-facing slope has scattered shrubs with a few coniferous trees. The site has thin sandy soil on the south-facing hillslope and thicker mineral soils overlaid with organic matter on the north-facing hillslope.

The sampling strategy for the Poudre site consisted of collecting manual TDR measurements immediately after a sequence of spring rainfall events and during subsequent periods of drying, which resulted in a total of 12 soil moisture patterns. Due to the shallow soils, particularly on the south-facing hillslope, surface soil moisture was measured in the top 5 cm of the soil once any litter layer was temporarily removed. After filtering to produce a consistent dataset for all sampling dates, a total of 350 locations remain for this dataset. The dates chosen for application of the interpolation methods are June 24, 2008, June 12, 2008 and May 28, 2008. The spatial average soil moisture values on these dates are 4.0, 11.2 and 19.2% V/V, respectively, which represent dry, medium and wet conditions for this dataset. The DEM for the catchment has 15 m resolution and a local coordinate system offset approximately 1 degree from north. The total relief for the Poudre site is approximately 124 m (Figure 1). The catchment consists of the headwater area for one incised channel with both steep and flat portions and the catchment area is approximately 8 ha. Visually, the soil moisture patterns exhibit slightly wetter conditions on the north-facing slope than the south-facing slope (Figure 1).

Topographic attributes

The form of the topography for each catchment was characterized using elevation and five additional topographic attributes: slope, cosine of the topographic aspect (cosA), the logarithm of the specific contributing area (logSCA), the sum of the plan and profile curvatures (Curv) and the PSRI. These attributes are related to different processes and variables that affect soil moisture. Specifically, surface slope is related to the horizontal hydraulic gradient of subsurface flows and to insolation, a primary driver of evapotranspiration and snowmelt (Western *et al.* 1999a). Aspect also affects insolation (Western *et al.* 1999a). The SCA is a measure of the upslope area that can potentially contribute flow to a unit length of contour on the surface (Western *et al.* 1999a). The profile curvature is related to the change in the hydraulic gradient and hence the velocity of flow (Mitasova & Hofierka 1993) and the plan curvature is related to the degree of surface and flow convergence along an elevation contour (Mitasova & Hofierka 1993). The number of attributes was intentionally limited to allow testing of all possible attribute combinations in the analyses

below. The wetness index (Beven & Kirkby 1979) was not included because it combines logSCA and slope in a predetermined way. If the wetness index is important, the nonlinear methods should be able to identify that importance using the underlying variables.

The SCA and the slope were calculated using the D_∞ algorithm (Tarboton 1997) as implemented in TauDEM (Tarboton 2008). SCA was initially used as a predictor variable but its large skewness resulted in numerical complications and led us to use log-transformed values instead. The aspect for each DEM cell was calculated according to the formula in Moore *et al.* (1991) with units of degrees clockwise from the north. The aspect values were then cosine-transformed, which provides a continuous range of values from –1 to 1 and separates north-facing aspects from south-facing aspects by sign (Green *et al.* 2007).

The sum of the profile and plan curvatures (Mitasova & Hofierka 1993) was the only curvature measure used. The profile curvature is the curvature of the intersection of the topographic surface with a vertical plane oriented in the downslope direction. The plan curvature is the curvature of the intersection of the surface with a horizontal plane. The sum of the two curvatures was chosen based on a screening analysis designed to select the single curvature measure with the highest overall potential for estimating soil moisture.

The PSRI depends on the day of the year, latitude, local slope and aspect (Western *et al.* 1999a; Dingman 2002). No allowances were made for other factors affecting actual insolation, such as shading by vegetation or atmospheric attenuation, in the PSRI calculation. The PSRI was calculated individually for each sampling date, so it is the only topographic attribute that changes between different dates at the same catchment.

RESULTS

Typical results

Figure 2 presents observed soil moisture patterns and moisture patterns estimated by each method when 25% of the available data is used for training. The patterns shown are from the set of topographic attributes that produces the highest median NSCE for each method and the training data that produce the results closest to that median NSCE among the different samples (using those topographic attributes). The P value used in the SANN is 45, which gave the best performance among all tested values. The MM method uses two component densities. In almost all situations, one component density provided slightly superior performance, but the MM method with one component is very similar to linear regression. The results for two components are shown here to highlight the method's ability as a potentially nonlinear method. All of the methods reproduce the main features observed in the patterns. The soil moisture pattern for Tarrawarra exhibits a pronounced organization, with wet areas occurring in the valley bottoms and on the south-facing hillslope, and all of the methods capture this general tendency. The Satellite Station pattern is more weakly organized than Tarrawarra, but the estimation methods capture some of this organization. The Cache la Poudre site has an aspect-dependent pattern, which is the main feature that is reproduced by the estimation methods. The NSCE values included in the figure are calculated only from the associated testing dataset. For all three sites, the SANN has the highest NSCE value although all methods perform very similarly for Tarrawarra. The NSCE values are all significantly lower for the other two sites than for Tarrawarra, but the SANN remains the best-performing method. The MM is the second-best method for Satellite Station, while MLR is the second-best method for Cache la Poudre. The SANN method likely performs better than the other methods because it has the most flexibility in the type of relationship that it can infer from the data. Such flexibility would allow the SANN to include subtleties in the relationships to the topographic attributes that are ignored by the other two methods.

