8-2014

Identifying sampling locations for field-scale soil moisture estimation using K-means clustering

Zachary J. Van Arkel
Iowa State University, zvanarkel@gmail.com

Amy L. Kaleita
Iowa State University, kaleita@iastate.edu

Identifying sampling locations for field-scale soil moisture estimation using K-means clustering

Abstract
Identifying and understanding the impact of field-scale soil moisture patterns is currently limited by the time and resources required to do sufficient monitoring. This study uses K-means clustering to find critical sampling points to estimate field-scale near-surface soil moisture. Points within the field are clustered based upon topographic and soils data and the points representing the center of those clusters are identified as the critical sampling points. Soil moisture observations at 42 sites across the growing seasons of 4 years were collected several times per week. Using soil moisture observations at the critical sampling points and the number of points within each cluster, a weighted average is found and used as the estimated mean field-scale soil moisture. Field-scale soil moisture estimations from this method are compared to the rank stability approach (RSA) to find optimal sampling locations based upon temporal soil moisture data. The clustering approach on soil and topography data resulted in field-scale average moisture estimates that were as good or better than RSA, but without the need for exhaustive presampling of soil moisture. Using an electromagnetic inductance map as a proxy for soils data significantly improved the estimates over those obtained based on topography alone.

Keywords
Clustering, K-means, Machine learning, Temporal stability

Disciplines
Agriculture | Bioresource and Agricultural Engineering | Water Resource Management

Comments
This article is from *Water Resources Research* 50 (2014): 7050–7057, doi:10.1002/2013WR015015. Posted with permission.

AGU PUBLICATIONS

Water Resources Research

TECHNICAL REPORTS: METHODS

10.1002/2013WR015015

Identifying sampling locations for field-scale soil moisture estimation using K-means clustering

Zach Van Arkel[1] and Amy L. Kaleita[1]

[1]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa, USA

Key Points:
- Soil moisture sampling points can be selected using topography and soils data
- Clustering method gives accurate results without prior observations of soil moisture

Correspondence to:
A. Kaleita,
kaleita@iastate.edu

Citation:
Van Arkel, Z., and A. L. Kaleita (2014), Identifying sampling locations for field-scale soil moisture estimation using K-means clustering, *Water Resour. Res.*, *50*, 7050–7057, doi:10.1002/2013WR015015.

Received 13 NOV 2013
Accepted 2 JUL 2014
Accepted article online 9 JUL 2014
Published online 19 AUG 2014

Abstract Identifying and understanding the impact of field-scale soil moisture patterns is currently limited by the time and resources required to do sufficient monitoring. This study uses K-means clustering to find critical sampling points to estimate field-scale near-surface soil moisture. Points within the field are clustered based upon topographic and soils data and the points representing the center of those clusters are identified as the critical sampling points. Soil moisture observations at 42 sites across the growing seasons of 4 years were collected several times per week. Using soil moisture observations at the critical sampling points and the number of points within each cluster, a weighted average is found and used as the estimated mean field-scale soil moisture. Field-scale soil moisture estimations from this method are compared to the rank stability approach (RSA) to find optimal sampling locations based upon temporal soil moisture data. The clustering approach on soil and topography data resulted in field-scale average moisture estimates that were as good or better than RSA, but without the need for exhaustive presampling of soil moisture. Using an electromagnetic inductance map as a proxy for soils data significantly improved the estimates over those obtained based on topography alone.

1. Introduction

The modeling of hydrologic processes is a key component in weather forecasting, crop growth simulation, and environmental performance prediction. Compared to other sinks in the hydrologic cycle, the volume of soil moisture (θ) is small, but it is of fundamental importance to many hydrological, biological, and biogeochemical processes. Thus, accurate θ information is of value to researchers in environmental modeling.

On a global scale, remote sensing of the Earth's brightness temperature can yield soil moisture estimates. The constant motion of the satellite allows coverage of large areas with frequencies adequate for weather and crop models needing the θ information. The launch of the SMOS (Soil Moisture Ocean Salinity) satellite and the upcoming launch of the SMAP (Soil Moisture Active Passive) satellite and Sentinel-1 will produce large amounts of θ data, and in the case of the latter two, with improved spatial and temporal resolution [*Vereecken et al.*, 2014]. However, in order to be confident in θ readings from these platforms, satellite estimates must be validated against measurements of "true" θ. Measuring "true" θ at a resolution comparable to a satellite pixel is nontrivial. Either spatially dense ground measurements are needed, which would require much time and money to collect, or representative sampling points throughout the landscape that adequately estimate θ at the satellite resolution must be identified.

