Multiscale Bayesian neural networks for soil water content estimation

Raghavendra B. Jana,[1] Binayak P. Mohanty,[1] and Everett P. Springer[2]

Received 29 January 2008; revised 14 May 2008; accepted 4 June 2008; published 5 August 2008.

[1] Artificial neural networks (ANN) have been used for some time now to estimate soil hydraulic parameters from other available or more easily measurable soil properties. However, most such uses of ANNs as pedotransfer functions (PTFs) have been at matching spatial scales (1:1) of inputs and outputs. This approach assumes that the outputs are only required at the same scale as the input data. Unfortunately, this is rarely true. Different hydrologic, hydroclimatic, and contaminant transport models require soil hydraulic parameter data at different spatial scales, depending upon their grid sizes. While conventional (deterministic) ANNs have been traditionally used in these studies, the use of Bayesian training of ANNs is a more recent development. In this paper, we develop a Bayesian framework to derive soil water retention function including its uncertainty at the point or local scale using PTFs trained with coarser-scale Soil Survey Geographic (SSURGO)-based soil data. The approach includes an ANN trained with Bayesian techniques as a PTF tool with training and validation data collected across spatial extents (scales) in two different regions in the United States. The two study areas include the Las Cruces Trench site in the Rio Grande basin of New Mexico, and the Southern Great Plains 1997 (SGP97) hydrology experimental region in Oklahoma. Each region-specific Bayesian ANN is trained using soil texture and bulk density data from the SSURGO database (scale 1:24,000), and predictions of the soil water contents at different pressure heads with point scale data (1:1) inputs are made. The resulting outputs are corrected for bias using both linear and nonlinear correction techniques. The results show good agreement between the soil water content values measured at the point scale and those predicted by the Bayesian ANN-based PTFs for both the study sites. Overall, Bayesian ANNs coupled with nonlinear bias correction are found to be very suitable tools for deriving soil hydraulic parameters at the local/fine scale from soil physical properties at coarser-scale and across different spatial extents. This approach could potentially be used for soil hydraulic properties estimation and downscaling.

Citation: Jana, R. B., B. P. Mohanty, and E. P. Springer (2008), Multiscale Bayesian neural networks for soil water content estimation, *Water Resour. Res.*, *44*, W08408, doi:10.1029/2008WR006879.

1. Introduction

[2] Point and nonpoint source contaminant transport models as well as global- and regional-scale circulation models for hydrologic and climate forecasting use soil hydrologic models. The prediction accuracy of these models depends largely on the quality of the input parameters, along with the accurate depiction of field realities such as boundary conditions. However, collection of these required soil hydraulic parameters by direct measurement at any model grid scale is expensive and time consuming. Additionally, in order to capture the variability of these spatially distributed parameters, a large number of samples need to be collected. All these factors make the direct measurement of the soil hydraulic parameters matching to the model grids highly impractical.

[3] The use of pedotransfer functions (PTFs) to obtain the required soil hydraulic parameters from other available or easily measurable soil properties has been advocated in the last two decades. A number of studies have been carried out in the recent past to develop such transfer functions and test them against available soil properties databases [e.g., *Rawls et al.*, 1991; *van Genuchten and Leij*, 1992; *Schaap et al.*, 1998; *Pachepsky et al.*, 1999; *Wösten et al.*, 2001; *Sharma et al.*, 2006, *Jana et al.*, 2007]. Soil texture (sand, silt, and clay percentages), organic carbon, and bulk density have been used to a large extent for prediction of soil hydraulic properties. *Schaap et al.* [1998] showed that the use of detailed particle-size distributions could increase the accuracy of soil hydraulic parameters predictions as compared with the predictions using soil textural class alone as inputs [*Clapp and Hornberger*, 1978]. However, such detailed particle-size data are not easily available in all cases. Additional parameters, such as topographic features and vegetation, have rarely been used in developing PTFs

[1]Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA.
[2]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

0043-1397/08/2008WR006879$09.00

[*Wösten et al.*, 2001]. Recently, *Pachepsky et al.* [2001], *Leij et al.* [2004] and *Sharma et al.* [2006] included certain available topographical and vegetation attributes in addition to soil physical parameters for developing PTFs. While the inclusion of more input parameters for the PTFs showed some improvement in the performance of the transfer function models, it was also seen that the basic soil properties had the maximum influence on the soil hydraulic properties predictions. Increasing the number of model input parameters also means increasing the complexity of the model including the inherent uncertainties associated with the input data.

[4] Artificial neural networks (ANNs) have been used as PTF tools for parameter estimation in hydrology. *Pachepsky et al.* [1996], *Schaap and Bouten* [1996], *Schaap and Leij* [1998], *Schaap et al.* [1998], *Sharma et al.* [2006], and *Jana et al.* [2007] are a few examples. However, one of the major drawbacks of using a conventional ANN approach is the lack of uncertainty estimates. This, in turn, brings to question the quality of the ANN predictions. *Schaap and Leij* [1998] used bootstrapping to generate uncertainty estimates for ANN predictions of water retention parameters (and subsequently the water retention curve) and K_{sat}. However, these are a posteriori estimates of uncertainty. Owing to the nature of conventional ANNs, this knowledge of the uncertainty in predictions is not explicitly available from the ANNs. Conventional training of ANN weights is done by adjusting the values over multiple iterations till a single "optimal" set is obtained. Further, the ANN is not based on any physical processes underlying the hydrology. Hence the training of the weights in ANNs is totally dependent on the inputs. Different sets of training data would give rise to different sets of weights. This is especially true in hydrology because most hydrologic systems are inherently stochastic [*Kingston et al.*, 2005]. Thus the idea of an "optimal" set of weights would appear to be inappropriate. If a sufficiently large training data set is available, then it may be assumed that the "optimal" set of weights obtained is representative of the entire population. However, this is very rarely the case. Overtraining is also a matter of concern in conventional ANNs.

[5] Bayesian training of neural networks (NNs), on the other hand, obtains a range of weights. This, in turn, gives a distribution of predicted values, rather than a single value, thus explicitly accounting for the uncertainty in the predictions. Use of Markov chain Monte Carlo (MCMC) simulation techniques further reduces the possibility of the training becoming stuck in local minima. Bayesian training also prevents overtraining of the ANN. As such, Bayesian ANNs incorporate the advantages of conventional ANNs while eliminating some of the drawbacks. In this regard, Bayesian ANNs may be considered as the next evolution of ANNs.

[6] Bayesian ANNs have been used in related water resources modeling fields such as forecasting river salinity [*Kingston et al.*, 2005] and rainfall-runoff modeling [*Khan and Coulibaly*, 2006]. However, not much is done in this aspect toward soil hydraulic parameter estimation in the vadose zone. Moreover, most previous PTF studies derive and adopt soil hydraulic parameters at matching spatial scale of input and target data. The primary objective of this study is to develop and test a methodology to derive soil water content values (at saturation, θ_{0bar}; field capacity, $\theta_{0.3bar}$; wilting point, θ_{15bar}) and van Genuchten soil water retention function at point/local (1:1) scale using Bayesian neural networks based PTFs trained with coarse-scale (1:24,000) Soil Survey Geographic (SSURGO) soil textural data.

2. Bayesian Neural Networks

[7] Artificial neural networks, used in previous PTF developments, do not account for the physical processes governing the variations in soil hydraulic properties at different spatial scales. However, the basis for their utility comes from the relationship between the inputs and the targets developed during the ANN training. Let y be the target and let x be the input data, and then the relation between x and y can be described as

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

where $f(x|w)$ is the functional approximation (described by the ANN) of the relationship between the input and the target, w is the vector of weights and biases for the layers of ANN neurons, and E is the error term. Conventional (standard/deterministic) ANN methodology attempts to find a single set of weights w such that given the training inputs x, the network reproduces the training targets y with minimal error E.

[8] Bayesian ANNs, however, generate a probability distribution of the layer 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, ... x_n), Y is the target vector (y_1, y_2, ... y_n), $P(Y|X) = \int P(Y|w, X)P(w)dw$ is the marginal distribution of y, $P(w)$ is the prior distribution of weights, and $P(Y|w,X)$ is the likelihood function [*Gelman et al.*, 1995]. The marginal distribution, also known as the normalizing constant, is a constant of proportionality. The prior distribution of weights is our knowledge of the weight values before being introduced to the data. A noninformative prior (with arbitrary values) can be assumed here. 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 "$n+1$" denotes a new set of input and output variables that are not a part of X and Y. Evaluation of this integral is generally considered to be virtually impossible using conventional analytical or numerical integration methods. This integral can be solved by numerical integration using Markov chain Monte Carlo (MCMC) methods [*Neal*, 1992].