Number of predictor variables

Figure 3 plots the median NSCE for the three methods as the number of topographic attributes used varies. All three wetness conditions are shown for the three application sites. In all cases, 25% of the observations were used as training data, and the lines in the figure indicate the median NSCE values from all 30 samples at that sampling rate (calculated from only the testing locations). For each

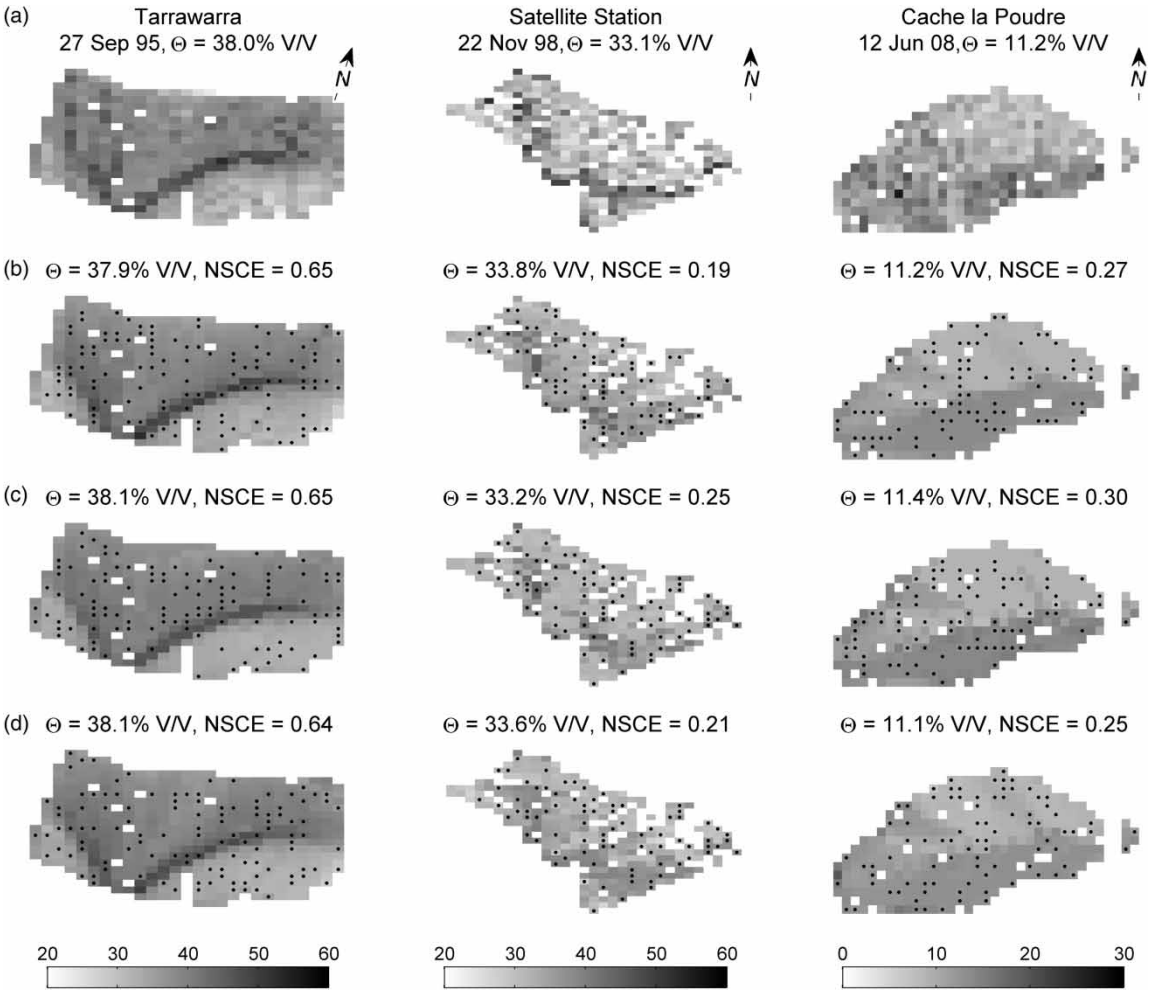

Figure 2 | (a) Observed soil moisture patterns at each study site, and soil moisture patterns estimated by (b) MLR, (c) SANN and (d) MM. The dots indicate training data locations.

subset size, the NSCE shown is from the attribute set with the highest median value. For the SANN and MM methods, the parameter values are those that produce the best possible performance in each case. No condition is used to determine whether the additional predictor variables explain a statistically significant portion of the variance in the soil moisture observations for any of the methods.