Current methods for field and larger-scale estimation require extensive time series θ measurements from a network of in situ sensors. One well-documented method for finding sampling points suitable for estimating θ at the field scale is the Rank Stability Analysis (RSA), or temporal stability analysis, introduced by *Vachaud et al.* [1985]. Given spatially extensive time series θ data, sampling points within the field are identified as optimal sampling locations if they have the smallest standard deviation of the mean difference between point-θ and field-average θ. These points are determined rank stable because they have the smallest variance with respect to the field mean θ; that is, their ranking relative to the field mean does not change very much, regardless of the absolute value of soil moisture in the field. Besides the time and monetary resources required to collect the extensive spatiotemporal θ data for analysis, the reliance on empirical data is a downfall of the method. Because the method is based solely on empirical

data, the ability to recognize why certain locations are better to sample than others is limited to the sampling points used to find the rank stable locations. Additionally, *Yang* [2010] demonstrated that choosing random points from the sampling grid within the field was as reliable in field-scale θ estimation as the RSA method and in fact superior when there was significant year-to-year variability in RSA results. Thus, it is not clear that the RSA method gives a high return of information on the investment in data collection.

Numerous researchers have attempted to quantitatively link soil moisture spatial and temporal variability to topographic indices. Influences on soil moisture patterns include both soil physical properties and topography [*Chang*, 2001; *Romano and Palladino*, 2002]; topography, soils, vegetation, and climate [*Famiglietti et al*., 1998; *Yeh and Eltahir*, 1998; *Western et al*., 1999]; land use and soil type [*Qiu et al*., 2001]; and some have observed that the more important factor (soils or topography) changed during drying phases [*Famiglietti et al*., 1998]. However, as noted by *Vanderlinden et al*. [2012], "No clear dominant controls can be identified that are consistent throughout the literature."

The complexity and variation of temporal and spatial θ behavior, and the variety of factors having an impact on θ patterns, suggest machine learning methods in modeling θ behavior may be particularly effective. For example, *Ahmad et al*. [2010] used Support Vector Machine (SVM) to model soil moisture from remotely sensed estimates of rainfall and vegetation. *Srivastava et al*. [2013] investigated various machine learning techniques, including Artificial Neural Networks (ANN) and SVM to downscale SMOS soil moisture estimates using MODIS data. Machine learning techniques have been used for soil classification [e.g., *Shukla et al*., 2004; *Twarakavi et al*., 2010], but not yet for identifying optimal soil moisture sampling sites. Because machine learning techniques are designed to handle large amounts of data from a variety of different variables, the numerous factors impacting spatiotemporal θ patterns can be used as inputs into the algorithm.

The ultimate goal of this research is to develop, with easily attainable and time-invariant data, a practical plan for designating critical θ sampling points within agricultural fields that can accurately estimate the field-scale θ, and eventually help in bridging the gap between point measurements and remotely sensed θ data. First and for the sake of comparison, given past time series θ information, critical sampling points are found using K-means clustering algorithms, a machine learning approach described below, and used to find a field-scale θ estimation. Second, K-means clustering algorithms are used to find critical sampling points depending only on topographic and soil physical data as inputs. The estimates of field-average θ are compared to estimates found using sampling locations identified by the RSA method. Finally, the utility of the preceding methods are explored.

2. Methods

2.1. Location and Data

This study analyzed in situ θ measurements from the Brooks research field in Story County, Iowa, USA. Soil moisture measurement values were taken in a 300 $\times$ 250 m grid on the field during the growing seasons (summers) of 2004, 2005, 2007, and 2008. The spacing between each sampling point is 50 m. The elevation in the field varies by approximately 5 m and the grid covers a variety of different landscape positions throughout the field (Figure 1), including ridges and closed depressions. According to the National Cooperative Soil Survey, there are six main soil types in this field; however, NCSS soil type delineations are not precise at this fine resolution [e.g., *Brevik et al*., 2003].

