Journal of Hydrology 399 (2011) 201–211

Contents lists available at ScienceDirect

Journal of Hydrology

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

Enhancing PTFs with remotely sensed data for multi-scale soil water retention estimation

Raghavendra B. Jana *, Binayak P. Mohanty

Department of Biological and Agricultural Engineering, Texas A&M University, TX 77843-2117, United States

ARTICLE INFO

Article history:
Received 14 April 2010
Received in revised form 27 September 2010
Accepted 30 December 2010
Available online 4 January 2011

This manuscript was handled by P. Baveye, Editor-in-Chief

Keywords:
Bayesian Neural Networks
Multiscale methods
Pedotransfer functions
Remote sensing
Data
Non-linear bias correction

SUMMARY

Use of remotely sensed data products in the earth science and water resources fields is growing due to increasingly easy availability of the data. Traditionally, pedotransfer functions (PTFs) employed for soil hydraulic parameter estimation from other easily available data have used basic soil texture and structure information as inputs. Inclusion of surrogate/supplementary data such as topography and vegetation information has shown some improvement in the PTF's ability to estimate more accurate soil hydraulic parameters. Artificial neural networks (ANNs) are a popular tool for PTF development, and are usually applied across matching spatial scales of inputs and outputs. However, different hydrologic, hydro-climatic, and contaminant transport models require input data at different scales, all of which may not be easily available from existing databases. In such a scenario, it becomes necessary to scale the soil hydraulic parameter values estimated by PTFs to suit the model requirements. Also, uncertainties in the predictions need to be quantified to enable users to gauge the suitability of a particular dataset in their applications. Bayesian Neural Networks (BNNs) inherently provide uncertainty estimates for their outputs due to their utilization of Markov Chain Monte Carlo (MCMC) techniques. In this paper, we present a PTF methodology to estimate soil water retention characteristics built on a Bayesian framework for training of neural networks and utilizing several in situ and remotely sensed datasets jointly. The BNN is also applied across spatial scales to provide fine scale outputs when trained with coarse scale data. Our training data inputs include ground/remotely sensed soil texture, bulk density, elevation, and Leaf Area Index (LAI) at 1 km resolutions, while similar properties measured at a point scale are used as fine scale inputs. The methodology was tested at two different hydro-climatic regions. We also tested the effect of varying the support scale of the training data for the BNNs by sequentially aggregating finer resolution training data to coarser resolutions, and the applicability of the technique to upscaling problems. The BNN outputs are corrected for bias using a non-linear CDF-matching technique. Final results show good promise of the suitability of this Bayesian Neural Network approach for soil hydraulic parameter estimation across spatial scales using ground-, air-, or space-based remotely sensed geophysical parameters. Inclusion of remotely sensed data such as elevation and LAI in addition to in situ soil physical properties improved the estimation capabilities of the BNN-based PTF in certain conditions.

1. Introduction

Remotely sensed data products have become increasingly available for use in earth surface/water resources related research. The National Aeronautic and Space Administration (NASA) operates a number of satellites which provide vital information about the earth's surface processes. For example, the AQUA mission satellite, part of NASA's Earth Observing System (EOS), has six instruments on board which provide data regarding atmospheric and sea surface temperatures, humidity profiles, cloud data, precipitation, radiation balance, terrestrial snow, sea ice, and soil moisture, among others. Of particular interest are the Moderate-Resolution Imaging Spectrometer (MODIS) and the Advanced Microwave Scanning Radiometer for EOS (AMSR-E) instruments. MODIS, which is also included on the TERRA satellite platform, provides products that include the Leaf Area Index (LAI) and other vegetative indices such as the Normalized Difference Vegetative Index (NDVI) and land use land cover. AMSR-E provides a soil moisture data product. With the easy availability of global scale remotely sensed data at different footprints or support scales, new applications and scaling techniques for their utilization to hydrologic problems have been investigated.

Pedotransfer functions (PTFs) have been used to obtain certain complex and expensive soil hydraulic parameters from other available or easily measurable soil properties in the last two

* Corresponding author. Address: Department of Biological and Agricultural Engineering, MS 2117, Texas A&M University, College Station, TX 77843-2117, United States. Tel.: +1 979 676 3427.
E-mail address: raghujana@tamu.edu (R.B. Jana).

0022-1694/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.12.043

decades. Studies have been conducted to develop such transfer functions and test them against available soil properties databases (e.g., Cosby et al., 1984; Rawls et al., 1991; van Genuchten and Leij, 1992; Schaap and Bouten, 1996, 1998; Schaap and Leij, 1998a,b; Pachepsky et al., 1999; Wösten et al., 2001; Sharma et al., 2006; Jana et al., 2007, 2008). Traditionally, soil texture (%sand, %silt, %clay), and bulk density have been the predominant inputs in these PTFs for prediction of soil hydraulic properties. However, usage of supplementary data in addition to texture and bulk density in developing pedotransfer functions has increased within the current decade. It has been shown that addition of topography and vegetation parameters enhance the predictive estimates of soil hydraulic parameters by PTFs to some extent (Pachepsky et al., 2001; Leij et al., 2004; Sharma et al., 2006). However, increasing the number of model input parameters also means increasing the complexity of the model including the inherent uncertainties associated with the input data, and, consequently, the PTF estimates.

Artificial neural networks (ANNs) have been a preferred tool for parameter estimation by PTFs in hydrology (e.g., Schaap and Bouten, 1996; Schaap et al., 1998; Schaap and Leij, 1998a; Sharma et al., 2006; Jana et al., 2007, 2008). However, one major drawback of using a conventional ANN approach is the inherent lack of uncertainty estimates. This, in turn, brings to question the confidence one may place on the accuracy of the ANN predictions. In one study, (Schaap et al., 1998) provided *a posteriori* estimates of the prediction uncertainties by generating multiple realizations of the ANN output. The resultant outputs are then bootstrapped and analyzed to provide confidence levels. Another study by Jana et al. (2008) provided uncertainty estimates of the predicted soil hydraulic properties by using Bayesian Neural Networks which are inherently designed to provide the confidence ranges. Conventionally, the weights of an ANN are obtained during training by iteratively adjusting the values till a single "optimal" set is obtained. However, the ANN methodology is not based on any physical processes underlying the hydrology. Rather, the training of the weights in ANNs is a statistical process that is totally dependent on the input values. Since most hydrologic systems are inherently stochastic, (Kingston et al., 2005), the existence of an "optimal" set of weights is questionable.

Bayesian Neural Networks (BNNs) are designed to overcome this deficiency in conventionally trained ANNs by obtaining a range of weights. Thus, a distribution of predicted values is generated, explicitly accounting for the uncertainty in the predictions. Markov Chain Monte Carlo (MCMC) simulation techniques which form a part of the BNN training also reduce the possibility of the training becoming stuck in local minima and overtraining of the network. As such, BNNs incorporate the best features of conventional ANNs such as their ability to form functional relationships between the inputs and the targets, while addressing some of the drawbacks such as the ability to provide stochastic limits. Thus, BNNs may be thought of as the next generation of neural network models.

While the use of BNNs in the field of water resources modeling is still new, relatively little has been done towards using them for PTF development in the vadose zone. The utility of BNNs has mostly been in surface hydrology applications where it has been used for forecasting river salinity (Kingston et al., 2005), rainfall–runoff (Khan and Coulibaly, 2006), or oxygen demand in estuaries and coastal regions (Borsuk et al., 2001). Most previous PTF studies derive and adopt soil hydraulic parameters at the same spatial scale of input and target data. (Jana et al., 2007, 2008) have demonstrated the usability of ANN- and BNN-based PTFs to estimate soil water contents at a scale different from that of the training data. The objective of this study is to develop and test the Bayesian Neural Network based PTF methodology to derive soil water retention values (at saturation, θ_{0bar}, and field capacity, $\theta_{0.3bar}$) at different scales using ground-based and remotely sensed data at multiple scales which include soil texture, bulk density, elevation and Leaf Area Index (LAI). Remotely sensed data such as brightness temperature have been used to derive soil state variables such as soil moisture (Chang and Islam, 2000; Das and Mohanty, 2006). The novelty of this study lies in the use of such satellite-based measurements of vegetation and elevation in addition to the ground based soil data for the estimation of relatively time-invariant parameters such as soil water retention.