In nearly all scenarios in Figure 3, the performance of the three methods is similar. The broadest range of performance between methods is observed at Satellite Station for the dry and medium dates. In both of those cases, the SANN has the best performance. The SANN also performs the best in most other cases in the figure and it typically performs the best for other sampling rates (not shown). All of the methods perform best for the moderate wetness condition at the Tarrawarra and Cache la Poudre sites, but at the Satellite Station site the methods perform similarly for the dry and moderate conditions and perform worst for the wettest condition. Reduced performance in dry and wet conditions is consistent with reduced spatial structure in the soil moisture patterns and greater importance of local controls (e.g. porosity) relative to topography under those conditions (Grayson *et al.* 1997; Western *et al.* 1999a). The similarity of the results of the MLR, MM and SANN indicates that allowing nonlinearity in the relationships between soil moisture and topographic attributes does not substantially improve pattern estimation. This result likely implies that the relationships (where meaningful) are close to linear. For most scenarios in the figure, the methods show little variation in NSCE as predictor

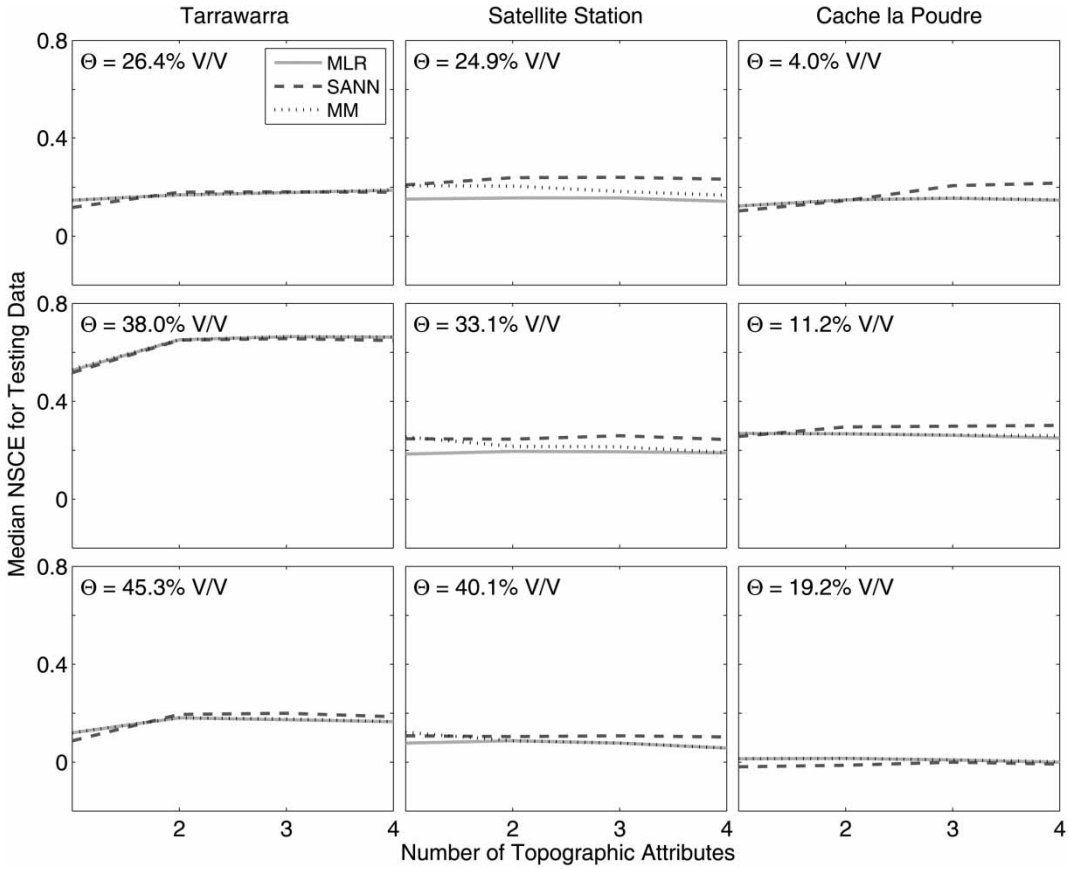

Figure 3 | Median NSCE for each estimation method as the number of topographic attributes used increases.

variables are added. The exceptions are the moderate and wet dates at Tarrawarra as well as the dry date at Cache la Poudre. For those two dates at Tarrawarra, the NSCE values increase substantially with the use of a second predictor variable and then remain roughly constant. For the dry date at Cache la Poudre, the SANN method improves consistently with the addition of more predictor variables while the other two methods show little change. Overall, the NSCE for the SANN usually remains roughly constant or improves with additional predictor variables, while the other two methods generally decrease slightly with additional predictor variables. The lack of substantial improvement with additional variables suggests that one or two attributes represent the major effects of the dominant physical processes. Where they occur, negative trends in NSCE with additional variables indicate overtraining, which leads to reduced generalization capability. When higher sampling rates are used (not shown), the NSCE values are less likely to decrease as the number of predictor variables increases. That result is expected because higher sampling rates likely include more information that could justify the inclusion of additional attributes.