The θ measurements were taken with an average interval of approximately 3 days. On each sampling day, measurements were made within a time window of at most 2 h, in order to reduce the θ differences due to drying. Each soil moisture observation is an average of three samples taken within a $\sim$0.5 m radius of each sampling location at a depth of 0–6 cm with a ThetaProbe moisture meter (Delta-T Devices, Cambridge UK, marketed in the United States by Dynamax, Inc., Houston, Texas). To reduce the effect of variations due to ridge and furrow patterns and plant-stem water transport [*Logsdon et al*., 2010], samples were taken between the crop rows; after seedbed preparation and planting in this conventionally farmed field, little to no plant debris was on the surface thus handling of ground cover was not needed between the rows. Values from the probe were then converted to estimates of volumetric θ using a calibration developed at this

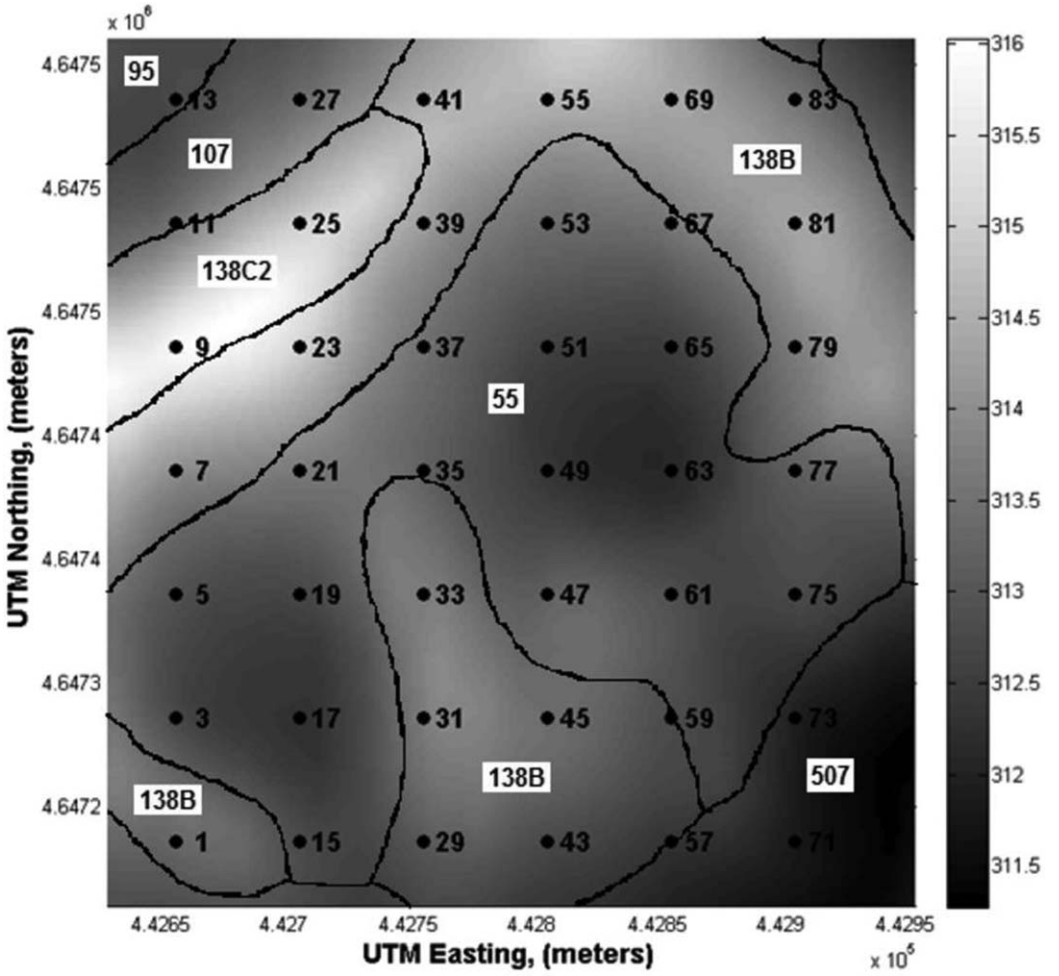

Figure 1. Brooks Field sampling grid with elevation (shading, in m) and soil types. Points are on 50 m intervals. Soil type indices, according to the National Cooperative Soil Survey: 55: Nicollet loam, 1−3% slopes; 95: Harps loam, 1−3% slopes; 107: Webster clay loam, 0−2% slopes; 138B: Clarion loam, 2−5% slopes; 138C2: Clarion loam, 5−9% slopes, moderately eroded; 507: Canisteo clay loam, 0−2% slopes.

site ($R^2 = 0.77$ compared to gravimetric-based observations) [*Kaleita et al.*, 2005]. The ThetaProbe measurements are considered the "true" θ values in this study.