In addition, we also study the dependency of the derived soil water retention values on the scale of the training data. In an earlier work, (Jana et al., 2007) tested the effect of varying the extent from which training data for an artificial neural network is extracted. The study showed that there was no significant improvement in the ANN predictions at the fine scale with increase in the number of training data points at the coarse scale resulting from widening of the spatial extent. In this study we test the effect of changing another component of the scale triplet (Blöschl and Sivapalan, 1995) – the measurement support. The support area is the region over which a measurement is valid. A test of the effect of the change in the measurement support was conducted by sequentially decreasing the support scale of the BNN training data from 1 km to 30 m.

2. Study areas and data

The Bayesian training methodology is tested in two different regions in USA. The first is the Las Cruces Trench site in the Rio Grande basin of New Mexico and the second is in the Southern Great Plain Experiment 1997 (SGP97) hydrology experiment region in Oklahoma. The test sites were chosen so as to provide variety in terrain, land use characteristics, vegetation, soil types and soil distribution patterns. At the same time, sufficient data at the fine scale is also available to validate the BNN predictions. These test beds have been used in previous multiscale PTF studies (Jana et al., 2007, 2008), and as such, provide a reference against which the results of this new study may be compared. A brief description of the test locations is given below for completeness.

2.1. Rio Grande Basin

The Las Cruces Trench (Fig. 1) is located on the New Mexico State University Ranch, roughly 40 miles northeast of the Las Cruces city. The trench is situated in undisturbed soil on a basin slope of Mt. Summerford, near the northern end of the Dona Ana Mountains. The region has a semi-arid climate and vegetation, with generally flat topography. The trench is 26.4 m long, 4.5 m wide and 6 m deep (Wierenga et al., 1991). Using in situ and laboratory methods, Wierenga et al. (1989) developed a comprehensive database of fine-scale soil properties using 594 disturbed soil samples and 594 associated soil cores taken from nine distinct soil layers identified on the north wall of the trench. Samples were also taken from three vertical transects on this wall. The data set included saturated hydraulic conductivity, soil water retention function, particle size distribution, and bulk density for each layer. Only the fifty data points from the top 6-cm layer of the Las Cruces Trench site database is used in this study.

2.2. Little Washita Watershed

Fig. 2 shows the Southern Great Plains 97 (SGP97) experimental region of approximately 40 km by 250 km (10,000 km^2) in the central part of the US Great Plains in the sub-humid environment of Oklahoma (Fig. 2). The region has a moderately rolling topography. Rangeland and pasture dominate the land use with patches of

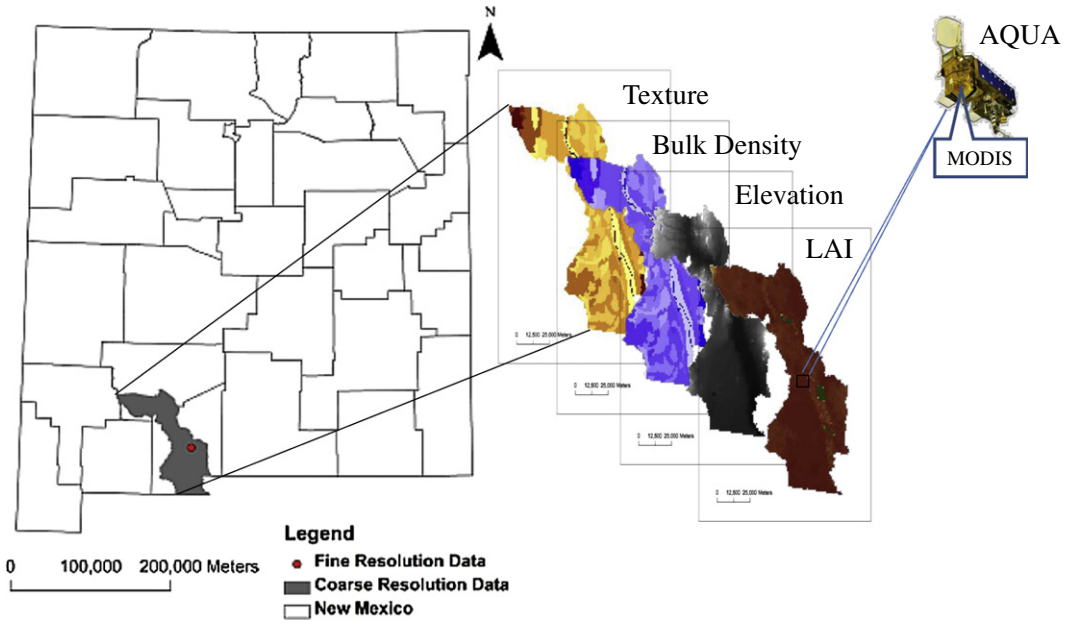

Fig. 1. Rio Grande Basin study area, New Mexico.

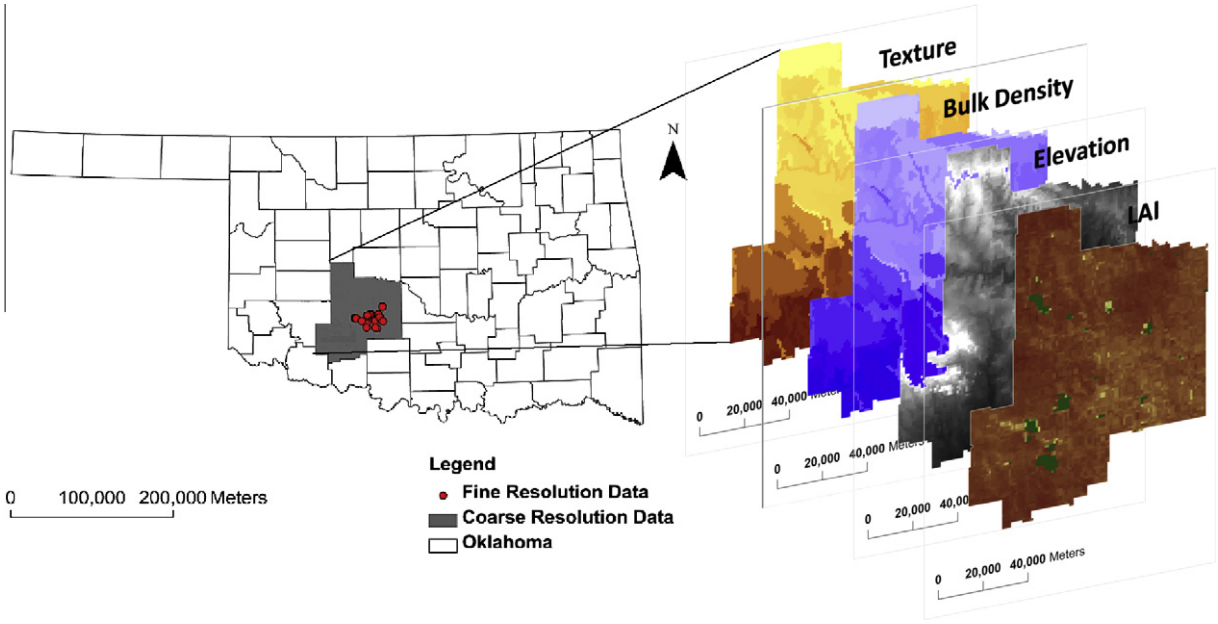

Fig. 2. Little Washita study area, Oklahoma.

winter wheat and other crops. A soil property database of the area developed by Mohanty et al. (2002) provided the fine-scale data at this site.

2.3. Coarse-resolution data

Coarse resolution (1 km) data for both test locations are obtained from a variety of sources. The soil texture, bulk density and water content details were obtained from the Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling (CONUS-SOIL), a database of soil characteristics for the conterminous United States based on the USDA-NRCS State Soil Geographic Database (STATSGO) (Miller and White, 1998). The STATSGO database was developed for use in regional scale models by generalizing soil survey maps where available, and Landsat imagery where soil survey maps were unavailable. STATSGO base maps were compiled state-wise at 1:250,000 scale. The soil physical properties used in this study are the sand, silt, and clay percentages, and the bulk density. The hydraulic parameters are the water content at satiation (θ_{0bar}), and the water content at 1/3 bar ($\theta_{0.3bar}$). While we wished to test the BNN methodology on the entire range of the soil water characteristic, non-availability of water content data at other pressures in the database to train the Bayesian Neural Networks restricted our choices to these two parameters.