Sample size

The effects of the size of the training dataset on the performance of the methods are presented in Figure 4. The figure shows the results for the best set of two predictor variables for each of the methods, where the best set is defined as the set that produces the maximum value for the median NSCE. The rows in Figure 4 show the results for a given method while the columns show the results for a given study site. Only the moderate wetness condition is shown for each site, but the results are consistent for all three conditions. The box-and-whisker plots in the figure characterize the variation in the performance of each

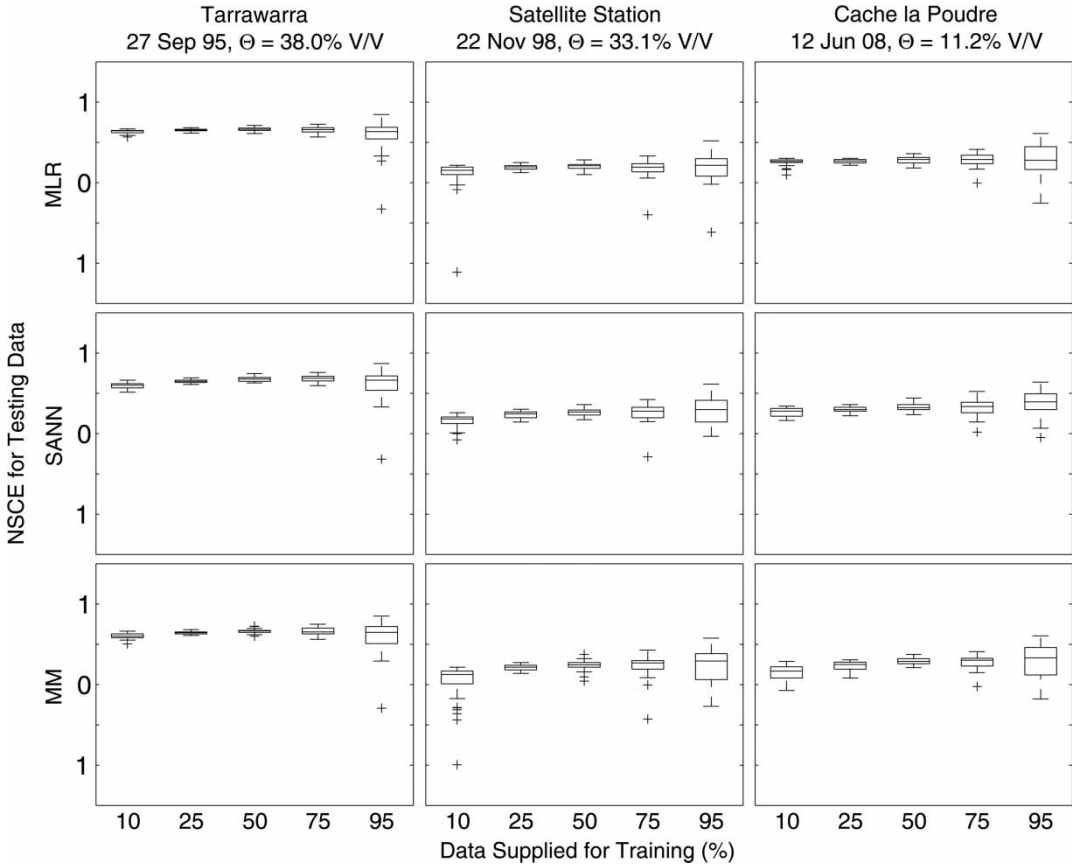

Figure 4 | Box-and-whisker plots characterizing the performance of each estimation method as the amount of training data used increases. Plus symbols indicate outliers, which are defined as values that are more than 1.5 times the distance between the upper and lower quartiles away from the box limits.

method among the 30 training sets generated at a given sampling density. The upper and lower limits of the boxes represent the 25th and 75th percentiles, respectively, for the NSCE and the horizontal line within the box represents the median value. Whiskers represent the limits of observed NSCE values that are not considered outliers while outliers are marked by plus signs.

For all the methods, the performance does not increase much as more observations are supplied for training, and in some cases performance can decline with increasing data due to overtraining. The highest median NSCE values for the SANN and MM that are achieved as one varies the sampling rate are higher than for MLR (for all study sites). As a specific example, at the Cache la Poudre site the highest median NSCE for MLR is 0.29, which occurs at the 75% sampling rate. For the SANN and MM, the values are 0.39 and 0.33, respectively, and they both occur at the 95% sampling rate. However, the sampling rates for which the median NSCE reaches its highest value with the MLR method for the Tarrawarra, Satellite Station and Cache la Poudre sites are 50, 95 and 75%, respectively, while the rates for the SANN are 75, 95 and 95%, and for MM they are 50, 95 and 95%. Overall, the nonlinear methods' best performances were obtained by using more data than the MLR, reflecting their ability to continue extracting useful information as the amount of data increases. However, the highest median NSCE occurs for all methods and sites with at least 50% of the data being used to train the estimator, which represents a significant data collection effort. The numbers of measurements represented by the 10 and 25% sampling rates are more feasible for regular manual collection. At those sampling rates the SANN has the highest median NSCE values at the Satellite Station and Cache la Poudre sites while MLR has the highest median values at Tarrawarra. Overall, the SANN achieves equivalent or higher NSCE values than MLR when the sampling rate is

at least 25%, and the MM also outperforms MLR in most cases when the sampling rate is at least 50%.