[9] MCMC methods aim to generate multiple samples from a continuous target density [*Bates and Campbell*, 2001]. The posterior weight distribution in multilayered ANNs is generally complex. Since it is difficult to sample directly from the complex posterior distribution, a simpler symmetrical distribution is used to generate the weight

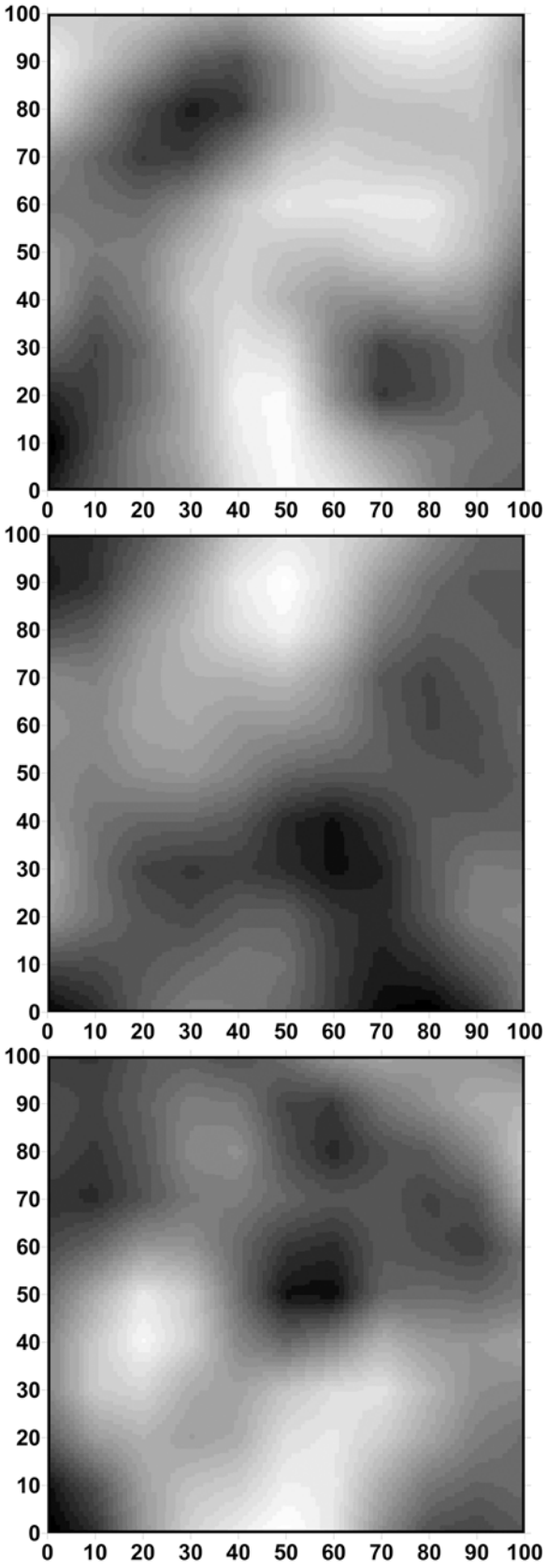

Figure 1. Samples of randomly generated coarse-scale parameter fields. One coarse pixel = 10 × 10 fine scale pixels.

vectors. This is called the "proposal" distribution and is considered to be locally Gaussian. In a "random walk" implementation of the Markov chain, the proposal distribution depends only upon the previous weights. Starting with the arbitrary values of w from the prior distribution, a series of values w^* are proposed by the Markov chain. These proposed values are accepted with a probability

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

Here w_{prev} is the previous weight value. If w^* is accepted, the value of w_{prev} is replaced by w^* and the procedure is iterated over again. An acceptance rate between 30% and 70% is considered to be optimal [*Bates and Campbell*, 2001]. Using a sufficiently large number of iterations ensures that the Markov chain converges to a stationary distribution. At that point, the weight vectors are considered to have been generated from the posterior distribution itself. A more detailed description and discussion of the Metropolis algorithm for the MCMC method used in this study is given by *Gelman et al.* [1995] and *Kingston et al.* [2005]. In our study, we generated 15,000 iterations for the Markov chain and discarded the first 5000 iterations as burn-in. This is done to allow the network suitable time to "understand" the relationship between the inputs and the outputs, and to attain stability.

[10] ANNs with one input layer, one hidden layer of four neurons, and one output layer were used in this study. The tangent hyperbolic transfer function was used between all neuron layers.

3. Numerical Study

[11] Spatially correlated Gaussian random fields for all input and target parameters were generated using the "Hydro_Gen" program [*Bellin and Rubin*, 1996]. One hundred fine-scale values were generated in a 10 × 10 grid for input (sand, silt, clay, and bulk density) and target ($\theta_{0\text{bar}}$, $\theta_{0.3\text{bar}}$, and $\theta_{15\text{bar}}$) parameters. A grid spacing of 10 units was used along with a correlation length of 20 units in both the x and y directions. These parameters are then individually averaged to form one coarse scale pixel of dimension 100 × 100 square units for each parameter. One thousand such coarse-scale pixels were generated per parameter. Arithmetic, harmonic, and geometric averaging methods were used in the coarsening procedure. However, it was observed that there was no significant difference between the three mean values. Hence the arithmetically averaged values were used as the coarse-scale training data in our study. Sample coarse-scale pixels are shown in Figure 1. One thousand fine-scale values were randomly selected from within the coarse-scale pixels to serve as the fine-scale testing data. Statistics of the coarse and fine scale data are given in Table 1.

[12] The Bayesian ANN was trained with the generated coarse-scale data and then shown the fine-scale inputs. The ANN outputs ($\theta_{0\text{bar}}$, $\theta_{0.3\text{bar}}$, and $\theta_{15\text{bar}}$) are then compared with the corresponding generated values. Figure 2 shows plots of the generated and ANN predicted θ values, and their comparative statistics are given in Table 2. It can be

Table 1. Statistics of Synthetically Generated Parameter Values (Numerical Study)

		Sand, %	Silt, %	Clay, %	BD, g/cm^3	θ_{0bar}, %	$\theta_{0.3bar}$, %	θ_{15bar}, %
Coarse scale	Mean	69.586	9.839	20.575	1.584	34.816	11.870	7.870
	SD	1.237	0.478	1.715	0.048	0.554	0.392	0.392
Fine scale	Mean	69.480	9.800	20.720	1.580	34.768	11.836	7.836
	SD	4.757	1.842	6.599	0.184	2.127	1.504	1.504

seen that the ANN predicted values, while being highly correlated with the generated values, are higher than the generated values for all three parameters. There is a systematic shift in the values.

[13] *Schaap and Leij* [1998] observed that significant differences exist between the mean values of similar parameters obtained from different databases (RAWLS, AHUJA, and UNSODA). This could be due to differences in measuring techniques or instrumental and/or human errors. Other theories attribute bias to differences in soil types or presence of heterogeneities. In our numerical study, however, these causes can be discounted. Since all parameters are synthetically generated, the issues of differing measurement techniques and instrument error are eliminated. Generating parametric values from Gaussian distributions also ensures that heterogeneities are not found in any of the coarse-scale pixels, unless intentionally introduced. As seen in Table 2, there is no significant difference between the mean and variability statistics of the training and testing data. Hence the only difference between the training and testing data sets is the scale. Since the training of the neural network is done by coarse-scale data, the model developed by the ANN is a coarse-scale model. When fine-scale inputs are fed to this model, the predictions obtained for the soil water contents are still at the coarser scale. This means that a bias exists between the ANN predicted values and the measured values at the point scale. This bias is attributed to the support-scale disparity in the model and the data. Different governing hydrologic processes dictate the soil water contents at different spatial scales. However, as previously mentioned, the ANN is not based on the physical processes underlying the hydrology. Hence a suitable bias correction technique needs to be applied to the predicted water content values.

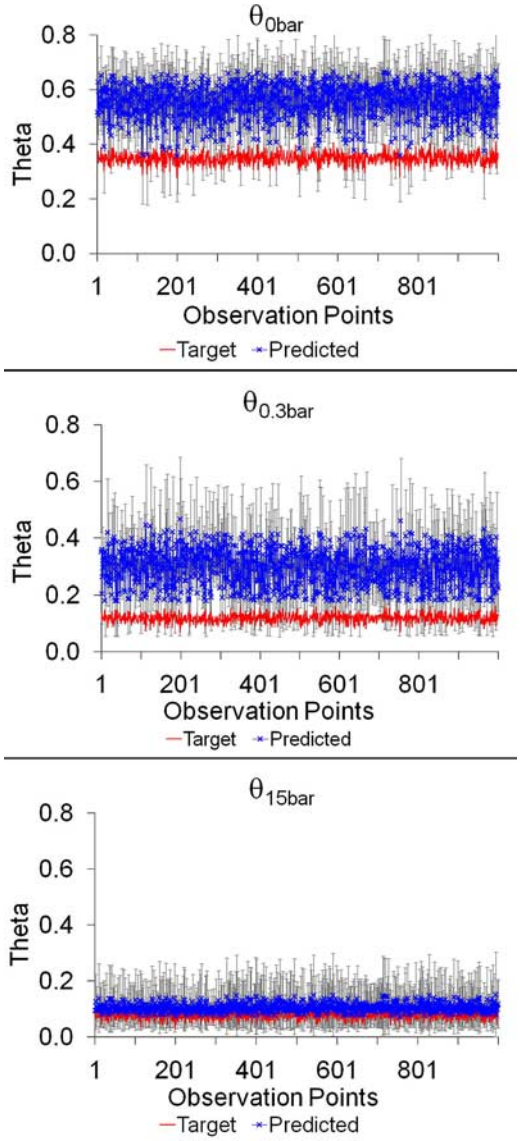

Figure 2. Graphs of target and artificial neural network (ANN) predicted water content values. Theta indicates soil water content.

4. Field Study

[14] We tested the Bayesian training methodology for deriving soil water retention at the local/fine scale from ANNs trained at a coarser scale in two different regions in the United States. The first site is in the Rio Grande basin of New Mexico, and the second site is the Southern Great Plain Experiment 1997 (SGP97) hydrology experiment region in Oklahoma.