In each season, data collection with the ThetaProbe began after planting of corn or soybeans (alternating years). In total, there were 99 measurement days during the study period. For reference, daily precipitation data were obtained from the Ames 8 WSW Station (UTM (Zone 15): 435912°E, 4652376°N; 42.0208° latitude, −93.7741° longitude), approximately 8 km from the Brooks field, from the National Oceanic and Atmospheric Administration website.

Sampling days from the time series data from each year were eliminated if any of the sampling locations had standing water at the time of collection, because ThetaProbe readings were not collected at those sites under such circumstances. Two days were eliminated from the 2004 temporal θ data, 4 days from 2005, 0 days from 2007, and 0 days from 2008. Eliminating these sampling days may have an impact on the observed pattern of θ behavior because the days with the highest average θ are not analyzed.

In the absence of high-resolution soils data, the electromagnetic inductance (EMI) is used as a proxy to identify changing soil properties. This noncontact sensor is sensitive to variations in several characteristics of the soil, including soil texture, soil moisture content, organic matter, and depth of clay pan. Consequently, the EMI data are not direct measures of any single soil property, but variations in EMI do reflect the heterogeneity in soil properties and for this reason are frequently used as a low-cost alternative to extensive soil sampling in applications where soil spatial variability is of interest [*Adamchuk et al.*, 2004]. Both horizontal (H-H)

and vertical (V-V) conductances in units of milliSiemens/meter were gathered using an EMI sled pulled by an all-terrain vehicle. EMI data were interpolated with inverse distance weighting for each of the θ sampling locations in the grid based upon the ~20 m resolution data found with the EMI sled.

Elevation data for the Brooks field were obtained using a GPS receiver mounted on the all-terrain vehicle that pulled the EMI sled. Using Surfer® (Golden Software, Inc., Golden, Colorado), a 10 m grid of elevation data was interpolated from this elevation data. Slope, planar curvature, and slope aspect were then derived using Surfer®. A 10 m grid was used based upon the finding by *Yang* [2010] that this scale was adequate to describe field-scale θ patterns at this site. The grid cell containing each of the sampling points was identified and the topographic indices for the sampling points were extracted from this information. At each location, then, the following data were available: a time series of θ observations, EMI (two polarizations), elevation, slope, planar curvature, and slope aspect.

2.2. RSA

The method introduced by *Vachaud et al.* [1985] was employed to compare the identification and prediction of sampling points from the methods proposed in this paper. Using time series θ data, the Rank Stability Analysis method finds the mean and standard deviation of relative differences from the areal mean for each sampling point. Points with small standard deviation are determined temporally rank stable because the differences in their θ behavior with respect to the larger-scale average vary the least in time; these points are thus considered optimal sampling locations (OSLs). Time series data from the 2004 season were used to find sampling points from the grid with the smallest standard deviation of mean relative difference to the field average. The temporal θ data from 2004 only were used, based on the assumption that in practice, RSA would be implemented on an initial time series of data to identify sampling locations for use in the future. Given the θ data for n sampling location(s) with the smallest standard deviation of the relative difference, the field mean θ for any observed day j is found with the following equation:

$$\overline{\theta}_j^{est} = \sum_{i=1}^{n} \frac{\theta_{OSL_i}}{1 + \overline{\delta}_{OSL_i}} \tag{1}$$

where θ_{OSL_i} is the measured volumetric soil moisture content from the ith OSL on the jth day, $\overline{\delta}_{OSL_i}$ is the mean relative difference of the ith OSL from the field-average θ, and $\overline{\theta}_j^{est}$ is the estimated mean soil moisture from these OSL(s) on the jth day. The mean relative difference of each point is found with the following equation:

$$\overline{\delta}_{OSL_i} = \left[\sum_{j=1}^{n} \frac{\theta_{OSL_i} - \overline{\theta}_j}{\overline{\theta}_j} \right] / n \tag{2}$$

where θ_{OSL_i} is the measured volumetric soil moisture content from the ith OSL on the jth day and $\overline{\theta}_j$ is the mean of the measured volumetric soil moisture content from all of the sampling points on the jth day.

2.3. K-Means Clustering

K-means clustering is used to separate the data into different clusters containing points with similar characteristics. In the K-means algorithm, each data location is initially assigned to one of k clusters at random. The centroid location (in n dimension, where n is the number of attributes in the input vector for each point) is computed for each of the k clusters. The distance from each point to each centroid is then computed by finding the smallest Euclidean distance between the input vector and the centroid vector. Each point is then reassigned to the cluster with the nearest centroid. This process continues until there is no change in cluster membership given additional iterations of the algorithm. Readers are referred to *MacQueen* [1967] for further explanation of the K-means algorithm.