Elevation data at the 1 km resolution was obtained from the GTOPO30 global digital elevation model provided by the US Geological Survey (USGS) Earth Resources Observation and Science

(EROS) (http://www.eros.usgs.gov/products/elevation/gtopo30.html). The data is available at a resolution of 30-arc seconds, which corresponds to approximately 1 km grids. For the RGB site training dataset, the elevation ranges between 1131 m and 2675 m, with an average elevation of 1362.49 m. For the LW training dataset, elevation ranges between 304 m and 683 m, with an average elevation of 401.78 m. Vegetation characteristics for the test regions were obtained in the form of the Leaf Area Index (LAI) data product from NASA's MODIS instruments on board the AQUA and TERRA satellite platforms (Myneni et al., 2002). LAI is a measure of the one-sided leaf area per unit ground area. LAI is derived from multiple products such as the surface reflectance, land cover type, and other associated surface characteristic information. It is a dimensionless (m^2/m^2) quantity ranging between 0 and 8. However, when represented in the raster format, the values are stretched between 0 and 255. The LAI dataset is available as 8-day composites. Since the fine-scale data was collected during the month of June 1997, and the MODIS sensor was launched later, for this study, we have selected an 8-day window in mid-June of 2005, so as to correspond with the general time of sampling of the fine scale dataset. At the 1 km resolution, the RGB training data had an average LAI value of 55.21, while at the LW site, an average LAI of 113.21 was observed. At the fine scale, the value of LAI from the coarse pixel corresponding to the data point location is taken as the LAI value.

Coarse scale data is obtained for a region surrounding the locations from where fine resolution data is available, as shown in Figs. 1 and 2. At the Rio Grande Basin (RGB) site, 5580 sets of 1 km resolution data values are used while 6356 sets are used for the Little Washita (LW) region. Different GIS layers of coarse scale data including remotely sensed observations are also shown in Figs. 1 and 2.

3. Multiscale Bayesian Neural Network analysis

Conventionally trained artificial neural networks, as used in most previous PTF applications, form a relationship between the inputs and the targets during the training. If y be the training target and x be the input data, then the relationship between x and y can be described as

$$y = f(x|w) + E \tag{1}$$

where $f(x|w)$ is the functional approximation of the relationship between the input and the target as described by the ANN, w is the vector of weights and biases for the layers of ANN neurons, and E is the error term. Here w is a single set of weights which provide outputs that best match the targets (i.e., least mean square error between outputs and targets). However, many such combinations of input and layer weights could exist which provide best-match outputs.

Unlike conventional ANNs, Bayesian Neural Networks generate a probability distribution of the weights which is dependent on the given input data. From Bayes' theorem,

$$P(w|y,X) = \frac{P(y|w,X)P(w)}{P(y|X)} \tag{2}$$

where X is the input vector $(x_1, x_2, \ldots, x_n)$, $P(y|X) = \int P(y|w,X)P(w)\,dw$, $P(w)$ is the prior distribution of weights, and $P(y|w,X)$ is the likelihood function (Gelman et al., 1995). As described by Kingston et al. (2005), the predictive distribution of y_{n+1} is given by

$$P(y_{n+1}|x_{n+1},y,X) = \int P(y_{n+1}|x_{n+1},w)P(w|y,X)dw \tag{3}$$

The subscript "$n+1$" for x connotes new data that has not been used in the training of the BNN. This integral can be solved by numerical integration using Markov Chain Monte Carlo (MCMC) methods (Neal, 1992).

MCMC methods are used to generate multiple samples from a continuous target density (Bates and Campbell, 2001). The posterior weight distribution is generally complex and difficult to sample from. Hence, a simpler symmetrical distribution is used to generate the weight vectors. This is called the "proposal" distribution and is considered to be locally Gaussian. This proposal distribution depends only upon the weights from the previous iteration in a random walk Markov chain implementation. Arbitrary values are chosen for the weight vector w to start with. A series of values w^* are then proposed by the Markov chain which are accepted with a probability given by

$$\alpha = min\left\{1, \frac{P(y|X,w^*)P(w^*)}{P(y|X,w_{prev})P(w_{prev})}\right\} \tag{4}$$

In the above equation, w_{prev} is the previous value of the weight vector. If w^* is accepted, the previous value w_{prev} is replaced by the proposed value w^* and the procedure is iterated over again. An acceptance rate between 30% and 70% is generally considered to be optimal (Bates and Campbell, 2001). Generating a large number of iterations ensures that the Markov chain is forced to converge to a stationary distribution. At that point, the weight vectors may be considered to have been generated from the posterior distribution itself. Detailed descriptions and discussions of the Metropolis algorithm for the MCMC method used in this study are given by (Gelman et al., 1995), and Kingston et al., 2005). We generated 15,000 Markov chain iteration samples and discarded the first 5000 samples as *burn-in*. This is done to allow the network suitable time to "understand" the relationship between the inputs and the outputs, and attain stability. Thus, 10,000 possible weight combinations, each of which satisfy the neural network's training requirements, are generated.

Coarse scale information, from grids of 1 km resolution, are fed to the BNN for training. The input parameters are the sand, silt and clay percentages, bulk density, elevation and LAI. The training targets are the soil water contents at matric potentials of 0 bar (satiation, θ_{0bar}), and 0.3 bar (field capacity, $\theta_{0.3bar}$). Using the BNNs trained with the coarse-resolution data sets, predictions of soil water contents were made at the point resolution for the corresponding point-scale data sets.

4. Non-linear bias correction

It has been shown by previous studies that a bias can exist between data sets due to difference in measurement techniques, instrument or operator errors, averaging methods, or due to the scale disjoint between the training and simulation datasets used in the BNN (Schaap and Leij, 1998b; Jana et al., 2007, 2008). Since the training of the neural network is done using coarse-scale (1:250,000, or 1 km resolution) data, the BNN model developed is a coarse-scale model. When point scale (1:1) inputs are fed to this model, the predictions obtained for the soil water contents are technically still at the coarser scale. This gives rise to a bias between the BNN-predicted values and the measured values at the point scale. Different hydrological governing processes dictate the soil water contents at different spatial scales. However, the BNN is not based on the physical processes underlying the hydrology. It only forms a relationship between the inputs and the targets based on the data provided for training, and thus cannot inherently account for the support scale disparity between the training and simulation datasets. Hence, a suitable bias correction technique needs to be applied to the predicted water content values for adjustment to the target scale.

A linear bias correction was applied by Jana et al. (2007) to provide a proportional shifting effect to the ANN predicted values that brings the mean of the ANN predicted values closer to that of the measured values. Linear bias correction, however, only accounts for the first moment (mean) of the data. No correction is applied to the second moment (spread) of the values. Also, it was observed that there is a smoothing effect on the soil water content distribution since only the systematic error is accounted for in this technique. This smoothing further decreases the variance of the predicted soil water content values. Using a linear bias correction technique would provide a good estimate for the mean of the entire data set (i.e., the effective soil water content values at the field scale). Further, parametric scaling being a non-linear process, application of a linear bias correction can be successful only to a certain degree. In a subsequent study, (Jana et al., 2008) applied a non-linear bias correction to the BNN predictions by matching the cumulative distribution functions (CDFs) of the BNN predicted and the target (measured) values. The CDF-matching technique is based on the idea of obtaining the predicted parameter values corresponding to the probability of values on the CDF of the target parameter (Calheiros and Zawadzki, 1987; Atlas et al., 1990; Reichle and Koster, 2004; Ines and Hansen, 2006). After ascertaining the type of distribution (e.g., normal, log-normal, gamma) of the parameters by statistical tests, CDFs are obtained for the target and predicted values for each parameter based on the type of distribution they follow. For each predicted soil water content value, there exists a particular probability of occurrence. Similarly, for a particular probability of occurrence, there exists a corresponding target soil water content value. CDF matching is achieved by forcing the predicted soil water content value with a particular occurrence probability towards the corresponding target soil water content value. For a normally distributed parameter, the CDF is given by

$$\mathrm{CDF}(\theta_i;\mu,\sigma) = \int_{-\infty}^{\theta_i} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\left(\frac{(\theta_i-\mu)^z}{2\sigma^2}\right)\right) \quad (5)$$