4.1. Rio Grande Basin

[15] A soil properties database for the Rio Grande basin in New Mexico [*Jana et al.*, 2005] was used in this study. Point-scale (1:1) soil physical and hydraulic properties measured at the Las Cruces Trench site [*Wierenga et al.*,

Table 2. Comparative Statistics of Target and Predicted Water Content Values (Numerical Study)

	θ_{0bar}		$\theta_{0.3bar}$		θ_{15bar}	
	Target	Predicted	Target	Predicted	Target	Predicted
n	1000	1000	1000	1000	1000	1000
Mean	0.348	0.556	0.118	0.300	0.078	0.106
SD	0.021	0.063	0.015	0.066	0.015	0.015
CV	0.061	0.113	0.127	0.220	0.192	0.140
R		0.985		0.984		0.980
R^2		0.970		0.968		0.961
RMSE		0.213		0.199		0.028

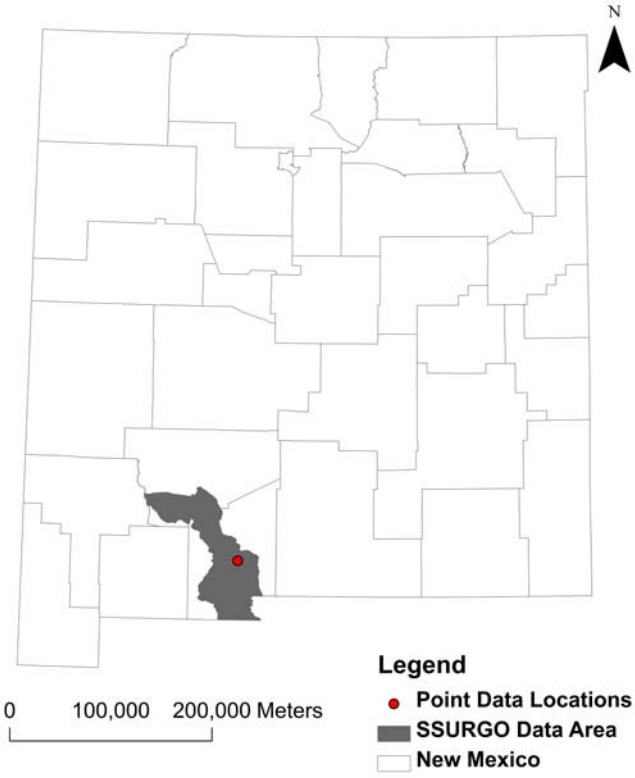

Figure 3. Rio Grande Basin study area, New Mexico.

1989] situated within the Rio-Grande-Mimbres subwatershed region (Figure 3) were used as the model inputs at local/point scale. The Las Cruces Trench (Figure 3) is located on the New Mexico State University ranch, roughly 40 miles northeast of the city of Las Cruces. The trench is located in undisturbed soil on a basin slope of Mount Summerford, near the northern end of the Dona Ana Mountains. 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 (1:1) 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. Since the data for the coarser scale are from the topsoil layer (0–6 cm), we use only the 59 sites from the top 6-cm layer of the Las Cruces Trench site database in this study.

4.2. Southern Great Plains

[16] A soil property database of the Southern Great Plains 1997 (SGP97) hydrology experiment area [*Mohanty et al.*, 2002] was used in the second case study. Figure 4 shows the experimental region of approximately 40 km $\times$ 250 km (10,000 km^2) in the central part of the U.S. Great Plains in the subhumid environment of Oklahoma. Data measured in the Little Washita Watershed area within this region was used for this study. The region has a moderately rolling topography. Rangeland and pasture dominate the land use with patches of winter wheat and other crops [*Allen and Naney*, 1991].

[17] At both the sites, the coarse-resolution (1:24,000) soil properties data were derived from the U.S. Department of Agriculture/Natural Resources Conservation Service Soil Survey Geographic (SSURGO) database http://soildatamart.nrcs.usda.gov). SSURGO is a public domain database containing georeferenced spatial and attribute data for soils. Since these surveys cover large spatial extents, soil property data are based on the 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

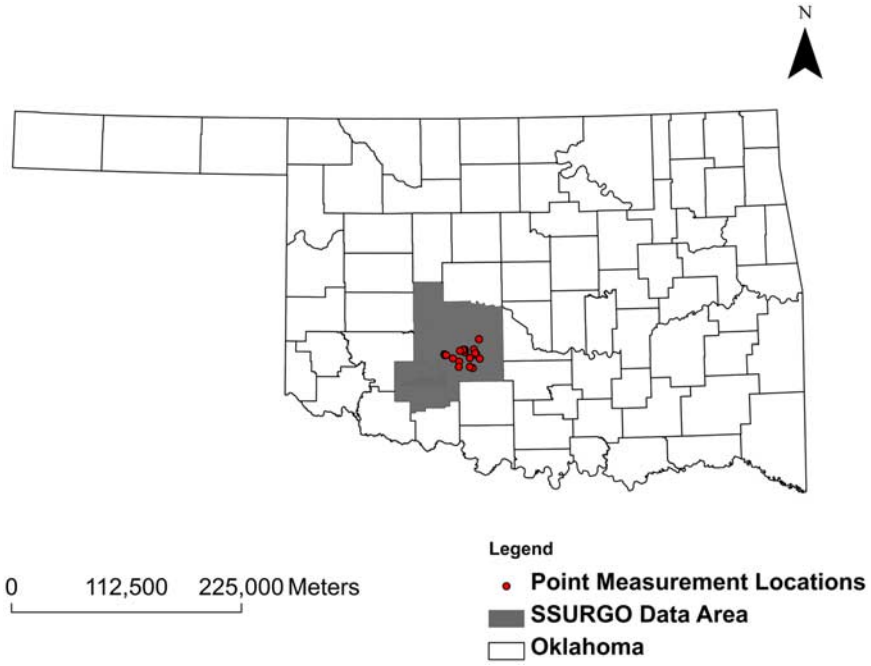

Figure 4. Southern Great Plains study area, Oklahoma.

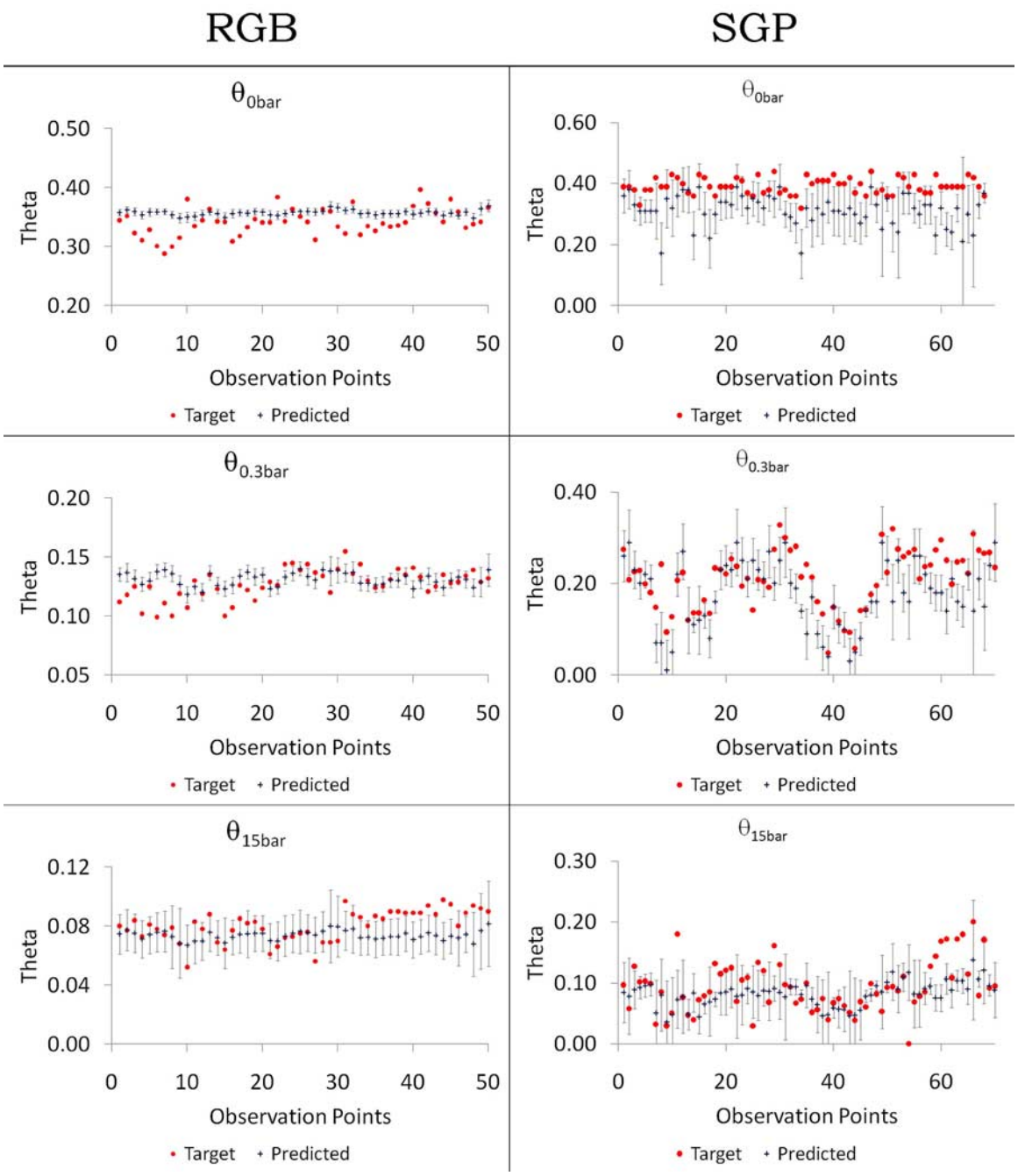

Figure 5. Target and predicted soil water content values. Theta indicates soil water content; RGB, Rio Grande Basin; SGP, Southern Great Plains.

by field transects [*Natural Resources Conservation Service*, 2007]. 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 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 the observed readings are included in the database for a particular soil type/map unit. These soil maps are made at scales ranging from 1:12,000 to 1:31,680 (http://www.nrcs.usda.gov/technical/soils/soilfact.html). In this study, we have used the representative value for the soil physical and hydraulic parameters from 1:24,000 resolution soil maps.

[18] As in the numerical study, the soil physical properties used are the % sand, % silt, % clay, and the oven-dry bulk density. The hydraulic parameters are the water content at satiation (θ_{0bar}), the water content at a pressure of 1/3 bar ($\theta_{0.3bar}$), and the water content at 15 bar (θ_{15bar}).