Predictor sets chosen by methods

The SANN and MM can model nonlinear relationships between topographic attributes and soil moisture, so it is possible that they perform best when using different topographic attributes than the MLR uses. To investigate this possibility, we analyzed the frequency with which each attribute is included in models with relatively high NSCE values for each estimation method. Specifically, the top 10% of the models in terms of NSCE were identified for each method when 25% of the observations were used as training data, and the fraction of those models that contains each topographic attribute was calculated. For example, for models with two predictor variables, there are 15 combinations of predictor variables, and 30 samples were tested for each combination, for a total of 450 results. From those results, the 45 cases with the highest NSCE values were analyzed.

In Figure 5, the height of the bar associated with each attribute represents the proportion of those cases that contain that attribute. For each study site, all three methods perform best using the same attribute in the one-variable case although the attribute is different for each site. That single attribute remains the most frequently used as the number of variables increases. The best-performing single attribute is logSCA for Tarrawarra, slope for Satellite Station and cosA for Cache la Poudre. The logSCA is related to the convergence of flow due to topography and suggests the importance of water redistribution at Tarrawarra. The Satellite Station topography is generally divided between highland areas with high slopes and lowland areas with low slopes, and the soil moisture values reflect that division as well. The preference for cosA at the Cache la Poudre site reflects the distinct vegetation on the two opposing hillslopes at this site. For the two-predictor variable scenario at Tarrawarra, the second selected variable is split approximately equally between cosA and PSRI for all methods, indicating the importance of radiation-driven

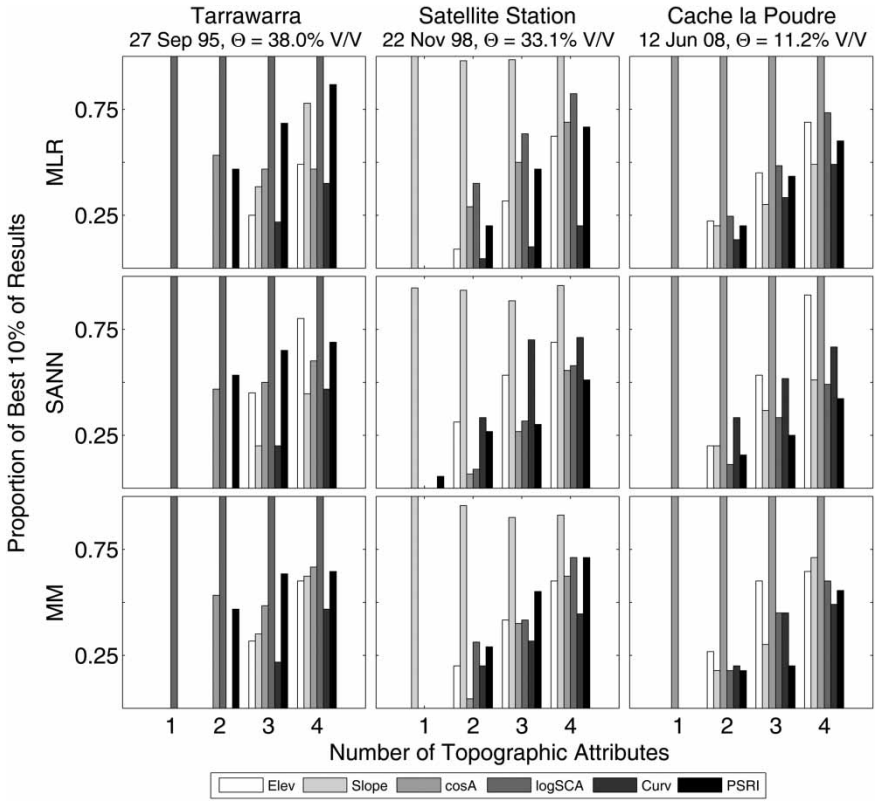

Figure 5 | Proportion of times each topographic attribute is selected by the best-performing models as a function of the number of topographic attributes used.

evapotranspiration variations at this site. There is not a dominant secondary variable chosen by any of the methods in the two-predictor variable scenario at the Satellite Station or Cache la Poudre sites. Beyond the two-variable case, no consistent preferences for additional attributes are observed at any site.