Here, θ_i is the water content value at which the CDF is calculated, μ is the mean and σ is the standard deviation of the soil water content values. In order to effectively scale the BNN-model predicted values to the measured dataset, it is required to find θ_i^{pred} such that

$$\mathrm{CDF}(\theta_i^{pred}) = \mathrm{CDF}(\theta_i^{target}) \quad (6)$$

This is achieved by computing the inverse of the cumulative probability values for the calibration dataset, but with the mean and standard deviation values of the target distribution. The inverse is the value of the soil water content that corresponds to a particular probability. This procedure effectively scales the distribution of the neural network predicted calibration dataset to approximate that of the target values. A schematic representation of the CDF-matching technique is shown in Fig. 3. It must be noted that the bias correction is applied not only to the mean BNN-predicted values, but to the entire band of uncertainty as well. This means that all the outputs from the MCMC algorithm are accounted for in the bias correction too. Since the bias correction is not the main focus of this study, the derivation and/or provenance of the method is not discussed in detail here. More detailed discussions about the method may be found elsewhere (e.g., Calheiros and Zawadzki, 1987; Atlas et al., 1990; Reichle and Koster, 2004; Ines and Hansen, 2006).

5. Multiple support scale analysis

(Jana et al., 2007) studied the effect of changing the extent from which coarse-resolution data is chosen to train the ANN. Extent is one of the components of the scale triplet – extent, spacing, and support – as described by Blöschl and Sivapalan (1995)). It was reported by Jana et al. (2007) that no significant change in the ANN prediction capability was found due to such an increase in the number of coarse-resolution data points beyond a certain extent. In this study, we tested the effect of changing the support scale of the data used for training the multiscale BNNs. Soil physical and hydraulic property data for training of the BNN were obtained at resolutions of 30 m, 90 m, 270 m, and 810 m from within the coarse resolution training data extent for the Rio Grande Basin study area (Fig. 1). The 30 m data was primarily obtained from

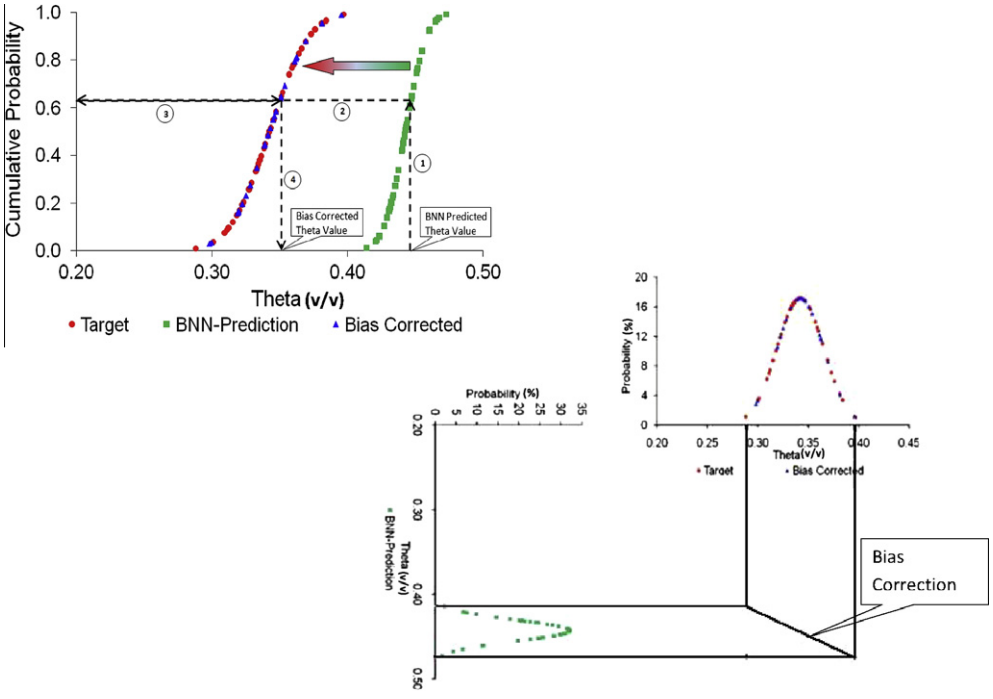

Fig. 3. Non-linear bias correction.

the Soil Survey Geographic (SSURGO) database compiled by the United States Department of Agriculture – Natural Resources Conservation Service (USDA-NRCS) (http://soildatamart.nrcs.usda.gov). This public domain database contains geo-referenced spatial and attribute data for soils compiled from soil surveys. These surveys cover large spatial extents (usually county-wide) and the soil property data are based on soil type rather than the spatial location. The SSURGO database was created by field methods, using observations along soil delineation boundaries and traverses, and determining map unit composition by field transects. Aerial photographs are interpreted and used as the field map base. Multiple readings are taken for each property within each map unit. The number of readings taken differs between map units that are based on factors such as the size of the soil polygon, the variation in topography and change in vegetation, among others. Low, high, and representative values for each soil physical and/or hydraulic property are provided in the database for each soil type/map unit at scales ranging between 1:12,000 and 1:31,680 (http://www.nrcs.usda.gov/technical/soils/soilfact.html). In this study, the representative values for the soil texture (sand, silt and clay percentages), bulk density, elevation, and soil water contents were obtained from 1:24,000 resolution SSURGO soil maps in a gridded format with a resolution of 30 m. The LAI values were re-sampled from the MODIS data described earlier.

The 3×3 grids of parametric data at the 30 m resolution are generalized to 90 m resolution using the mean aggregation feature in ArcMap™ software by ESRI®. The 270 m resolution data was obtained by aggregating 3×3 grids of 90 m data, and so on. Since the soil property data are in grid format, changing the support area of the parameter causes a corresponding change in the spacing too. Hence, in reality, two components of the scale triplet are being modified here.

The BNN methodology, along with the non-linear bias correction technique, was applied with the BNN being trained with data at each coarse resolution. Predictions of the soil water contents at saturation and field capacity at the point (1:1) resolution were obtained, and corrected for bias by the CDF mapping method.

6. Upscaling study

In order to investigate the multi-scale nature of the Bayesian Neural Networks, a study was conducted to upscale the soil water retention parameters from the 30 m resolution to the 1 km resolution at the Little Washita Watershed site. Training data at the 30 m resolution for this study consisted of the soil texture (sand, silt and clay percentages), and the bulk density from the USDA-NRCS SSURGO database. Elevation data obtained at the 30 m resolution from the United States Geological Survey's (USGS) National Elevation Dataset was also used as a training input. Training targets were the water content at satiation (θ_{0bar}) and the water content at 1/3 bar ($\theta_{0.3bar}$). Simulation input data for soil texture and bulk density at the 1 km resolution were obtained from the STATSGO database, while the elevation data were from the GTOPO30 global digital elevation model mentioned earlier.

Twelve coarse resolution (1 km) pixels were taken from the LW region for this study. Fine (30 m) resolution training input and target data from within these areas were taken from the SSURGO and National Elevation databases. The BNN methodology was applied with the networks being trained at the fine scale. Predictions of the soil water contents at saturation and field capacity at the coarse (1 km) resolution were then obtained.