[19] From the SSURGO data for each region, we trained the ANNs for estimating the saturation, field capacity, and wilting point soil water contents (θ_{0bar}, $\theta_{0.3bar}$, and θ_{15bar}, respectively). One thousand random sets of data values were selected for the ANN training from this data pool by a

Table 3. Descriptive Statistics for θ Values (Field Studies)[a]

	θ_{0bar}				$\theta_{0.3bar}$				θ_{15bar}			
	Target	Pred	LBC	NLBC	Target	Pred	LBC	NLBC	Target	Pred	LBC	NLBC
RGB												
N	50	50	50	50	50	50	50	50	50	50	50	50
Mean	0.342	0.357	0.340	0.342	0.127	0.131	0.126	0.127	0.080	0.074	0.078	0.080
SD	0.023	0.004	0.002	0.023	0.013	0.006	0.003	0.013	0.011	0.003	0.003	0.010
SGP												
N	68	68	68	68	70	70	70	70	70	70	70	70
Mean	0.393	0.312	0.391	0.393	0.209	0.175	0.196	0.208	0.093	0.082	0.094	0.093
SD	0.027	0.051	0.010	0.027	0.065	0.072	0.038	0.063	0.042	0.020	0.023	0.041

[a]Abbreviations are RGB, Rio Grande Basin; SGP, Southern Great Plains; N, number of data values; Pred, ANN-predicted; LBC, linear bias corrected; NLBC, nonlinear bias corrected.

bootstrapping process. Each data set consists of training inputs (% sand, % silt, % clay, and oven-dry bulk density) and the corresponding target outputs (θ_{0bar}, $\theta_{0.3bar}$, and θ_{15bar}). It was observed by conducting several replicated model runs that further increase in the size of the training data set (>1000 and within the available data pool) did not provide any further improvement in the training. Moreover, a low ratio of selected to available data sets ensured randomness of the bootstrapped selections. Finally, using the trained ANNs with the SSURGO-based coarse-resolution data sets, predictions of soil water contents (θ_{0bar}, $\theta_{0.3bar}$, θ_{15bar}) were made at the point resolution for the corresponding point-scale data sets of % sand, % silt, and % clay, and oven-dry bulk density.

5. Nonlinear Bias Correction

[20] Linear bias correction, as applied by *Hansen and Ines* [2005] and *Jana et al.* [2007], provides a proportional shifting effect to the ANN-predicted curve and brings the mean of the ANN-predicted values closer to that of the target values. However, this is a first-moment correction only. Only the mean of the predictions is brought closer to that of the target by this method, while no correction is applied to the second moment (spread) of the values. Moreover, parametric scaling is a nonlinear process. Applying a linear bias correction to approximate this process can be successful only to a certain degree. Hence a nonlinear technique needs to be adopted for the bias correction.

[21] Matching of cumulative distribution functions (CDFs) is a technique that has been used to correct for nonlinear bias [*Calheiros and Zawadzki*, 1987; *Atlas et al.*, 1990; *Anagnostou et al.*, 1999; *Wood et al.*, 2002; *Reichle and Koster*, 2004; *Ines and Hansen*, 2006; *Baigorria et al.*, 2007; *Hashino et al.*, 2007]. The technique is based on the idea to obtain the predicted parameter values corresponding to the probability of occurrence of values on the CDF of the target parameter. Statistical tests are conducted to test for the type of distribution (e.g., normal, lognormal, gamma) of the parameters. CDFs are then obtained for the target and predicted values for each parameter based on the type of distribution they follow. For a particular 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 toward the corresponding target soil water content value.

6. Results and Discussions

[22] Figure 5 shows the targets and the outputs with the uncertainty band predicted by the Bayesian ANNs for the three soil water retention (θ_{0bar}, $\theta_{0.3bar}$, θ_{15bar}) parameters in the two different regions. The error bars are obtained from the Markov chain Monte Carlo simulations and represent the uncertainty in the neural network predictions. Bayesian neural networks, as mentioned earlier, use a range of weight sets instead of a single set. The uncertainty band (error bars) show the limits to which the predictions could have varied based on the weight set used. The final predicted soil water content value is the average of all such simulations. In the Rio Grande Basin (RGB), we have 50 point-scale inputs/outputs for each of the three parameters. In the Southern Great Plains (SGP), we have 70 point-scale measured values for θ_{15bar} and $\theta_{0.3bar}$. However, we used only 68 point-scale values for θ_{0bar} since two values were discarded as outliers. Table 3 gives the target and predicted means and standard deviations for the three parameters at the two sites.

[23] It is observed that the Bayesian ANN approximates the water contents slightly better for the SGP case, as compared with the RGB. This fact is attributed to the difference in the site conditions. The RGB point-scale data were collected from a single 26.4-m-long, 4.5-m-wide trench near the city of Las Cruces, New Mexico. The soil type shows little variation over this area, as can be inferred from Table 4. The SGP data, on the other hand, were

Table 4. Soil Physical Properties at Coarse and Fine Scale for the RGB Region (Field Study)

	Sand		Silt		Clay		Bulk Density	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (g/cm^3)	SD (g/cm^3)
SSURGO Data	49.43	21.49	29.20	19.39	21.37	10.87	1.60	0.12
Point Data	81.46	1.67	9.76	1.17	8.78	1.20	1.66	0.05

Table 5. Soil Physical Properties at Coarse and Fine Scale for SGP Region (Field Study)

	Sand		Silt		Clay		Bulk Density	
	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (%)	SD (%)	Mean (g/cm^3)	SD (g/cm^3)
SSURGO Data	43.31	21.82	37.74	17.31	19.17	8.80	1.78	0.42
Point Data	51.91	21.11	33.61	16.41	14.48	6.04	1.40	0.10

collected at points covering an area of approximately 10,000 km^2 in the state of Oklahoma. Large variations in soil type can occur across this big area (Table 5), as compared with the Las Cruses trench in RGB. While the SSURGO data have comparable statistics for the four input soil physical properties, the point-scale data are vastly different in the two cases (Tables 4 and 5). The standard deviation, the measure of variability in the texture, is very

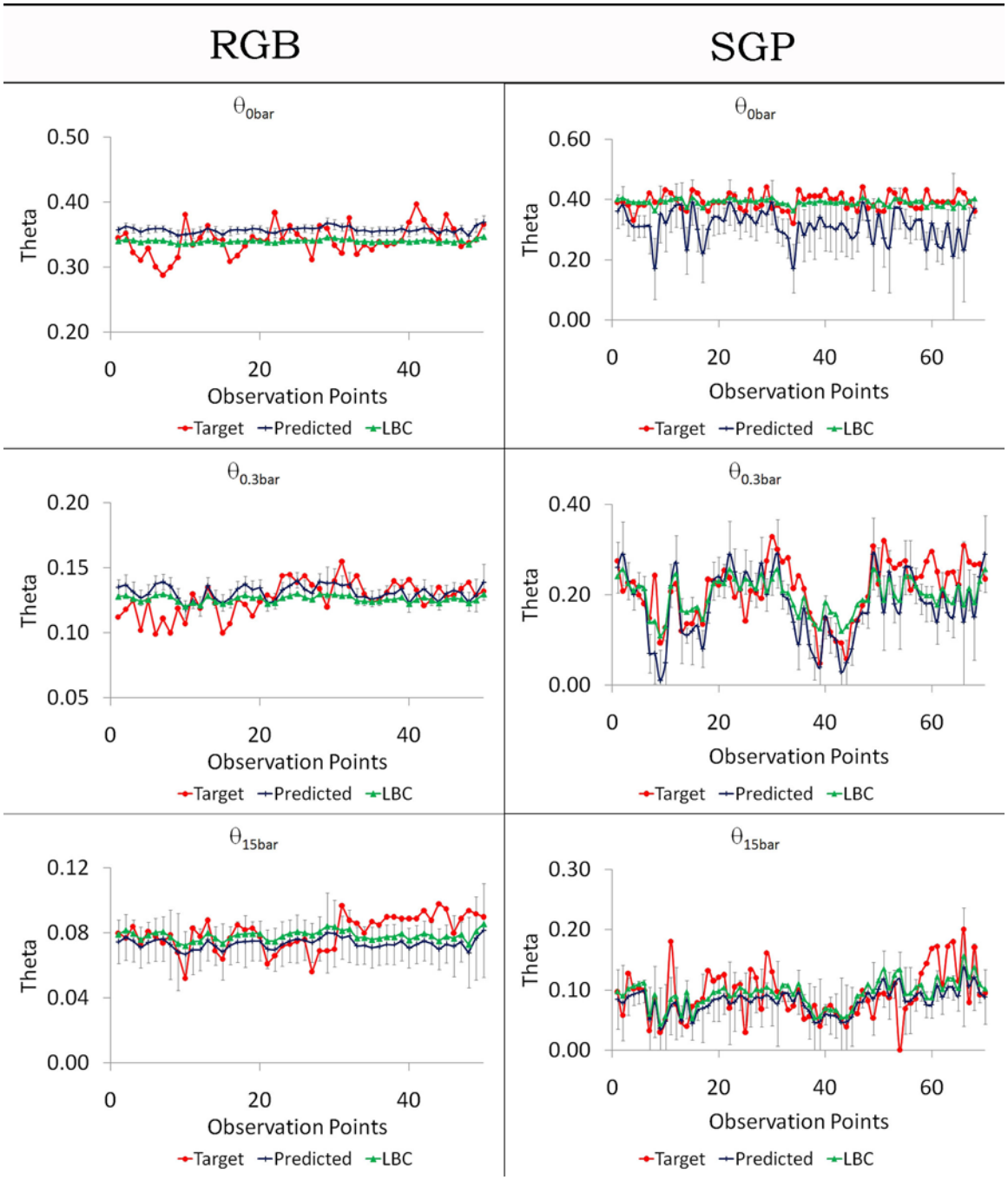

Figure 6. Graph of target, predicted, and linear bias corrected (LBC) soil water content values. Theta indicates soil water content; RGB, Rio Grande Basin; SGP, Southern Great Plains.