This analysis was repeated for the dry and wet dates at each site. For the dry date (not shown), the most significant difference from the moderate wetness date is that the models now perform best when using PSRI at Tarrawarra, which indicates greater importance of evapotranspiration compared with flow convergence. Additionally, elevation is a clear secondary attribute selected at Satellite Station by the SANN and MM but not the MLR. For the wet date (not shown), the only noteworthy difference from the moderate wetness scenario is the lack of a clearly preferred attribute at the Cache la Poudre site for the SANN. However, the MLR and MM methods preferentially selected logSCA as the first predictor. Overall, the methods tend to select the same attributes for soil moisture estimation at each catchment. Therefore, performance differences between the methods are due to differences in the forms of the modeled relationships between the attributes and soil moisture rather than due to the selection of different attributes.

We also implemented each method in a stagewise manner using the same topographic attributes and sampling scheme. In this implementation, the best single predictor variable was chosen first and its estimate of soil moisture was retained. Then, this predictor variable was removed from subsequent consideration, and the remaining predictor variables were evaluated in their ability to explain the residuals. This process was repeated until all predictor variables were used. We found that the stagewise method does not offer any improvement in performance over the original implementation.

CONCLUSIONS

In this paper, we investigated the efficacy of nonlinear methods for estimating soil moisture patterns from sparse observations and compared them to MLR. The nonlinear estimation methods considered are an SANN and a Gaussian MM. These methods were applied to three different study sites and three wetness conditions for each site using several different sizes for the training datasets.

The SANN method consistently outperforms the MM method and MLR and is the best overall method tested. For the majority of locations, wetness conditions and levels of training data, it provides higher NSCE values than the other two methods. In most scenarios, the increase in performance is not large, but the improvement is consistent and the method never performs much worse than the other methods. Another positive aspect of the SANN method is its superior performance when using multiple predictor variables. All the methods tend to perform the best at a given site when using the same topographic attribute as a single predictor variable, but a different attribute is best for each of the sites. However, because SANN performs better than the other methods when using multiple predictor variables, the *a priori* selection of one or two attributes for soil moisture estimation would not be required for that method. Thus, one could use a single, larger set of topographic attributes at a variety of sites. MLR may also be able to be used in a similar fashion if a suitable test for statistical significance is evaluated before a predictor variable is added to the model.

The estimation accuracy of all of the methods depends on both the catchment and wetness condition for which estimates are made. All methods perform best at Tarrawarra under the moderate wetness condition and perform worst at Cache la Poudre for the wet condition. Thus, not only do individual site characteristics determine how well the methods estimate soil moisture from topographic attributes, but those characteristics also affect the wetness condition under which the methods perform best.

In order for the SANN and MM to consistently perform better than MLR, relatively large training datasets are required. More training data leads to higher NSCE values for most scenarios tested. For all scenarios except the wet condition at Cache la Poudre, SANN performs equivalently or better than MLR for sampling rates greater than or equal to 25%, and for most scenarios MM outperforms MLR at the 50% sample rate and above. The superior performance of the SANN and MM methods suggest that the relationships between soil moisture and topographic attributes might be weakly nonlinear, but more data are required for the

estimation methods to discern any nonlinearity and for the improved performance of these methods to be realized.

Several directions are open for future research. These include testing the abilities of the methods to estimate soil moisture under circumstances (e.g. wetness conditions, locations or DEM resolutions) different from those for which the methods were trained. In addition, other statistical learning techniques such as support vector machines and genetic algorithms could be evaluated for estimating soil moisture patterns.

ACKNOWLEDGEMENTS

The authors acknowledge the Army Research Office and the Center for Geosciences and Atmospheric Research for their financial support. The authors thank all parties involved in MARVEX, including researchers from NIWA and the University of Melbourne, as well as Tim Green from the Agricultural Research Service for helpful suggestions. The authors also thank Brandon Lehman, Josh Melliger, Jonathan Freed and Rob Erskine for collecting the Cache la Poudre data, and the anonymous reviewers for their suggestions for improving this paper.

REFERENCES

Ahmad, S., Kalra, A. & Stephen, H. 2010 Estimating soil moisture using remote sensing data: a machine learning approach. *Adv. Water Res.* **33** (1), 69–80.

Bardossy, A. & Lehmann, W. 1998 Spatial distribution of soil moisture in a small catchment. Part 1: geostatistical analysis. *J. Hydrol.* **206**, 1–15.

Beven, K. J. & Kirkby, M. J. 1979 A physically-based, variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* **24** (1), 43–69.

Bishop, C. M. 1995 *Neural Networks for Pattern Recognition*, 1st edition. Oxford University Press, New York.

Bishop, C. M. 2006 *Pattern Recognition and Machine Learning*, 1st edition. Springer, New York.

Chiles, J.-P. & Delfiner, P. 1999 *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, New York.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. Ser. B–Methodol.* **39** (1), 1–38.

Dingman, S. L. 2002 *Physical Hydrology*, 2nd edition. Prentice Hall, Englewood Cliffs, NJ.