2.4. Data Analysis

Temporal θ data from 2004 and topographic and EMI data were used to construct three matrices: $\mathbf{M}_\theta$, $\mathbf{M}_T$, and $\mathbf{M}_E$. Each matrix contained the points in the sampling grid as rows and the rows then represent the input vectors into the algorithms. The matrix $\mathbf{M}_\theta$ contained the 2004 θ sampling days as columns, with θ observations as data elements. Thus, $\mathbf{M}_\theta$ is a 42 $\times$ 24 matrix of θ values (corresponding to 42 sampling locations over 24 sampling days from the 2004 season). This is the data used to find optimal sampling locations based upon RSA, and using K-means on the soil moisture data.

The columns of $\mathbf{M}_T$ contained elevation, slope, slope aspect (flow direction or downhill direction), and planar curvature (curvature in the direction perpendicular to the flow). The columns of $\mathbf{M}_E$ contained elevation, slope, slope aspect, planar curvature, H-H EMI, and V-V EMI. Thus, $\mathbf{M}_T$ is a 42 $\times$ 4 matrix, and $\mathbf{M}_E$ is a 42 $\times$ 6 matrix. These data were classified into clusters using the K-means approach.

The centroid vector of each cluster in each method was then identified. Using the Euclidean distance formula, the input vector (corresponding to a single sampling location) with the smallest distance from each centroid was identified. This input vector (sampling location) was deemed the best matching unit (BMU) to the cluster centroid. These BMUs were then used as the critical sampling locations identified for each data set. Identification of sampling points using K-means on $\mathbf{M}_T$ and $\mathbf{M}_E$ are thus independent of any soil moisture observations.

To find the estimated average of the field θ (grid average) using the sampling points identified by the clustering approach, a weighted average was found using the BMUs and the number of points in the corresponding cluster:

$$\overline{\theta}_j^{est} = \frac{\sum_{i=1}^{k} \theta_{BMU_{ij}} * n_i}{N} \tag{3}$$

where $\overline{\theta}_j^{est}$ is the estimated mean θ on the *j*th day, $\theta_{BMU_{ij}}$ is the θ value on the *j*th day for the BMU to the centroid of the *i*th cluster, n_i is the number of sampling locations in the *i*th cluster, N is the total number of sampling points, and *k* is the number of clusters.

Finally, random points were also selected for the purpose of comparison. One-hundred random realizations of *k* points were generated, and observed soil moisture at the same *k* random points on each day were averaged together to generate the estimate of field average moisture content.

In this study, we explored selection of 1, 2, 3, and 4 observation locations; in the K-means approach this corresponds to 1, 2, 3, and 4 clusters, respectively. This represents the range of optimal sampling location numbers from the ideal of one (only one point needed to adequately capture the field mean) through roughly 10% of the observed data locations.

To compare the accuracies of the estimated field average from the different methods, the average bias (AB), estimation coefficient of determination (R^2), and correlation coefficient (r) were calculated, comparing the estimated field averages to the corresponding "true" field average, which was assumed to be the arithmetic average of all the observations for that day. Average bias indicates the extent to which the method consistently over or underestimates the field average. Estimation R^2 indicates to what extent the method is better than simply using the long-term field average; a positive value indicates the method improves over the long-term field mean (a maximum value of one indicates that the method perfectly matches the observed data), whereas a negative value indicates that the method is worse than assuming the long-term field mean. Estimation R^2 is calculated from the sums of squared errors as

$$R^2 = 1 - \frac{\sum_{j=1}^{J} \left(\bar{\theta}_j - \bar{\theta}_j^{est}\right)^2}{\sum_{j=1}^{J} \left(\bar{\theta}_j - \overline{\bar{\theta}_j}\right)^2} \tag{4}$$

where $\bar{\theta}_j$ is the mathematical average of observed θ, and $\bar{\theta}_j^{est}$ is the estimated average θ on the *j*th day, and $\overline{\bar{\theta}_j}$ is the mean of all the daily observed averages. The correlation coefficient indicates the relative agreement between the method results and the actual field mean, or how well the method is able to match trends in the truth data.