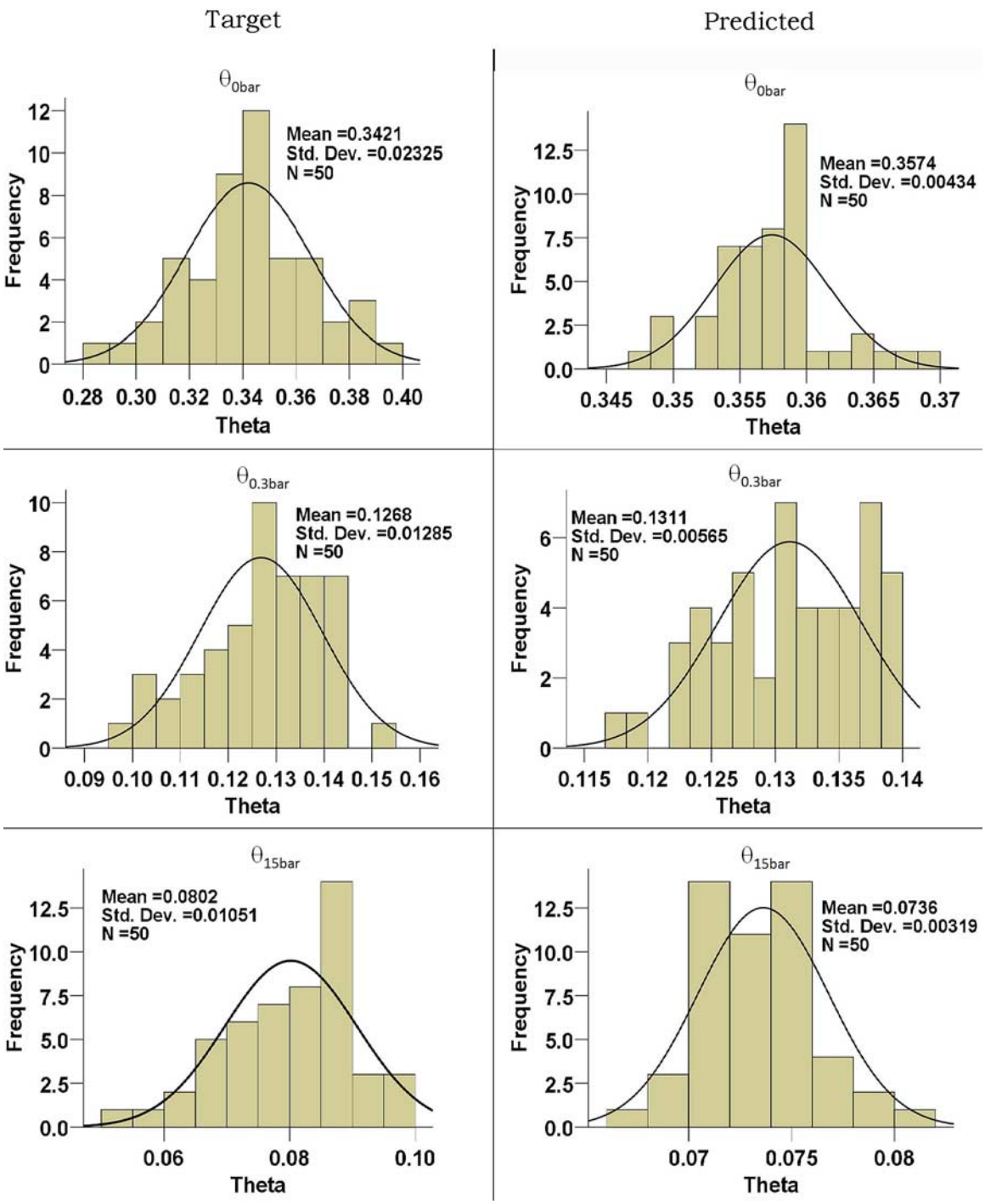

Figure 7. Histograms for target and ANN-predicted θ values for the RGB area.

low for the RGB data. The corresponding statistics for the SGP region show that the Bayesian ANN is trained using a larger variety of soil types from the SGP data and hence is able to capture a larger variability in soil water content values. The ANN trained with the RGB data has only limited variation and is thus restricted in its scope. This argument is supported by the tightness of the uncertainty band for all parameters for the RGB case. The band for the predictions does not encompass the target values in most cases. On the contrary, the uncertainty band for the predicted θ for the SGP region mostly includes the target values. This observation supports the notion for collecting more point-scale data from larger spatial extents. Having data from a larger extent ensures that the Bayesian ANN can train itself with a wider range of water content values and hence result in better predictions.

[24] A discussion on the usage of bootstrapping and MCMC techniques is necessary here. The bootstrapping process is used independent of the ANN in our study. It is used as a random selection method to extract the 1000 data

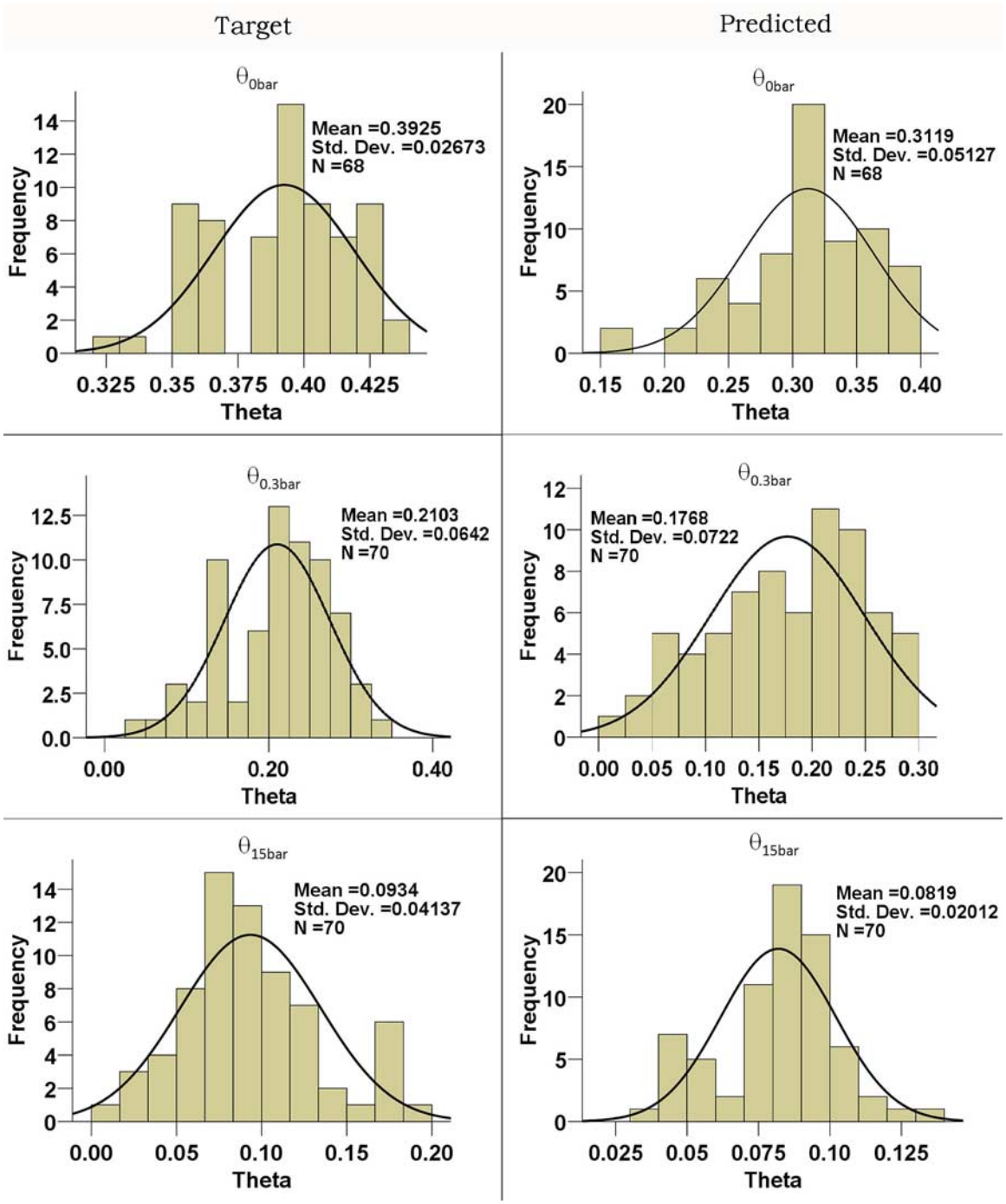

Figure 8. Histograms for target and ANN-predicted θ values for SGP area.

values from the data pool. While any random selection algorithm would have sufficed for the current data, the bootstrapping process was chosen bearing in mind the utility of this method when such a large data pool is not available. *Schaap and Leij* [1998] use the bootstrapping method to generate multiple replicates for the ANN calibration and validation sets. These replicates were used to generate predictions of water retention parameters and K_s. Confidence intervals were then computed using the generated predictions' standard deviations. Uncertainty of the ANN is assessed in our study using the MCMC method, which is a part of the Bayesian framework for ANNs. MCMC evaluates the ANN's predictions with respect to the expected targets from the training and provides feedback to the ANN to enable it to modify the weights structure in order to obtain better predictions. Only those predictions that exceed a certain given probability of being similar to the required distribution are accepted. The range of these accepted values provides the uncertainty estimates.