7. Results and discussion

Data from multiple scales, from satellite-based remote sensing footprints to ground-based point scale measurements, were applied in a multiscale Bayesian Neural Network methodology to obtain fine-scale soil water content values after being trained with coarse scale data. The resultant outputs from the BNN were plotted along with their respective expected values (Fig. 4). The error bars, obtained from the Markov chain Monte Carlo simulations, represent the uncertainty in the neural network predictions. BNNs, as mentioned earlier, generate a distribution of weights instead of a single set. The uncertainty band (error bars) show the limits to which the predictions could have varied based on the combination

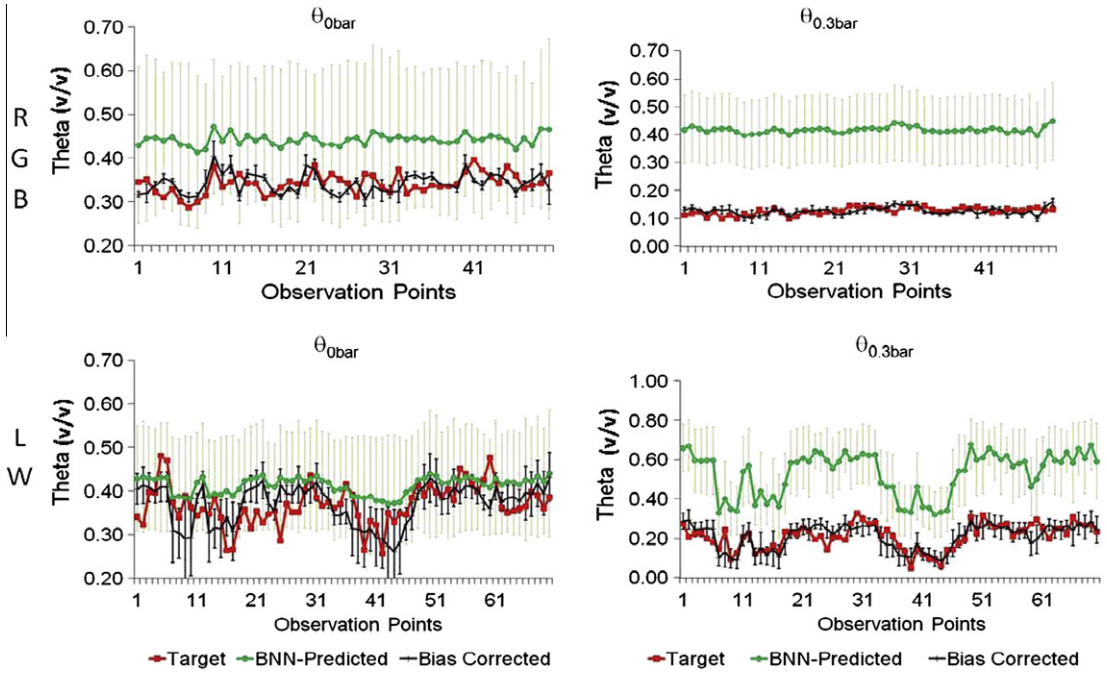

Fig. 4. Target, BNN-predicted, and bias corrected soil water content values.

of weights used. The final predicted soil water content value is an average of all such possible values from the 1000 Monte Carlo simulations. Fifty sets of point scale inputs/outputs are available for each parameter at the Rio Grande Basin (RGB), while in the Little Washita Watershed (LW), we have seventy values measured at the point scale. The comparative statistics for the target and predicted parameters at the two sites are given in Table 1. As expected, BNN predictions at both test sites and for both water content parameters were biased from the expected target values. As previously noted, this bias is an artifact of the scale disjoint between the training and simulation data sets, and is eliminated by application of the non-linear bias correction technique.

As in the study by Jana et al. (2008), it is apparent that the outputs at the RGB site showed less variations than those at the LW site (Fig. 4). This relative invariance is attributed to the small range in variations of the corresponding inputs at the fine scale. Descriptive statistics for the soil physical properties at the coarse and fine scales from the two test sites are presented in Table 2. There is a significant difference in the amount of variation of the texture between the coarse and fine scales at RGB. The fine-scale soil is predominantly sandy and almost uniform. Further, the Las Cruces trench has a dimension of 26 m. This means that all the observation points at this site lie within one coarse scale pixel, thus showing invariance in topography and vegetation too. Like any other model, neural network outputs too are dependent on the quality of input data. The invariance in the inputs at the fine scale is reflected in the soil water content estimates produced by the BNN. In contrast, the LW data are spread over a much wider area (approximately 10,000 km^2). Fine scale observation points lie in different coarse scale pixels. The variability in the soil physical properties are also comparable at the two scales (Table 2). This results in a better estimation of the water content values at the LW site as compared to the RGB site, as can be seen from the R values in Table 1.

Table 1
Descriptive and comparative statistics of target, BNN-predicted, and Bias Corrected soil water content values.

	θ_{0bar} (v/v)			$\theta_{0.3bar}$ (v/v)		
	Target	BNN-predicted	Bias corrected	Target	BNN-predicted	Bias corrected
RGB						
Mean	0.342	0.443	0.342	0.127	0.418	0.127
Std. dev.	0.024	0.012	0.024	0.013	0.011	0.013
R		0.328	0.333		0.223	0.257
RMSE		0.103	0.026		0.246	0.015
LW						
Mean	0.370	0.414	0.370	0.210	0.533	0.210
Std. dev.	0.046	0.019	0.046	0.064	0.111	0.064
R		0.416	0.421		0.762	0.764
RMSE		0.061	0.050		0.331	0.044

RGB: Rio Grande Basin; LW: Little Washita; BNN: Bayesian Neural Network; *R*: Correlation coefficient; RMSE: Root mean square error.

From Table 1, for the RGB region, we see that the BNN estimation of the water content at saturation (θ_{0bar}) is slightly better than that for the field capacity ($\theta_{0.3bar}$). A correlation coefficient value of 0.333 is observed for the θ_{0bar} value, while the value is 0.257 for $\theta_{0.3bar}$. Earlier studies (Jana et al., 2008) have shown similar correlation values. No significant improvement or deterioration in the BNN's prediction capabilities were seen for the RGB site although the resolution of the training data is much coarser (1 km) in this study as compared to the earlier study (30 m), and considering that topography and vegetation have been added to the training factors. We suggest that this is due to the general conditions of the RGB site. This site exhibited more uniformity in soil texture, topography, and vegetation with large spatial correlation length scale when compared to the LW site.

For the LW site, the R value for the θ_{0bar}, at 0.421, remains similar to the earlier study (Jana et al., 2008). The correlation value for the $\theta_{0.3bar}$, however, is greatly improved (0.764). This improvement in the correlation can be attributed to the use of the topographic information in this model, as against using only texture and bulk density. Field capacity ($\theta_{0.3bar}$) is defined as the available water content in the soil after gravity draining. Drainage by gravity, especially from the wet end of the soil water characteristic, is greatly influenced by the topography. By including the elevation in the BNN model, a better estimate of the variability is obtained for this parameter as compared to the earlier study (Jana et al., 2008).

Kolmogorov–Smirnov tests showed that the BNN predicted and field measured fine scale water content values are normally distributed. Hence, the CDF matching algorithm for non-linear bias correction could be carried out using the normal CDF equation (Eq. (5)). Normal CDFs were plotted for the target and predicted θ values for both the regions (Fig. 5). It can be seen that the CDFs of the model predicted soil water content values and those of the measured (target) values do not match for either region. Probability Density Functions (PDFs) are also plotted for the target and BNN-predicted θ values for both the regions (Fig. 6). The difference in the mean and spread of the target and BNN-predicted distributions is apparent here. The means of the BNN-predicted values are consistently higher than the measured values for each water content at either location. The BNN-predicted θ values were randomly split into two halves, one half for model calibration and the other for validation of the bias correction scheme. The cumulative probabilities for each point value are computed using the mean and standard deviation of the calibration dataset. The calibrated (bias corrected) soil water content CDF values (Fig. 5) now follow the target CDF closely.

To test the calibration of the bias correction scheme, the remaining half of the neural network predicted soil water content values (the validation dataset) is used. The calibration is found to be correct as the distributions of validation data are aligned with the target distributions (Fig. 5). This suggests that our bias correction scheme for the predicted θ values approximates the target values well. In concurrence with the CDFs, the PDFs of the calibration

Table 2
Bayesian Neural Network input parameters at coarse and fine scales.

	Sand (%)		Silt (%)		Clay (%)		Bulk density (–)		Elevation (m)		LAI (m^2/m^2)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
RGB												
Coarse scale	54.641	13.973	30.603	7.796	14.752	7.751	1.520	0.174	1362.490	167.060	55.210	19.874
Fine scale	81.458	1.666	9.760	1.170	8.780	1.200	1.660	0.050	1200.000	0.000	250.000	0.000
LW												
Coarse scale	41.969	17.280	43.875	14.034	14.226	5.726	1.437	0.015	401.783	45.431	113.210	15.272
Fine scale	51.913	21.113	33.606	16.412	14.482	6.040	1.396	0.099	391.000	25.213	12.400	2.815

RGB: Rio Grande Basin; LW: Little Washita; LAI: Leaf Area Index.