3. Results and Discussion

Table 1 gives the points identified by each of the approaches described above. Points selected by RSA are equally weighted when averaged to the field scale, where points selected by K-means are weighted proportionally to the size of the cluster they represent. It should be noted that because the K-means approach uses random starts to the algorithms, the results may converge to somewhat different results each time the algorithm is run. For these data, the one-point and two-point results were always the same, but the three-

Table 1. Points Identified for Sampling by the Rank Stability Analysis and K-Means Algorithms Using $\mathbf{M}_\theta$ (2004 Soil Moisture) as Input Data, and for the K-Means Algorithms Using $\mathbf{M_T}$ (Topo) and $\mathbf{M_E}$ (Topo/EMI)

Method	One Point	Two Points	Three Points	Four Points
RSA M_θ	55	23, 55	3, 59, 61	11, 47, 51, 61
K-means M_θ	77	29, 47	65, 67, 83	3, 47, 55, 77
K-means M_T	77	51, 67	21, 31, 51,	43, 59, 67, 81
K-means M_E	21	35, 67	51, 59, 67	1, 15, 51, 67

point and four-point selections were slightly different each time, with three or four of the same six points being selected each time. The selections given in Table 1 are one realization, but other realizations may have had one or two different points selected. This behavior is less likely with larger data sets, but when the number of clusters is relatively large compared to the total number of data, small differences in the cluster membership result in different sampling points being closest to the cluster centroid.

The RSA method identified sampling locations that were, for the most part, different from those identified through the K-means approach. This was even the case for K-means on $\mathbf{M}_\theta$, the same data as used in the RSA method. Points 47, 51, and 59, however, were selected by both RSA and one or more K-means. The K-means approaches on all data sets identified sampling locations that were different depending on the input data used, but points 51 and 67 were selected four and five times, respectively. The single-point selections for K-means on $\mathbf{M}_\theta$ and $\mathbf{M_T}$ were also the same.

Table 2 gives the performance indices from all methods. For the RSA method, with increasing number of points used to estimate the field mean θ in validation, the performance improves: AB decreases and R^2 increases. For the K-means on $\mathbf{M}_\theta$ and $\mathbf{M_T}$, however, this was not the case; K-means on $\mathbf{M}_\theta$ gave results that were fairly consistent from one to four sampling locations, while K-means on $\mathbf{M_T}$ gave the worst results with two samples.

None of the single-point approaches performed well on the validation data, but the RSA approach performed the worst, with the lowest R^2 and correlation, and largest AB of all methods. For three and four-point samplings, RSA generally had better performance across all metrics than K-means on the moisture data $\mathbf{M_\theta}$, and similar performance to K-means on the topographic and EMI data $\mathbf{M_E}$.

Table 2. Average Bias (AB), Estimation R^2, and Correlation Coefficient for Mean Field θ Estimate From Critical Sampling Points Identified With M_θ (2004 Soil Moisture), M_T (Topo), and M_E (Topo/EMI), Validated Against Soil Moisture From 2005, 2007, and 2008[a]

	Number of Points	RSA $M_{\theta\ \text{Calibration}}$	RSA $M_{\theta\ \text{Validation}}$	K-means $M_{\theta\ \text{Calibration}}$	K-means $M_{\theta\ \text{Validation}}$	K-means $M_{T\ \text{Validation}}$	K-means $M_{E\ \text{Validation}}$	Random $_{\text{Validation}}$
AB (cm^3/cm^3)	1	−0.000	0.025	−0.005	−0.010	−0.010	−0.004	−0.003 (−0.027 to −0.061)
	2	0.000	0.022	−0.001	0.004	0.018	0.002	−0.005 (−0.025 to 0.041)
	3	0.000	0.003	0.001	−0.011	0.012	0.005	−0.001 (−0.022 to 0.041)
	4	−0.000	0.005	−0.001	0.001	−0.005	0.004	−0.002 (−0.017 to 0.027)
R^2	1	0.966	−0.325	0.957	0.668	0.668	0.552	0.382 (−4.91 to 0.796)
	2	0.981	0.110	0.975	0.734	0.265	0.798	0.641 (−1.41 to 0.895)
	3	0.994	0.813	0.985	0.689	0.599	0.794	0.710 (−1.22 to 0.917)
	4	0.997	0.867	0.982	0.771	0.869	0.872	0.831 (−0.100 to 0.934)
Correlation	1	0.935	0.802	0.927	0.904	0.904	0.879	0.833 (0.504 to 0.916)
	2	0.970	0.896	0.946	0.899	0.873	0.911	0.913 (0.790 to 0.963)
	3	0.989	0.914	0.972	0.924	0.920	0.925	0.930 (0.761 to 0.966)
	4	0.994	0.961	0.967	0.942	0.958	0.954	0.953 (0.805 to 0.976)

[a]For the random method, the median value from 100 realizations is given, as well as the range.