Draper, N. R. & Smith, H. 1981 *Applied Regression Analysis*, 2nd edition. John Wiley & Sons, New York.

Dunne, T. & Black, R. D. 1970 Partial area contributions to storm runoff in a small New England watershed. *Water Resour. Res.* **6** (5), 1296.

Dunne, T., Moore, T. R. & Taylor, C. H. 1975 Recognition and prediction of runoff-producing zones in humid regions. *Hydrol. Sci. Bull.* **20**, 305–327.

Eagleson, P. S. 1978 Climate, soil and vegetation. 3. Simplified model of soil-moisture movement in liquid-phase. *Water Resour. Res.* **14** (5), 722–730.

Elshorbagy, A. & El-Baroudy, I. 2009 Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. *J. Hydroinf.* **11** (3–4), 237–251.

Entekhabi, D., Rodriquez-Iturbe, I. & Castelli, F. 1996 Mutual interaction of soil moisture state and atmospheric processes. *J. Hydrol.* **184**, 3–17.

Gill, M. K., Asefa, T., Kemblowski, M. W. & McKee, M. 2006 Soil moisture prediction using support vector machines. *J. AWRA* **42** (4), 1033–1046.

Grayson, R. B., Western, A. W., Chiew, F. H. S. & Bloschl, G. 1997 Preferred states in spatial soil moisture patterns: local and nonlocal controls. *Water Resour. Res.* **33** (12), 2897–2908.

Green, T. R. & Erskine, R. H. 2004 Measurement, scaling, and topographic analyses of spatial crop yield and soil water content. *Hydrol. Process.* **18**, 1447–1465.

Green, T. R., Salas, J. D., Martinez, A. & Erskine, R. H. 2007 Relating crop yield to topographic attributes using Spatial Analysis Neural Networks and regression. *Geoderma* **139**, 23–37.

Hewlett, J. D. & Hibbert, A. R. 1967 Factors affecting the response of small watersheds to precipitation in humid areas. In: *Forest Hydrology: Proc. of a National Science Foundation Advanced Science Seminar. Pennsylvania State University* (W. E. Sopper & H. W. Lull, eds). Pergamon Press, New York, p. 813.

Hillel, D. 1998 *Environmental Soil Physics*. Academic, New York.

Hirsch, R. M., Helsel, D. R., Cohn, T. A. & Gilroy, E. J. 1993 Statistical analysis of hydrologic data. In: *Handbook of Hydrology* (D. R. Maidment, ed.). McGraw-Hill, New York, pp. 17.11–17.55.

Journel, A. G. & Huijbregts, C. J. 1978 *Mining Geostatistics*. Academic, New York.

Kitanidis, P. K. 1993 Geostatistics. In: *Handbook of Hydrology* (D. R. Maidment, ed.). McGraw-Hill, New York, p. 1424.

Lehman, B. M. & Niemann, J. D. 2008 Spatial patterns of in a semi-arid montane catchment with aspect-dependent vegetation. Paper presented at *1st Int. Conf. on Hydropedology*, State College, Pennsylvania, PA.

Lin, J. Y., Cheng, C. T. & Chau, K. W. 2006 Using support vector machines for long-term discharge prediction. *Hydrol. Sci. J.–J. Sci. Hydrol.* **51** (4), 599–612.

Liu, H. B., Xie, D. & Wu, W. 2008 Soil water content forecasting by ANN and SVM hybrid architecture. *Environ. Monitor. Assess.* 143 (1–3), 187–193.

Martinez, A., Salas, J. D. & Green, T. R. 2004 Sensitivity of spatial analysis neural network training and interpolation to structural parameters. *Math. Geol.* **36** (6), 721–742.

Mathworks 2009 *Matlab.* The Mathworks, Natick, MA.

McLachlan, G. & Peel, D. 2000 *Finite Mixture Models*. John Wiley & Sons, New York.

Mitasova, H. & Hofierka, J. 1993 Interpolation by regularized spline with tension: II. Application to terrain modeling and surface geometry analysis. *Math. Geol.* **25** (6), 657–669.

Moore, I. D., Burch, G. J. & Mackenzie, D. H. 1988 Topographic effects on the distribution of surface soil water and the location of ephemeral gullies. *Trans. Am. Soc. Ag. Engr.* **31** (4), 1098–1107.

Moore, I. D., Grayson, R. B. & Ladson, A. R. 1991 Digital terrain modeling – a review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **5** (1), 3–30.

Nadaraya, E. A. 1964 On estimating regression. *Theory Prob. Appl.* **9** (1), 141–142.

Nash, J. E. & Sutcliffe, J. V. 1970 River forecasting through conceptual models: Part I, A discussion of principles. *J. Hydrol.* **10**, 282–290.

Nyberg, L. 1996 Spatial variability of soil water content in the covered catchment at Gardsjon, Sweden. *Hydrol. Process.* **10**, 89–103.

Perry, M. A. & Niemann, J. D. 2007 Analysis and estimation of soil moisture at the catchment scale using EOFs. *J. Hydrol.* **334** (3–4), 388–404.