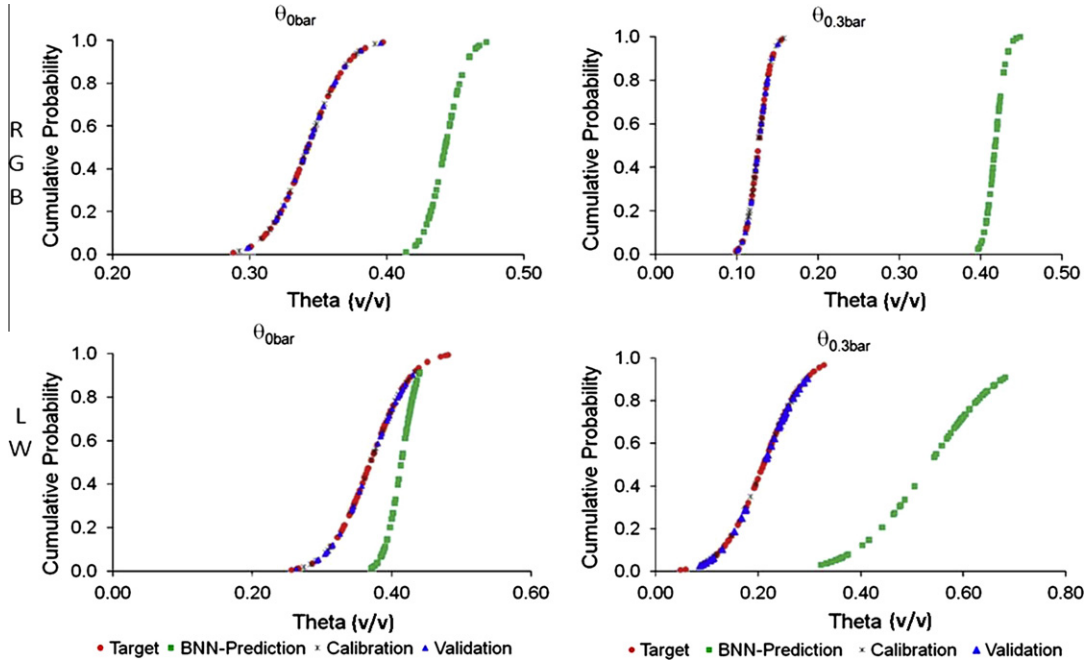

Fig. 5. Cumulative probability distributions of soil water content values.

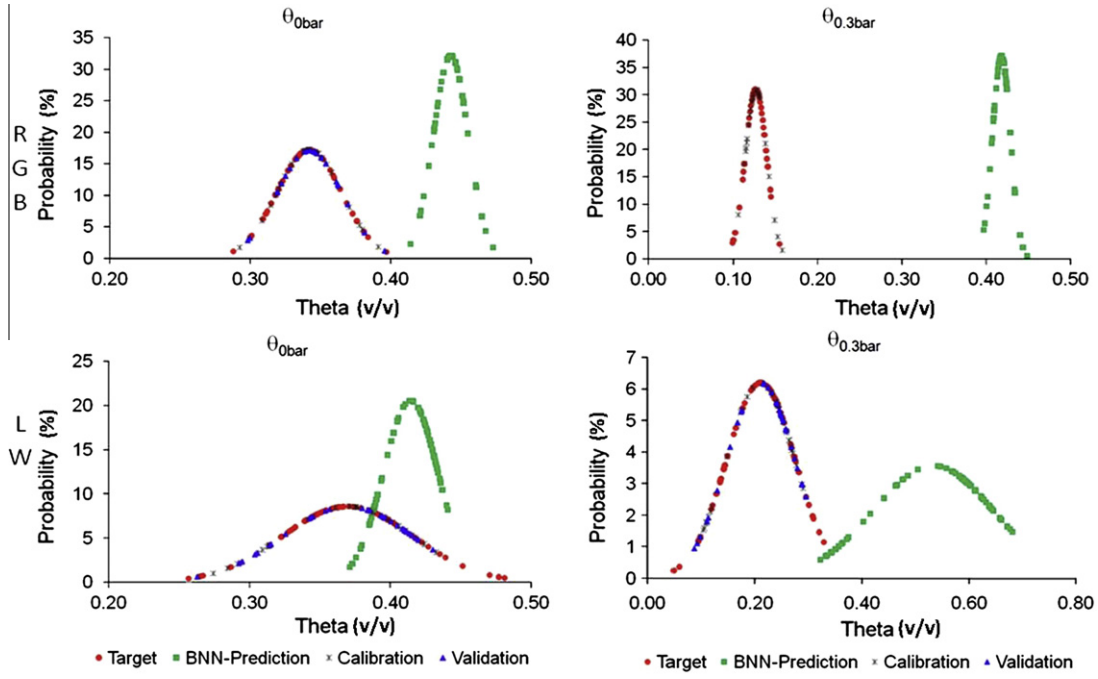

Fig. 6. Probability distributions of soil water content values.

and validation datasets plot on top of the target PDF (Fig. 6). Scatter plots of the target, BNN-Prediction, calibration and validation values for both water contents at both sites are also plotted (Fig. 7) to provide a sense of how the bias correction procedure shifts the BNN predictions closer to the targets.

From Fig. 4 it can be seen that the variability of the target values is largely approximated by the non-linear bias-corrected θ values. However, point-to-point matching of values is still not obtained. The bias-corrected values are being sampled from the same distribution as the target values, but at a different probability. Uncertainties introduced into the observations of any point-scale data due to measurement and operator errors, and other influencing factors such as the presence of macropores or roots debris have not been considered here. These factors make the approximation of the particular point values dictated by stochastic natural processes a near-impossible task, given the current inputs. Since the individual values may also be considered as being sampled from a distribution (to cover the uncertainties), the point-to-point match of the values is neither practically achievable nor really necessary. In other words, if we select values from the normal

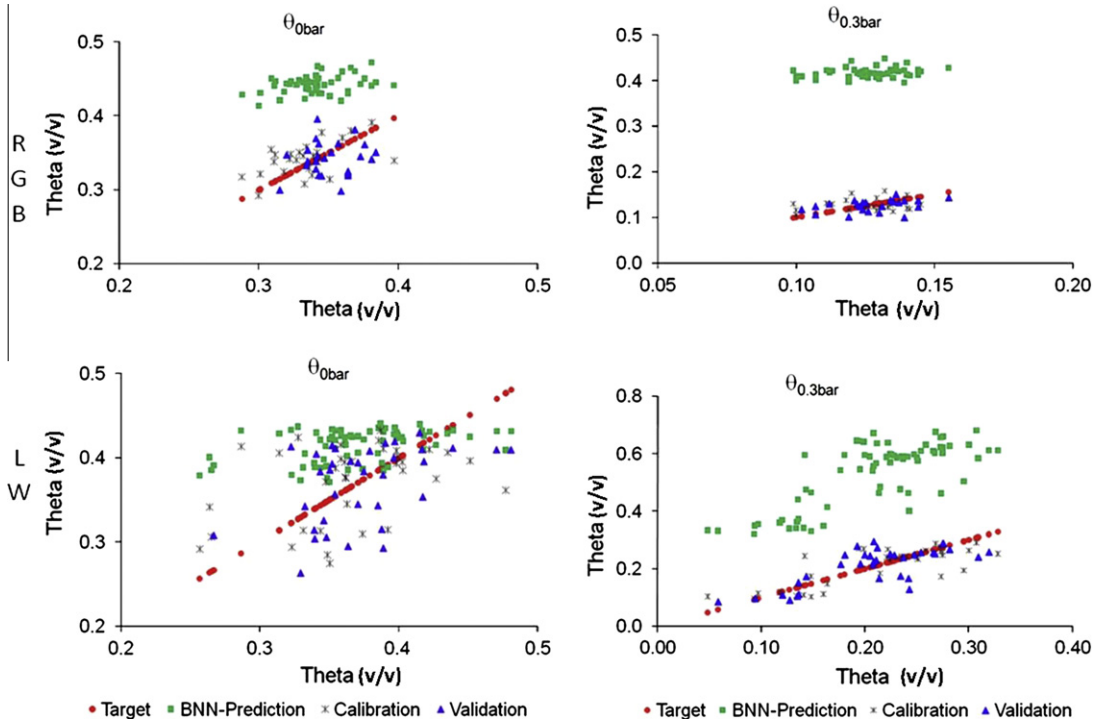

Fig. 7. Scatter plots of soil water content values.

distribution of the bias-corrected values for the exact probabilities as those of the target values, we would get a much better match at each point. Further, uncertainties in fine-scale data due to factors such as measurement and/or operator errors and presence of macropores or organic debris make it nearly impossible to precisely estimate the observed values. The non-linear bias correction approach provides θ values for any probability, which is not possible by using a linear bias correction. Matching of the distributions, along with the Bayesian nature of the neural network model, ensures that the above-mentioned uncertainties are incorporated into the estimation scheme.