Compared to the other approaches, K-means on the topography data gave erratic results, with two-point estimation performing considerably worse across all metrics than one, three, or four-point estimation. Overall, the performance of K-means on $\mathbf{M}_T$ suggests that these topography data alone are not sufficient to confidently select critical sampling points, and that if topography data are the only data available, a higher number of sampling locations may need to be selected. On the other hand, K-means on $\mathbf{M}_E$ gave generally good results that consistently improved with additional sampling locations included. While the single-point sampling identified by K-means on $\mathbf{M}_E$ had lower estimation R^2 than other single-point methods, the bias was low and the correlation was better than RSA and only slightly worse than the other two K-means approaches.

Random sampling creates somewhat erratic results. In general, the median values of average bias are similar to those of the other methods, though some realizations result in AB values as high as 0.04–0.06 cm^3/cm^3, more than double the AB magnitude of other methods. Median correlation values for random sampling are as high as or higher than other methods, but the low end of the range is notably lower than the correlation with other methods. Median R^2 values from random sampling are lower than the best-performing alternative method, and the worst R^2 values are less than zero. Overall, as one might expect, random sampling can give good results, but can also give results that are less reliable than other methods. In the absence of any site data, random sampling can be a reasonable choice, but these limitations should be recognized when using the resulting data.

The statistical indices of K-means on $\mathbf{M}_E$ support the use of the K-means method for identifying critical sampling points from topography and soils data together. Because this approach requires no a priori observations of soil moisture, it could be applied to any new study area with a one-time collection of topography and EMI, neither of which are costly nor time consuming.

4. Conclusion

The RSA method performed well on the 2004 calibration data, but when applied to validation data from subsequent seasons, did not perform as well when only one or two points were selected. RSA results for three and four points were generally good, having estimation R^2 values above 0.8 and correlations above 0.9. However, RSA requires a substantial amount of soil moisture data before the method can even be used, making it costly to implement in practice.

The K-means approach using soil moisture data identified one and two sampling locations that, when weighted to reflect their relative representativeness of the whole-field data set, provided a better estimate of the field average than RSA for the same number of sampling points, but performed generally similarly to RSA for three and four points. However, this approach still requires a large calibration data set.

The K-means approach on topography and soils data resulted in field average estimates that were much better than the RSA approach with one and two sampling locations, and similar performance to RSA with three and four sampling locations. A major advantage of the K-means approach on these data is that it does not require a priori observations of soil moisture. Topography data are readily obtained at relatively low cost from a number of sources, including on-site survey, LiDAR, and other sources, and topographic indices such as slope, aspect, and curvature, as used here, can be derived. Soils data at the field scale are more costly to obtain if relying on physical soil samples. However, in this study, use of a low-cost electromagnetic inductance map as a proxy for soil physical data was sufficient. Using EMI as a proxy for soils data in the K-means approach significantly improved the estimates over those obtained based on topography alone.

Another advantage of the K-means approach is that, because the clusters are built on multidimensional data, any a priori knowledge of key drivers of soil moisture patterns could be used to determine the most appropriate input data to the clustering algorithm. A test of this approach on existing data sources for which a full analysis of the controls of local soil moisture patterns would be useful to confirm or refute this potential advantage.

One limitation of this study is that the clustering algorithm generated slightly different results for three and four sampling points. However, in practice, this outcome would be less likely if the K-means method was applied to topography and soils data for the purpose of identifying representative sampling locations. In

application of this approach from high-resolution topography and soils data in the absence of calibration or validation data, a larger number of sampling points per unit area would be used compared to what was used here, where we constrained our analysis only to points with corresponding soil moisture observations. Without that constraint, the number of data locations would be much larger, and the K-means approaches would be expected to converge to the same best matching units for larger number of clusters.

In this study, the K-means approach was able identify critical sampling points using topographic and soil physical data that can be used to estimate mean field-scale θ values. Results suggest that fewer critical sampling points are needed if EMI data are included in the field physical data for identifying critical sampling points as opposed to only using topographic data to identify sampling points.

Acknowledgments
Data in this paper and MATLAB scripts for generating the results are available upon request from the corresponding author.

References

Adamchuk, V. I., J. W. Hummel, M. T. Morgan, and S. K. Upadhyaya (2004), On-the-go soil sensors for precision agriculture, *Comput. Electron. Agric.*, *44*(1), 71–91, doi:10.1016/j.compag.2004.03.002.