[25] There is a significant difference in the means and standard deviations of the target and the predicted soil water content values (Table 3). As shown in the numerical study, this deviation of the prediction from the target can be attributed to the scale/extent effect. We have trained the

Table 6. One-Sample Kolmogorov-Smirnov Normality Test Results[a]

	θ_{0bar}		$\theta_{0.3bar}$		θ_{15bar}	
	Target	Predicted	Target	Predicted	Target	Predicted
RGB						
N	50	50	50	50	50	50
Mean	0.342	0.357	0.127	0.131	0.080	0.074
SD	0.023	0.004	0.013	0.006	0.011	0.004
Kolmogorov-Smirnov Z	0.776	0.950	0.659	0.662	0.664	0.620
Asymptotic sigma (two-tailed)	0.584	0.327	0.779	0.773	0.802	0.837
SGP						
N	68	68	70	70	70	70
Mean	0.393	0.312	0.210	0.177	0.093	0.082
SD	0.027	0.051	0.064	0.072	0.041	0.020
Kolmogorov-Smirnov Z	1.156	0.995	0.803	0.825	0.668	0.941
Asymptotic sigma (two-tailed)	0.138	0.276	0.540	0.504	0.764	0.339

[a]Abbreviations are RGB, Rio Grande Basin; SGP, Southern Great Plains; N, number of data values.

neural network at a coarse scale using SSURGO soil properties data. This means that the model for the relationship between the inputs and outputs in the network's memory is a coarse-resolution model. Despite our simulation inputs being at the point scale, the predictions are technically at the coarse scale. Hence we observe the difference in the target and predicted values. Application of a bias correction technique is necessary to minimize the difference between the target and predicted values.

[26] A linear bias correction, as used by *Hansen and Ines* [2005] and *Jana et al.* [2007], was applied to the predicted water content values (θ_{0bar}, $\theta_{0.3bar}$, θ_{15bar}). As mentioned earlier, this technique only corrects for the mean of the distribution, and not for the spread. Figure 6 shows the mean-corrected values of the water content after linear bias correction for the RGB and the SGP regions. It is observed that because only the systematic error is accounted for in this technique, there is a smoothing effect on the soil water content (θ_{0bar}, $\theta_{0.3bar}$, θ_{15bar}) distribution. This smoothing further decreases the standard deviation of the predicted soil water content values. With such a bias correction technique, we obtain a good estimate for the mean of the entire data set (i.e., we have a good estimate for the effective soil water content values at the field scale). At the point scale, however, the estimation is not as effective.

[27] The CDF mapping method, which provides corrections to the first four moments of the distribution, is used for nonlinear bias correction. The three soil water contents (θ_{0bar}, $\theta_{0.3bar}$, θ_{15bar}) are hypothesized to be normally distributed. Histograms are plotted for each parameter, target and predicted, for both the RGB and SGP regions in Figures 7 and 8. In addition to the visual examination, we conducted the Kolmogorov-Smirnov test to check for normality. Table 6 shows the normal distribution at a confidence level of 0.05 for all the three soil water contents.

[28] Normal CDFs are plotted for the target and predicted θ values for both the regions (Figure 9). The CDF for a normal distribution is given by

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

where θ_i is the value at which the CDF is calculated, and μ is the mean and σ is the standard deviation of the soil water content values. 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.

[29] The Bayesian ANN-predicted θ values are randomly split into two halves: one for the model calibration and the other for the validation of the bias correction scheme. The cumulative probabilities for each point value are computed using the mean and standard deviation of the calibration data set. In order to effectively scale the model predicted values to the measured data set, we now find θ_i^{pred} such that

$$CDF\left(\theta_i^{pred}\right) = CDF\left(\theta_i^{target}\right). \quad (6)$$

This is achieved by computing the inverse of the cumulative probability values for the calibration data set, 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 data set to approximate that of the target values. The calibrated (bias corrected) soil water content CDF values are plotted (Figure 9). It can be seen that the calibrated CDF follows the target CDF closely.

[30] To test the calibration scheme, the remaining half of the neural network predicted soil water content values (the validation data set) are used. The calibration scheme is found to be correct as the validation data plot on top of the target (Figure 9). Results from two-sample Kolmogorov-Smirnov tests comparing the calibrated/validated data with the target data are given in Table 7. The test results confirm that the null hypothesis cannot be rejected at the 5% significance level, and hence the calibrated/validated data are not significantly different from the target data. This means that our bias correction scheme for the predicted θ values approximates the target values well.

[31] Probability density functions (PDFs) are plotted for the target, predicted, calibrated, and validated θ values for both the regions (Figure 10). The difference in the location (mean) and scale (spread) of the target and predicted distributions is apparent here. In concurrence with the CDFs, the PDFs of the calibrated and validated data sets plot on top of the target PDF.

[32] Figure 11 shows the target, Bayesian ANN-predicted, and nonlinear bias-corrected θ values. It can be readily seen that the variability of the target values is largely approximated by the corrected θ values. The point-to-point matching, however, still appears to be lacking. It must be noted here that the bias-corrected and target θ values are being sampled from normal distributions that are close to one another. They are, however, being sampled at different probabilities, and hence there is apparent mismatch in the point values. If we select values from the normal distribution of the bias-corrected values for the exact probabilities as those of the target values, we would expect a very good match at each point.

[33] There are always uncertainties in the observations of any point-scale data because of measurement errors and operator errors. There are also other factors influencing the soil water content values that have not been considered here, such as the presence of macropores or roots debris.

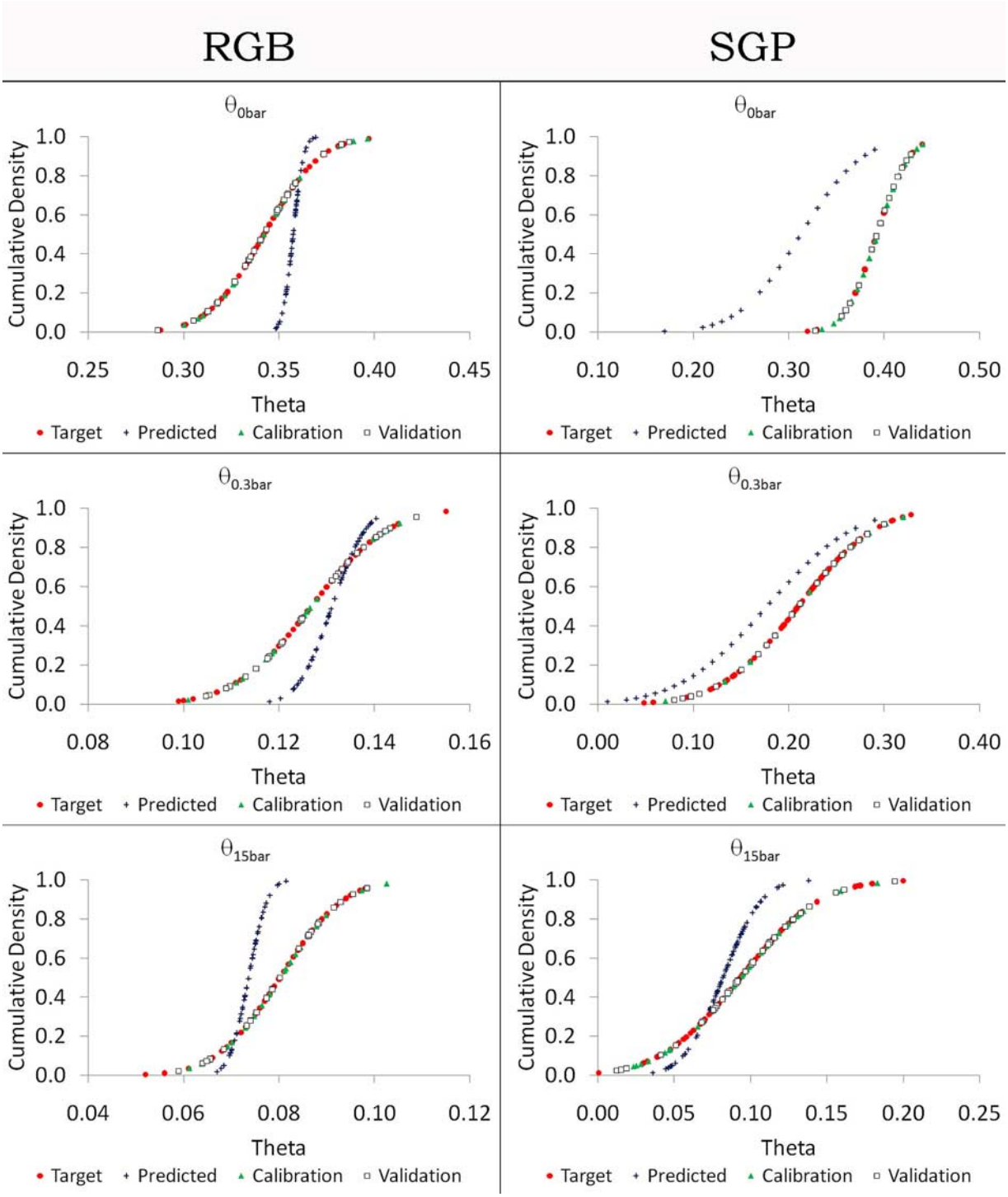

Figure 9. Normal cumulative density function plots of target, predicted, and nonlinear bias corrected calibration and validation soil water content values. Theta indicates soil water content; RGB, Rio Grande Basin; SGP, Southern Great Plains.

These factors make the approximation of the particular point values a near-impossible task, given the current inputs. Note, however, that this approach provides us θ values for any probability whereas the linear bias correction technique could not give us this capability. Matching of the distributions ensures that the above mentioned uncertainties are incorporated into the estimation scheme. In other words, the approach is better for field-scale mean parameter estimation as compared with matching particular point values.