The mean, standard deviation, root mean square error (RMSE) and average bias correction applied to the predicted θ values from BNNs trained with data from different coarse resolutions (support scales) in the multiple support scale analysis are given in Table 3. The average bias corrections applied at different resolutions are graphed in Fig. 8. It is found that a fourth order polynomial fits the average bias correction curve with an R^2 of 1 for both θ values. We have used a non-linear bias correction based on the assumption that scaling is a non-linear process. The findings shown in Fig. 8 support this assumption. If the effect of scale on the BNN prediction were to be linear, we would find that the average bias correction rises linearly with increase in resolution. Further, the non-monotonic nature of the curve may be an indicator of the fractal/self-similar nature of the hydraulic property itself.

The target (from STATSGO) and BNN-predicted values for the water content values at the 1 km resolution from the upscaling study are plotted in Fig. 9. It can be seen that the BNN-predicted water content values are close to the target values at all locations. The target θ_{0bar} values fall within the band of uncertainty at all locations, while the target $\theta_{0.3bar}$ values lie within the uncertainty band at most (10/12) locations. Comparative statistics between the target and BNN-predicted values are given in Table 4. It can be seen that the correlation values are much higher for the upscaling study than for the downscaling case, and the RMSE is much smaller for both water content values. This shows that the upscaling was successful.

It may be observed that the bias correction was not applied in the upscaling study. It was not considered necessary since upscaling is an interpolative exercise for the BNN. Neural networks perform better at interpolation than they do at extrapolation. This inherent property of BNN's means that they are naturally better

Table 3
Average bias correction necessary for different scales of training data.

θ_{0bar} (v/v)	Target	30 m		90 m		270 m		810 m		1 km	
		Pred	BC	Pred	BC	Pred	BC	Pred	BC	Pred	BC
RGB											
Mean	0.342	0.482	0.342	0.447	0.342	0.486	0.342	0.424	0.342	0.443	0.342
Std. dev.	0.023	0.006	0.023	0.012	0.023	0.016	0.023	0.011	0.023	0.012	0.024
Avg. BC			−0.140		−0.105		−0.144		−0.082		−0.100
LW											
Mean	0.127	0.405	0.127	0.582	0.127	0.537	0.127	0.570	0.127	0.418	0.127
Std. dev.	0.013	0.003	0.013	0.012	0.013	0.012	0.013	0.014	0.013	0.011	0.013
Avg. BC			−0.278		−0.455		−0.410		−0.443		−0.012

θ: Soil water content; Pred: BNN Predicted value; BC: Bias-corrected value; Avg. BC: Average bias correction applied.

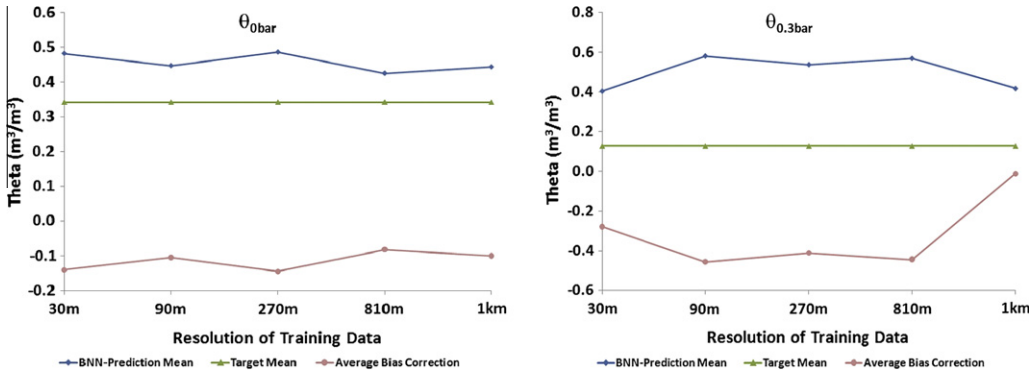

Fig. 8. Average bias correction necessary for different scales of training data.

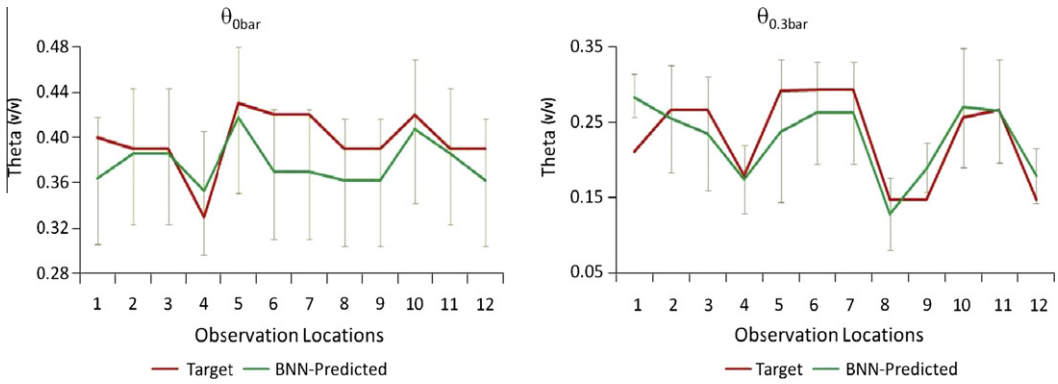

Fig. 9. Target and BNN-predicted soil water content values at 1 km resolution from upscaling study.

Table 4
Descriptive and comparative statistics of target, and BNN-predicted soil water content values at 1 km resolution from upscaling study.

	θ_{0bar} (v/v)		$\theta_{0.3bar}$ (v/v)	
	Target	BNN-predicted	Target	BNN-predicted
Mean	0.397	0.377	0.230	0.228
Std. dev.	0.026	0.020	0.060	0.049
R		0.598		0.802
RMSE		0.028		0.035

at upscaling exercises than downscaling, where an additional bias correction step would be necessary to account for the scale disjoint.

Remotely sensed data products are becoming increasingly easy to obtain and newer applications are being developed. The quality of remotely sensed data is improving all the time. However, at the present time, the resolutions at which such data are available are still rather coarse. This results in our having to resample the coarse resolution pixels to finer resolutions. Such simple resampling methods are not a substitute to rigorous scaling techniques, and introduce errors in parametric values. Empirical PTFs based on statistical techniques such as neural networks are inherently site-specific as they need to be trained to recognize the patterns particular to that site. So a network would need to be trained fresh if estimating soil hydraulic parameters at a site outside the area from which the coarse scale data is provided. Alternatively, the network would need to be trained with a very comprehensive dataset encompassing all possible variability in soil physical properties in order to be considered as a generic pedotransfer function application. Further, using this methodology, a few representative measurements are necessary at the fine scale for the bias correction procedure when downscaling the soil water retention parameters. Using these measurements, an estimate of the amount of correction to be applied can be obtained which can then be used for the entire extent of interest. However, this step is not necessary for upscaling the water retention parameters.