Ahmad, S., A. Kalra, and H. Stephen (2010), Estimating soil moisture using remote sensing data: A machine learning approach, *Adv. Water Resour.*, *33*(1), 69–80, doi:10.1016/j.advwatres.2009.10.008.

Brevik, E. C., T. E. Fenton, and D. B. Jaynes (2003), Evaluation of the accuracy of a central Iowa soil survey and implications for precision soil management, *Precis. Agric.*, *4*(3), 331–342, doi:10.1023/A:1024960708561.

Chang, D. H. (2001), Analysis and modeling of space-time organization of remotely soil moisture, PhD thesis, Univ. of Cincinnati, Cincinnatti, Ohio.

Famiglietti, J., J. Rudnicki, and M. Rodell (1998), Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, *J. Hydrol.*, *210*(1-4), 259–281, doi:10.1016/S0022–1694(98)00187-5.

Kaleita, A., J. Heitman, and S. Logsdon (2005), Field calibration for the Theta probe for Des Moines Lobe soils, *Appl. Eng. Agric.*, *21*(5), 865–870, doi:10.13031/2013.19714.

Logsdon, S. D., T. J. Sauer, G. Hernandez-Ramirez, J. L. Hatfield, A. Kaleita, and J. H. Prueger (2010), Effect of corn or soybean row position on soil water, *Soil Sci.*, *175*(11), 530–534, doi:10.1097/SS.0b013e3181fae168.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observation, in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, Univ. of Calif. Press, Berkeley, Calif.

Qiu, Y., B. Fu, J. Wang, and L. Chen (2001), Soil moisture variation in relation to topography and land use in a hillslope catchment of the Loess Plateau, China, *J. Hydrol.*, *240*(3-4), 243–263, doi:10.1016/S0022–1694(00)00362-0.

Romano, N., and M. Palladino (2002), Prediction of soil water retention using soil physical data and terrain attributes, *J. Hydrol.*, *265*(1-4), 56–75, doi:10.1016/S0022–1694(02)00094-X.

Shukla, M. K., B. K. Slater, R. Lal, and P. Cepuder (2004), Spatial variability of soil properties and potential management classification of a chernozemic field in lower Austria, *Soil Sci.*, *169*(12), 852–860.

Srivastava, P. K., D. Han, M. R. Ramirez, and T. Islam (2013), Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application, *Water Resour. Manage.*, *27*(8), 3127–3144, doi:10.1007/s11269-013-0337-9.

Twarakavi, N. K. C., J. Simunek, and M. G. Shaap (2010), Can texture-based classification optimally classify soils with respect to soil hydraulics?, *Water Resour. Res.*, *46*, W01501, doi:10.1029/2009WR007939.

Vachaud, G., A. Passerat de Silans, P. Balabanis, and M. Vauclin (1985), Temporal stability of spatially measured soil water probability density function, *Soil Sci. Soc. Am. J.*, *49*(4), 822–828, doi:10.2136/sssaj1985.03615995004900040006x.

Vanderlinden, K., H. Vereecken, H. Hardelauf, M. Herbst, G. Martinez, M. H. Cosh, and Y. A. Pachepsky (2012), Temporal stability of soil water contents: A review of data and analyses, *Vadose Zone J.*, *11*(4), doi:10.2136/vzj2011.0178.

Vereecken, H., J. A. Huisman, Y. Pachepsky, C. Montzka, J. van der Kruk, H. Bogena, L. Weihermüller, M. Herbst, G. Martinez, and J. Vanderborght (2014), On the spatio-temporal dynamics of soil moisture at the field scale, *J. Hydrol.*, *516*, 76–96, doi:10.1016/j.jhydrol.2013.11.061.

Western, A. W., R. B. Grayson, G. Bloschl, G. R. Willgoose, and T. A. McMahon (1999), Observed spatial organization of soil moisture and its relation to terrain indices, *Water Resour. Res.*, *35*(3), 797–810, doi:10.1029/1998WR900065.

Yang, L. (2010), Spatio-temporal patterns of field-scale soil moisture and their implications for in situ soil moisture network design, PhD thesis, Iowa State Univ., Ames.

Yeh, P., and E. Eltahir (1998), Stochastic analysis of the relationship between topography and the spatial distribution of soil moisture, *Water Resour. Res.*, *34*(5), 1251–1263, doi:10.1029/98WR00093.