[34] Other limitations of the proposed method warrant some discussion. PTFs based on ANNs are inherently site-specific as they need to be trained to recognize the patterns particular to that site. Also, a few measurements are necessary at the fine scale for the bias correction procedure. This correction gives us the estimate of the amount of

Table 7. Two-Sample Kolmogorov-Smirnov Test Results for Similarity of Distributions[a]

	Data	Number of Points	Minimum	Maximum	Mean	SD	p Value
RGB							
θ_{0bar}	Target	50	0.288	0.397	0.342	0.023	
	Calibrated	25	0.300	0.396	0.342	0.023	0.954[b]
	Validated	25	0.287	0.387	0.342	0.023	0.759[b]
$\theta_{0.3bar}$	Target	50	0.099	0.155	0.127	0.013	
	Calibrated	25	0.101	0.145	0.127	0.013	0.864[b]
	Validated	25	0.105	0.149	0.127	0.013	0.870[b]
θ_{15bar}	Target	50	0.052	0.098	0.080	0.011	
	Calibrated	25	0.061	0.103	0.080	0.011	0.740[b]
	Validated	25	0.059	0.098	0.080	0.011	0.992[b]
SGP							
θ_{0bar}	Target	68	0.320	0.440	0.393	0.027	
	Calibrated	34	0.328	0.440	0.393	0.027	0.071[b]
	Validated	34	0.328	0.428	0.393	0.027	0.157[b]
$\theta_{0.3bar}$	Target	70	0.049	0.328	0.210	0.064	
	Calibrated	35	0.071	0.320	0.210	0.064	0.966[b]
	Validated	35	0.080	0.300	0.210	0.064	0.810[b]
θ_{15bar}	Target	70	0.001	0.200	0.093	0.041	
	Calibrated	35	0.000	0.183	0.093	0.041	0.384[b]
	Validated	35	0.013	0.194	0.093	0.041	0.714[b]

[a]Abbreviations are RGB, Rio Grande Basin; SGP, Southern Great Plains; p-value, KS test statistic.
[b]Significant at the 0.05 level.

correction to be applied, which can then be extrapolated to the entire extent of interest.

[35] Field experts may question the suitability of using values from the NRCS-SSURGO database as the coarse-scale data. These reservations are based on the fact that the SSURGO data are also measured at the point scale and hence have a measurement "support" similar to that of the RGB or SGP point-scale data. However, it must be noted that while the initial measurements are point measurements, the values in the database are averaged values across the soil type or map boundaries. Averaging several representative point scale values is, in itself, a simple form of upscaling. Hence the SSURGO database gives us upscaled (coarse resolution, 5 m) values for the soil water function parameters. Second, the term scale consists of three components: support, extent, and spacing [*Blöschl and Sivapalan*, 1995]. For the initial measurements, two components of the scale triplet, extent and spacing, are much larger for the SSURGO data, as compared with either the RGB or the SGP data. According to *Western et al.* [2002], scaling often involves changing more than one component of the scale triplet at a time. On the basis of the above arguments, it is suitable to consider the SSURGO data as upscaled or coarse-resolution values as compared with either the RGB or the SGP data.

[36] Further, it may be argued that the bias between the target and ANN predicted water content values are due to the point-scale data being measured in a patch of heterogeneous material within the soil map unit. Such heterogeneous values would be smoothed out in the averaging process to obtain the coarse-scale SSURGO values. In the numerical study, we can see that the training and testing data sets have similar statistics, showing that the testing data set is not from a heterogeneous region within the coarse-scale pixel. In our opinion, this question has been answered by applying the methodology to the two diverse test beds. While patchy heterogeneity may be an argument in case of the RGB data, the same cannot be argued for in case of the SGP data. The SGP data have been collected over a much larger extent, and the possibility of hitting heterogeneous patches most of the time is low. At the RGB site, the point-scale measurements were in reasonable agreement with the SSURGO values for the corresponding soil map units. For example, data from SSURGO show that the soil map units covering the Las Cruces trench site have sand contents of 84.3% and 68.1%, providing an average of 76.2%. The point-scale measured data at the location have an average of 81.46% sand. This means that the point measurements were not from a heterogeneous patch within the soil map unit, and a comparable set of coarse-scale inputs were available in the data pool for training.

7. Conclusions

[37] Using coarse-scale soil properties data from the SSURGO database 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 different study regions (SGP, RGB) that differed greatly in soil, topography, vegetation, and climate and in the spatial extent from which the point data was collected. It has been shown that the Bayesian ANN predictions are superior when the data are from a larger region. This is due to the variety encompassed in the training process. The Bayesian ANN-predicted θ values were further corrected for bias. While a linear bias correction scheme can only correct for the difference in the means between the target and predicted value, a nonlinear correction scheme also corrects for the spread of the predicted distribution. The CDF matching technique was used to obtain the nonlinear bias correction. Applying a Bayesian ANN to this task also gives us an estimate of the uncertainty involved in the prediction scheme, which is an improvement over traditional ANN methods. Overall, the Bayesian ANN, coupled with a nonlinear bias correction scheme, appears to work well for estimation of soil hydraulic properties at a fine scale from data at coarser scales.

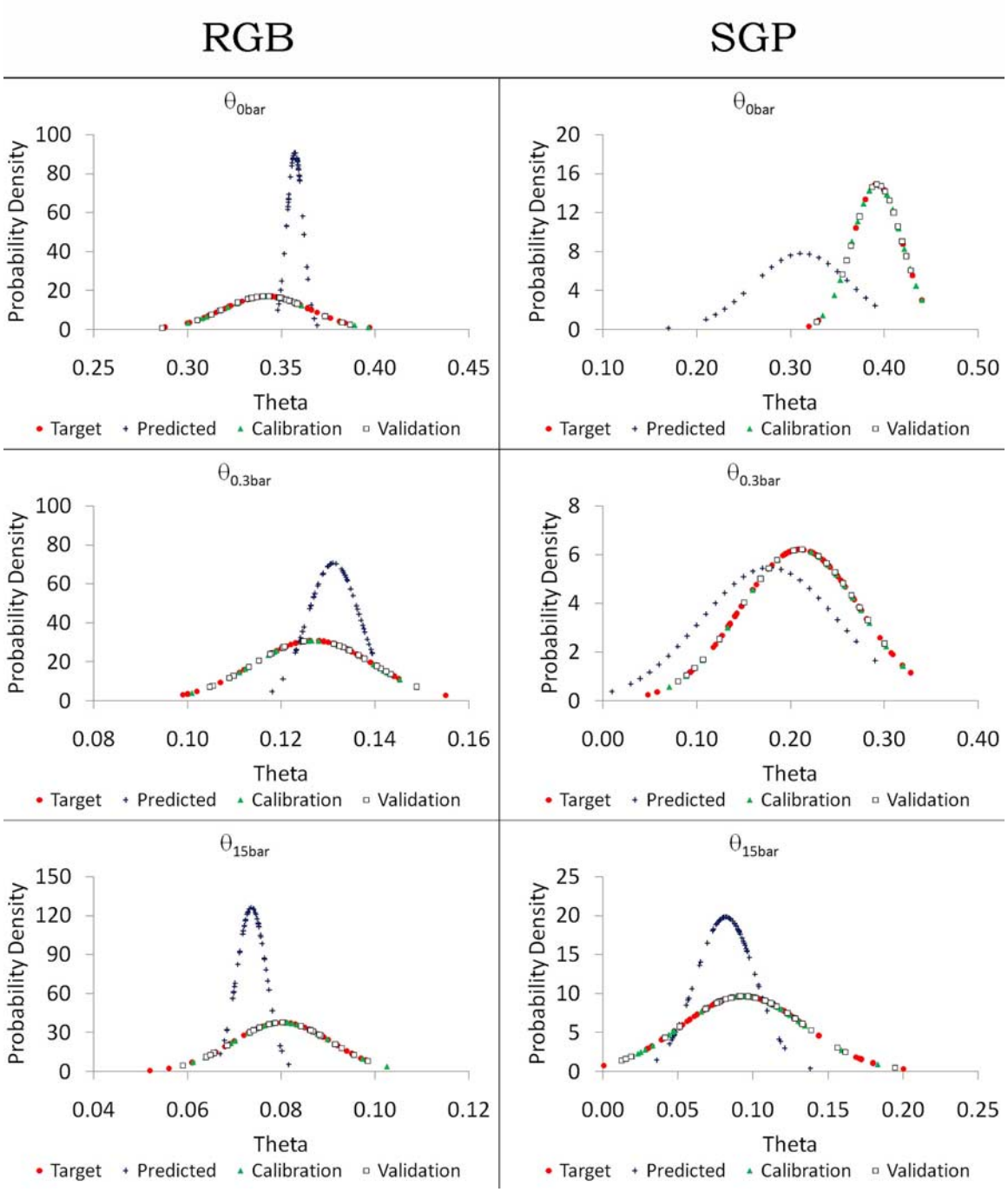

Figure 10. Normal probability density function plots of target, predicted, and nonlinear bias corrected calibration and validation soil water content values. Theta indicates soil water content; RGB, Rio Grande Basin; SGP, Southern Great Plains.