8. Conclusions

Using coarse scale soil properties data from ground-based and remote sensing platforms up to 1 km resolution and point scale measured soil properties data, we have shown that a Bayesian Neural Network can be applied across spatial scales to approximate fine-scale soil hydraulic properties. The study was conducted for two regions which are greatly different in soil, topography, vegetation, climate, and in the spatial extent from which the point data was collected. It has been shown that the BNN predictions are superior when the training data covers a larger region. This is due to the large scale heterogeneity encompassed in the training process. Using remotely sensed topographical and vegetation parameters in the training showed improvements in the Little Washita region where the point scale inputs are from a widely dispersed region. On the contrary, no significant improvement was found by the inclusion of additional parameters in the BNN training at the Rio Grande Basin site where it is limited to a small trench/plot. A marked improvement in predicted $\theta_{0.3bar}$ values was found at the LW site, and is attributed to the inclusion of the additional factors such as topography which represents the gravity draining component. As expected, the scale disjoint between training and simulation data made the application of a bias correction

procedure necessary. The non-linear technique of CDF matching was used to obtain the bias correction. The average bias correction necessary to be applied was found to vary as a fourth order polynomial based on the resolution of the training data. BNNs also easily provide an estimate of the uncertainties involved in the prediction scheme. Traditional ANN methods would involve a few further steps to obtain an *a posteriori* estimate of the same uncertainties. Overall, the Bayesian Neural Network, coupled with a non-linear bias correction scheme, appears to work well for estimation of soil hydraulic properties at a fine scale from data at coarser scales (downscaling). The Bayesian Neural Networks performed better at upscaling of water retention parameters than at downscaling, due to the inherent properties of the networks. However, this study also underlines the necessity of better input and training data using remote sensing techniques for better predictions, as also the fact that no single model is applicable at all geographical locations. Also, the currently available coarse scale data allows for testing of this approach only at the wet end of the soil water characteristic. If more points on the soil water characteristic curve are known at the coarse scale, then a comprehensive test of the methodology would be possible.

Acknowledgements

We acknowledge the partial support of Los Alamos National Lab, NASA Earth System Science Fellowship (NNX06AF95H), National Science Foundation (CMG/DMS Grant 0621113) and NASA THP (Grant 35410) grants. The Los Alamos portion of this work was supported by the Los Alamos National Laboratory, Laboratory Directed Research and Development Project "High-Resolution Physically-Based Model of Semi-Arid River Basin Hydrology" and in collaboration with SAHRA (Sustainability of semi-Arid Hydrology and Riparian Areas) under the STC Program of the National Science Foundation under Agreement No. EAR-9876800.

References

Atlas, D., Rosenfeld, D., Wolff, D.B., 1990. Climatologically tuned reflectivity-rain rate relations and links to area-time integrals. J. Appl. Meteorol. 29, 1120–1135.

Bates, B.C., Campbell, E.P., 2001. A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall–runoff modeling. Water Resour. Res. 37 (4), 937–947.

Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling – a review. Hydrol. Process 9, 251–290.

Borsuk, M.E., Higdon, D., Stow, C.A., Reckhow, K.H., 2001. A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal areas. Ecol. Model. 143 (3), 165–181. doi:10.1016/S0304-3800(01)00328-3.

Calheiros, R.V., Zawadzki, I.L., 1987. Reflectivity rain-rate relationships for radar hydrology in Brazil. J. Climate Appl. Meteorol. 26, 118–132.

Chang, D.-H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sens. Environ. 74, 534–544.

Cosby, B.J., Hornberger, G.M., Clapp, R.B., Ginn, T.R., 1984. A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resour. Res. 20 (6), 682–690.

Das, N.N., Mohanty, B.P., 2006. Root zone soil moisture assessment using remote sensing and vadose zone modeling. Vadose Zone J. 5 (1), 296–307. doi:10.2136/Vzj2005.0033.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. CRC Press, Boca Raton, Fla.

Ines, A.V.M., Hansen, J.W., 2006. Bias correction of daily GCM rainfall for crop simulation studies. Agric. Forest Meteorol. 138, 44–53.

Jana, R.B., Mohanty, B.P., Springer, E.P., 2007. Multiscale pedotransfer functions for soil water retention. Vadose Zone J. 6 (4), 868–878.

Jana, R.B., Mohanty, B.P., Springer, E.P., 2008. Multiscale Bayesian neural networks for soil water content estimation. Water Resources Research 44 (8), W08408. doi:10.1029/2008wr006879.

Khan, M.S., Coulibaly, P., 2006. Bayesian neural network for rainfall–runoff modeling. Water Resour. Res. 42 (7). doi:10.1029/2005wr003971.

Kingston, G.B., Lambert, M.F., Maier, H.R., 2005. Bayesian training of artificial neural networks used for water resources modeling. Water Resour. Res. 41, W12409. 12410.11029/12005WR004152.

Leij, F.J., Romano, N., Palladino, M., Schaap, M.G., Coppola, A., 2004. Topographical attributes to predict soil hydraulic properties along a hillslope transect. Water Resour. Res. 40 (2). doi:10.1029/2002wr001641.

Miller, D.A., White, R.A., 1998. A conterminous United States multi-layer soil characteristics data set for regional climate and hydrology modeling. Earth Interact. 2. <http://EarthInteractions.org>.

Mohanty, B.P., Shouse, P.J., Miller, D.A., van Genuchten, M.T., 2002. Soil property database: Southern Great Plains 1997 hydrology experiment. Water Resour. Res. 38 (5). doi:10.1029/2000wr000076 (Artn 1047).

Myneni, R.B., Hoffman, S., Knyazikhin, Y., Privette, J.L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G.R., Lotsch, A., Friedl, M., Morisette, J.T., Votava, P., Nemani, R.R., Running, S.W., 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. Remote Sens. Environ. 83 (1–2), 214–231.

Neal, R.M., 1992. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report CRG-TR-92-1, Dept. of Computer Science, University of Toronto, 21 pages.

Pachepsky, Y.A., Rawls, W.J., Timlin, D.J., 1999. The current status of pedotransfer functions, their accuracy, reliability, and utility in field- and regional-scale modelling. In: Corwin, D.L., Loague, K., Ellsworth T.R., (Eds.), Assessment of Non-Point Source Pollution in the Vadose Zone, American Geophysical Union, Washington, DC, pp. 223–234.

Pachepsky, Y.A., Timlin, D.J., Rawls, W.J., 2001. Soil water retention as related to topographic variables. Soil Sci. Soc. Am. J. 65, 1787–1795.

Rawls, W.J., Gish, T.J., Brakensiek, D.L., 1991. Estimating soil water retention from soil physical properties and characteristics. Adv. Soil Sci. 16, 213–234.

Reichle, R.H., Koster, R.D., 2004. Bias reduction in short records of satellite soil moisture. Geophys. Res. Lett. 31, L19501. doi:10.1029/2004GL020938.

Schaap, M.G., Bouten, W., 1996. Modeling water retention curves of sandy soils using neural networks. Water Resour. Res. 32 (10), 3033–3040.

Schaap, M.G., Leij, F.J., 1998a. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil Tillage Res. 47 (1-2), 37–42.

Schaap, M.G., Leij, F.J., 1998b. Database-related accuracy and uncertainty of pedotransfer functions. Soil Sci. 163 (10), 765–779.

Schaap, M.G., Leij, F.J., van Genuchten, M. Th., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62, 847–855.

Sharma, S.K., Mohanty, B.P., Zhu, J., 2006. Including topography and vegetation attributes for developing pedo transfer functions in southern great plains of USA. Soil Sci. Soc. Am. J. 70, 1430–1440.

van Genuchten, M.T., Leij, F.J., 1992. On estimating the hydraulic properties of unsaturated soils, in indirect methods for estimating the hydraulic properties of unsaturated soils. In: Proceedings of the International Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils, van Genuchten, M.T., Leij, F.J., Lund, L.J., (Eds.), Department of Soil and Environmental Sciences, University of California, Riverside, California, pp. 1–14.

Wierenga, P.J., Hudson, D., Vinson, J., Nash, M., Toorman, A., Hills, R.G., 1989. Soil Physical Properties at the Las Cruces Trench Site. NUREG/CR-5441, US Nuclear Regulatory Commission.

Wierenga, P.J., Hills, R.G., Hudson, D.B., 1991. The Las Cruces Trench Site: Experimental results and one dimensional flow predictions. Water Resour. Res. 27, 2695–2705.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123–150.