Perry, M. A. & Niemann, J. D. 2008 Generation of soil moisture patterns at the catchment scale by EOF interpolation. *Hydrol. Earth Syst. Sci.* **12** (1), 39–53.

Robinson, D. A., Jones, S. B., Wraith, J. M., Or, D. & Friedman, S. P. 2003 A review of advances in dielectric and electrical conductivity measurement in soils using time domain reflectometry. *Vadose Zone J.* **2** (4), 444–475.

Rodriguez-Iturbe, I. 2000 Ecohydrology: a hydrologic perspective of climate-soil-vegetation dynamics. *Water Resour. Res.* **36** (1), 3–9.

Rodriguez-Iturbe, I., Vogel, G. K., Rigon, R., Entekhabi, D., Castelli, F. & Rinaldo, A. 1995 On the spatial organization of soil moisture fields. *Geophys. Res. Lett.* **22** (20), 2757–2760.

Shin, H.-S. & Salas, J. D. 2000a Spatial analysis of hydrologic and environmental data based on artificial neural networks. In: *Artificial Neural Networks in Hydrology* (R. S. Govindaraju & A. R. Rao, eds). Kluwer Academic, Dordrecht, pp. 259–286.

Shin, H.-S. & Salas, J. D. 2000b Regional drought analysis based on neural networks. *J. Hydrol. Eng.* **5** (2), 11.

Specht, D. F. 1991 A general regression neural network. *IEEE Trans. Neural Net.* **2** (6), 568–576.

Sulebak, J. R., Tallaksen, L. M. & Erichsen, B. 2000 Estimation of areal soil moisture by use of terrain data. Geog. *Ann. Ser. A Phys. Geog.* **82A** (1), 89–105.

Tarboton, D. G. 1997 A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resour. Res.* **33** (2), 309–319.

Tarboton, D. G. 2008 *TauDEM, 3.1*. Utah State University, Logan, UT.

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A. & Hopmans, J. W. 2008 On the value of soil moisture measurements in vadose zone hydrology: a review. *Water Resour. Res.* **44**, W00D06.

Wang, W. C., Chau, K. W., Cheng, C. T. & Qiu, L. 2009 A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *J. Hydrol.* **374** (3–4), 294–306.

Watson, G. S. 1964 Smooth regression analysis. *Sankhya Indian J. Stat. A* **26**, 359–372.

Western, A. W., Bloschl, G. & Grayson, R. B. 2001 Toward capturing hydrologically significant connectivity in spatial patterns. *Water Resour. Res.* **37** (1), 83–97.

Western, A. W. & Grayson, R. B. 1998 The Tarrawarra data set: soil moisture patterns, soil characteristics, and hydrological flux measurements. *Water Resour. Res.* **34** (10), 2765–2768.

Western, A. W., Grayson, R. B., Bloschl, G., Willgoose, G. R. & McMahon, T. A. 1999a Observed spatial organization of soil moisture and its relation to terrain indices. *Water Resour. Res.* **35** (3), 797–810.

Western, A. W., Grayson, R. B. & Green, T. R. 1999b The Tarrawarra project: high resolution spatial measurement, modelling and analysis of soil moisture and hydrological response. *Hydrol. Process.* **13**, 633–652.

Wilson, D. J., Western, A. W. & Grayson, R. B. 2005 A terrain and data-based method for generating the spatial distribution of soil moisture. *Adv. Water Res.* **28** (1), 43–54.

Wilson, D. J., Western, A. W., Grayson, R. B., Berg, A. A., Lear, M. S., Rodell, M., Famiglietti, J. S., Woods, R. A. & McMahon, T. A. 2003 Spatial distribution of soil moisture over 6 and 30 cm depth, Mahurangi river catchment, New Zealand. *J. Hydrol.* **276** (1–4), 254–274.

Woods, R. A., Grayson, R. B., Western, A. W., Duncan, M., Wilson, D. J., Young, R. I., Ibbit, R., Hsenderson, R. & McMahon, T. A. 2001 Experimental design and initial results from the Maharungi River Variability Experiment: MARVEX. In: *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling, Water Science and Application* (V. Lakshmi, J. D. Albertson & J. Schaake, eds). American Geophysical Union, Washington, DC, pp. 201–213.

Yates, S. R. & Warrick, A. W. 1987 Estimating soil water content using cokriging. *Soil Sci. Soc. Am. J.* **51** (1), 23–30.

Zaslavsky, D. & Sinai, G. 1981 Surface hydrology: 1. Explanation of phenomena. *J. Hydrol. Div.* **107** (1), 1–16.

Zhu, H. & Journel, A. G. 1993 Formatting and integrating soft data: stochastic imaging via the Markov–Bayes algorithm. In: *Geostatistics Troia '92* (A. Soares, ed.). Kluwer, Dordrecht, pp. 1–12.

First received 23 November 2010; accepted in revised form 5 October 2011. Available online 27 January 2012