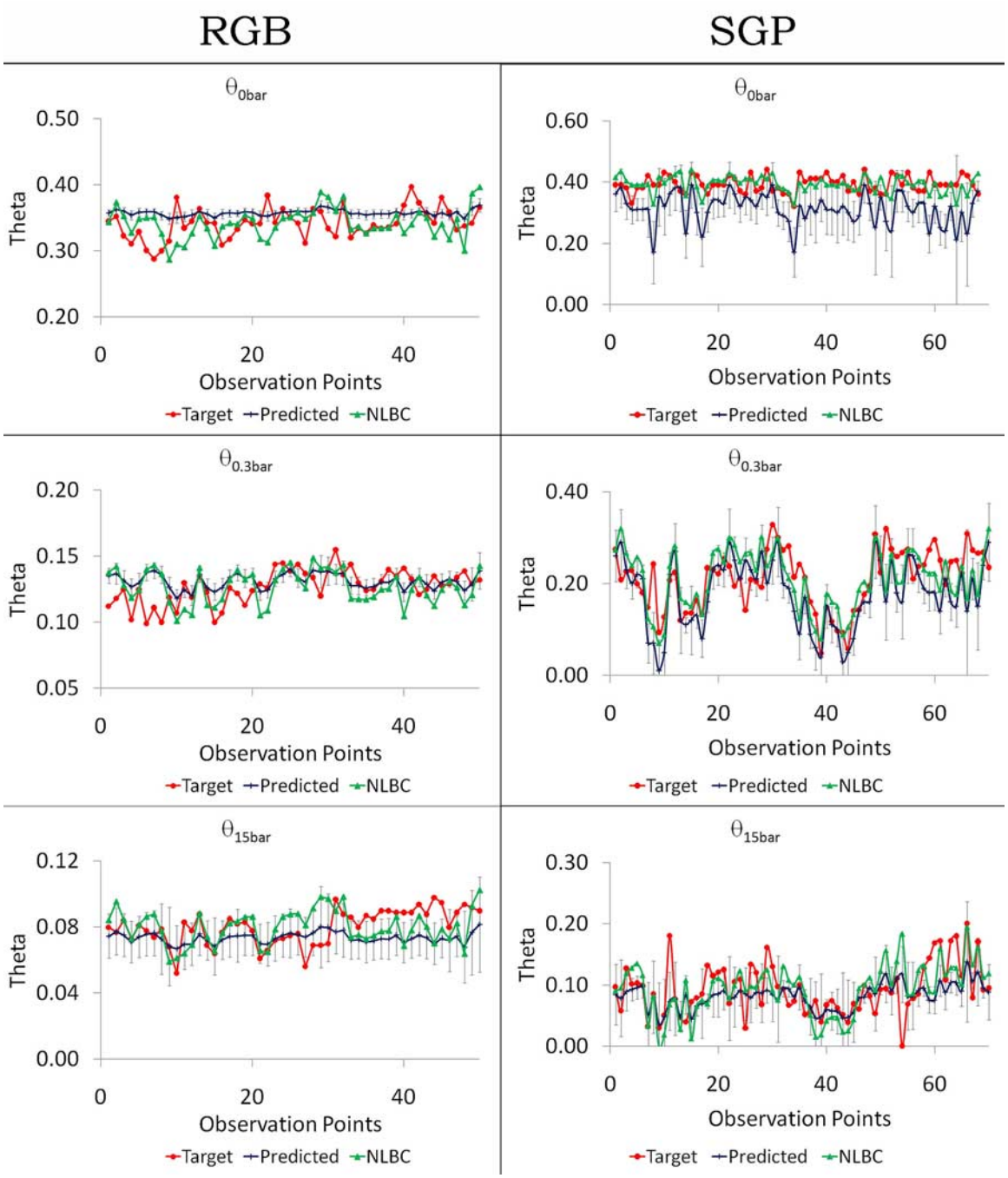

Figure 11. Graph of target, predicted, and nonlinear bias corrected soil water content values. NLBC, nonlinear bias corrected values; theta indicates soil water content; RGB, Rio Grande Basin; SGP, Southern Great Plains.

[38] **Acknowledgments.** We acknowledge the partial support of Los Alamos National Laboratory, NASA Earth System Science Fellowship (NNX06AF95H), NSF CMG (TEES 37050), and NASA THP (TEES 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 EAR-9876800.

References

Allen, P. B., and J. W. Naney (1991), *Hydrology of Little Washita River Watershed: Oklahoma—Data and Analyses, ARS-90*, U.S. Dep. of Agric., U.S. Gov. Print. Off., Washington, D.C.

Anagnostou, E. N., A. J. Negri, and R. F. Adler (1999), Statistical adjustment of satellite microwave monthly rainfall estimates over Amazonia, *J. Appl. Meteorol.*, *38*(11), 1590–1598, doi:10.1175/1520-0450(1999)038<1590:SAOSMM>2.0.CO;2.

Atlas, D., D. Rosenfeld, and D. B. Wolff (1990), Climatologically tuned reflectivity-rain rate relations and links to area-time integrals, *J. Appl. Meteorol.*, *29*, 1120–1135, doi:10.1175/1520-0450(1990)029<1120:CTRRRR>2.0.CO;2.

Baigorria, G. A., J. W. Jones, D.-W. Shin, A. Mishra, and J. J. Obrien (2007), Assessing uncertainties in crop model simulations using daily bias-corrected regional circulation model outputs, *Clim. Res.*, *34*, 211–222, doi:10.3354/cr00703.

Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resour. Res.*, *37*(4), 937–947, doi:10.1029/2000WR900363.

Bellin, A., and Y. Rubin (1996), HYDRO_GEN: A spatially distributed random field generator for correlated properties, *Stochast. Hydrol. Hydraul.*, *10*(4), 253–278, doi:10.1007/BF01581869.

Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modeling: A review, *Hydrol. Processes*, *9*, 251–290, doi:10.1002/hyp.3360090305.

Calheiros, R. V., and I. L. Zawadzki (1987), Reflectivity rain-rate relationships for radar hydrology in Brazil, *J. Clim. Appl. Meteorol.*, *26*, 118–132, doi:10.1175/1520-0450(1987)026<0118:RRRRFR>2.0.CO;2.

Clapp, R. B., and G. M. Hornberger (1978), Empirical equations for some hydraulic properties, *Water Resour. Res.*, *14*, 601–604, doi:10.1029/WR014i004p00601.

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

Hansen, J. W., and A. V. M. Ines (2005), Stochastic disaggregation of monthly rainfall data for crop simulation studies, *Agric. For. Meteorol.*, *131*, 233–246, doi:10.1016/j.agrformet.2005.06.006.

Hashino, T., A. A. Bradley, and S. S. Schwartz (2007), Evaluation of bias-correction methods for ensemble streamflow volume forecasts, *Hydrol. Earth Syst. Sci.*, *11*, 935–950.

Ines, A. V. M., and J. W. Hansen (2006), Bias correction of daily GCM rainfall for crop simulation studies, *Agric. For. Meteorol.*, *138*, 44–53, doi:10.1016/j.agrformet.2006.03.009.

Jana, R. B., B. P. Mohanty, and E. P. Springer (2005), *Soil hydrologic properties for simulation of semi-arid river basin water balance: A report*, Tex. A&M Univ. and Los Alamos Natl. Lab., College Station, Tex.

Jana, R. B., B. P. Mohanty, and E. P. Springer (2007), Multiscale pedotransfer functions for soil water retention, *Vadose Zone J.*, *6*(4), 868–878, doi:10.2136/vzj2007.0055.

Khan, M. S., and P. Coulibaly (2006), Bayesian neural network for rainfall-runoff modeling, *Water Resour. Res.*, *42*, W07409, doi:10.1029/2005WR003971.

Kingston, G. B., M. F. Lambert, and H. R. Maier (2005), Bayesian training of artificial neural networks used for water resources modeling, *Water Resour. Res.*, *41*, W12409, doi:10.1029/2005WR004152.

Leij, F. J., N. Romano, M. Palladino, M. G. Schaap, and A. Coppola (2004), Topographical attributes to predict soil hydraulic properties along a hillslope transect, *Water Resour. Res.*, *40*, W02407, doi:10.1029/2002WR001641.

Mohanty, B. P., P. J. Shouse, D. A. Miller, and M. T. van Genuchten (2002), Soil property database: Southern Great Plains 1997 Hydrology Experiment, *Water Resour. Res.*, *38*(5), 1047, doi:10.1029/2000WR000076.

Natural Resources Conservation Service (2007), National soil survey handbook, *430-VI*, U.S. Dep. of Agric., Washington D.C. (Available at http://soils.usda.gov/technical/handbook/)

Neal, R. M. (1992), Bayesian training of backpropagation networks by the hybrid Monte Carlo method, *Tech. Rep. CRG-TR-92-1*, 21 pp., Dep. of Comput. Sci., Univ. of Toronto, Toronto, Ont., Canada.

Pachepsky, Y. A., D. Timlin, and G. Várallyay (1996), Artificial neural networks to estimate soil water retention from easily measurable data, *Soil Sci. Soc. Am. J.*, *60*, 727–773.

Pachepsky, Y. A., W. J. Rawls, and D. J. Timlin (1999), The current status of pedotransfer functions, their accuracy, reliability, and utility in field- and regional-scale modelling, in *Assessment of Non-Point Source Pollution in the Vadose Zone, Geophys. Monogr. Ser.*, vol. 108, edited by D. L. Corwin et al., pp. 223–234, AGU, Washington, D. C.

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

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

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

Schaap, M. G., and W. Bouten (1996), Modeling water retention curves of sandy soils using neural networks, *Water Resour. Res.*, *32*(10), 3033–3040, doi:10.1029/96WR02278.

Schaap, M. G., and F. J. Leij (1998), Database-related accuracy and uncertainty of pedotransfer functions, *Soil Sci.*, *163*(10), 765–779, doi:10.1097/00010694-199810000-00001.

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

Sharma, S. K., B. P. Mohanty, and J. Zhu (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, doi:10.2136/sssaj2005.0087.

van Genuchten, M. T., and F. J. Leij (1992), On estimating the hydraulic properties of unsaturated soils, in *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils: Proceedings of the International Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, edited by M. T. van Genuchten et al., pp. 1–14, Dep. of Soil and Environ. Sci., Univ. of Calif., Riverside.

Western, A. W., R. B. Grayson, and G. Blöschl (2002), Scaling of soil moisture, *Annu. Rev. Earth Planet. Sci.*, *30*, 149–180, doi:10.1146/annurev.earth.30.091201.140434.

Wierenga, P. J., D. Hudson, J. Vinson, M. Nash, A. Toorman, and R. G. Hills (1989), *Soil physical properties at the Las Cruces Trench site, NUREG/CR-5441*, U.S. Nucl. Regul. Comm., Rockville, Md.

Wierenga, P. J., R. G. Hills, and D. B. Hudson (1991), The Las Cruces Trench site: Experimental results and one-dimensional flow predictions, *Water Resour. Res.*, *27*, 2695–2705, doi:10.1029/91WR01537.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier (2002), Long-range experimental hydrologic forecasting for the eastern United States, *J. Geophys. Res.*, *107*(D20), 4429, doi:10.1029/2001JD000659.

Wösten, J. H. M., Y. A. Pachepsky, and W. J. Rawls (2001), Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics, *J. Hydrol.*, *251*, 123–150, doi:10.1016/S0022-1694(01)00464-4.

R. B. Jana and B. P. Mohanty, Department of Biological and Agricultural Engineering, Texas A&M University, 301C Coates Hall, College Station, TX 77843-2117, USA. (bmohanty@tamu.edu)

E. P. Springer, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87845, USA.