A method for upscaling *in situ* soil moisture measurements to satellite footprint scale using Random Forests

Daniel Clewley, Jane B. Whitcomb, Ruzbeh Akbar, Agnelo R. Silva, Aaron Berg, Justin Adams, Todd Caldwell, Dara Entekhabi, Mahta Moghaddam

Abstract

Geophysical products generated from remotely sensed data require validation to evaluate their accuracy. Typically *in situ* measurements are used for validation, as is the case for satellite-derived soil moisture products. However, a large disparity in scales often exists between *in situ* measurements (covering meters to 10s of meters) and satellite footprints (often hundreds of meters to several kilometers), making direct comparison difficult. Before using *in situ* measurements for validation, they must be 'upscaled' to provide the mean soil moisture within the satellite footprint. There are a number of existing upscaling methods previously applied to soil moisture measurements, but many place strict requirements on the number and spatial distribution of soil moisture sensors difficult to achieve with permanent/semi-permanent ground networks necessary for long-term validation efforts.

A new method for upscaling is presented here, using Random Forests to fit a model between *in situ* measurements and a number of landscape parameters and variables impacting the spatial and temporal distributions of soil moisture. The method is specifically intended for validation of the NASA Soil Moisture Active Passive (SMAP) products at 36, 9 and 3 km scales. The method was applied to *in situ* data from the SoilSCAPE network in California, validated with data from the SMAPVEX12 campaign in Manitoba, Canada with additional verification from the TxSON network in Texas. For the SMAPVEX12 site the proposed method was compared to extensive field measurements and was able to predict mean soil moisture over a large area more accurately than other upscaling approaches.

Index Terms

Soil Moisture, Random Forests, Scaling, SMAP, AirMOSS

I. Introduction

VALIDATION of satellite and airborne geophysical data products is a critical task in supporting the success of Earth science missions. However, this task is often hindered by two important issues: (a) the difference in scales between coarse resolution ($\sim$ km scale) mission products and localized point-scale *in situ* sensor measurements and (b) the inability of *in situ* networks to fully capture the spatial heterogeneity of geophysical variables. The variable of interest in this paper is soil moisture, though the analysis and methods presented could as easily apply to other products. In particular to validate coarse-resolution soil moisture estimates from missions such as the NASA Soil Moisture Active Passive (SMAP) [1] and the European Space Agency's (ESA) Soil Moisture Ocean Salinity (SMOS) missions [2], soil moisture estimates from a broad and diverse set of validation sites must be appropriately combined to develop scale-representative soil moisture aggregates. This is to say that point-wise *in situ* observations have to be upscaled to properly estimate the true mean of the soil moisture fields at the coarse-scale (several kilometers) of the satellite products.

In situ networks typically consist of a number of soil moisture sensor probes installed throughout a study domain. Often the probes are installed only within a few centimeters of the surface, but in some instances, they are also distributed along a vertical profile to capture the so-called root-zone soil moisture. Each profile may comprise 3 – 5 individual probes installed at different depths within the soil, with probe measurements typically collected every 15 – 30 minutes. Given the small scale (tens of meters) variations in soil moisture expected at many locations, the measurements collected by each soil moisture sensor profile generally only characterize soil moisture in the immediate vicinity of the sensors. Satellite soil moisture products, on the other hand, due to the spaceborne observatory's orbit geometry and antenna field of view, are typically on the order of at least a few kilometers. For example, radiometer-only soil moisture estimates from SMAP (L2SM_P) are reported at 36 km resolution and estimates from SMOS are on the order of 40 km. *In situ* networks must be specifically deployed and tailored towards a mission's validation strategy to ensure proper alignment of sensor with validation pixels and proper spatial sampling of sub-pixel variability within that pixel [3].

This work was supported by a grant from the National Aeronautics and Space Administration, Earth Science Technology Office, Advanced Information Systems Technologies program.

D. Clewley is with Plymouth Marine Laboratory, UK, PL1 3DH (email dac@pml.ac.uk). J. Whitcomb and M. Moghaddam are with the University of Southern California, CA, USA, 90089. R. Akbar and D. Entekhabi are with Massachusetts Institute of Technology, MA, USA, 4307, A. R. Silva is with METER Group Inc., Pullman, WA, USA, A. Berg is with Department of Geography, University of Guelph, Ontario, Canada, J. Adams is with Wilfrid Laurier University, Ontario, Canada, T. Caldwell is with University of Texas at Austin.

Surface soil moisture content ($<$ 5 cm depth) has a high degree of spatial variation and is in general a complex function of local topography, soil texture, precipitation, and vegetation cover. Therefore, uniformly distributed *in situ* sensors are likely to lack spatial representativeness and may not fully capture soil moisture heterogeneity and dynamics within a study domain. Crow *et. al.* [4] in a comprehensive review of soil moisture upscaling strategies highlighted the fact that progressively a larger number of point-scale sensors ($N >$10) are needed to achieve field aggregates of soil moisture close to the true mean which also meet the SMAP mission Root Mean Squared Error (RMSE) accuracy goals of 0.04 m^3m^{-3}. However, from a practical point of view, installation and maintenance of many spatially uniformly distributed sensors becomes time consuming and costly.

In the past, numerous soil moisture upscaling methods, with varying degrees of complexity and accuracy, have been introduced and are divided into two general classes of techniques: (1) sensor-only methods and (2) model-driven approaches. The most simplistic sensor-only upscaling approach is the arithmetic mean of all existing sensors within the study domain or satellite footprint. Unless a large number of well-distributed sensors exists, linear averaging of sensors within a site does not capture the true field mean. Block Kriging [5], on the other hand, by estimating correlation and covariance between individual sensors, usually via time-series analysis, is able to derive non-equal linear weights to calculate upscaled soil moisture. Temporal stability concepts, introduced in [6], attempt upscaling soil moisture by first identifying regions with persistent soil moisture patterns then aggregating a select subset of sensors to field size, typically by simple averaging. The challenge here is identification of desirable sensor locations prior to installation. Inverse distance weighting, forms weighted average of sensor measurements, using weights that diminish as function of distance from the sensors. Theissen polygons are formed by joining sensor nodes with line segments, finding perpendicular bisectors of those segments, and extracting a set of polygons with sides formed by perpendicular bisectors.

Sensor-data driven upscaling methods incorporate static and dynamic spatial data, field campaign data, and sometimes land surface models into an overall model of soil moisture in the study domain. These approaches have the potential to produce significantly more accurate upscaled soil moisture estimates than can be obtained using sensor-only approaches.

Distributed land surface models such as TOPmodel-based Land Atmosphere Transfer Scheme (TOPLATS; [7], [8]) are an example of a model-driven approach. Such models incorporate a number of static and dynamic data sources and provide spatially distributed estimates of soil moisture from which upscale soil moisture estimates can be generated. Although distributed land surface models are able to capture the spatial pattern of soil moisture errors in model assumptions and input parameters can lead to discrepancies when compared to measured values [9]. The accuracy can be improved by constraining modeling with *in situ* measurements [10] but there is still a requirement for the model to be representative of the underlying spatial patterns of soil moisture.

When collectively considering (a) resolution disparity, (b) sensor representativeness, and (c) soil moisture spatial variability, the need for new and robust upscaling techniques becomes clear, and development of such approaches are of keen interest. Furthermore, methods in which other static and dynamic ancillary data sources or data layers, are included along with *in situ* sensors to upscale soil moisture can overcome the deficiencies of sensor-only upscaling techniques and are not impacted by modeling errors. Preventing propagation of land surface modeling errors into upscaled estimates reduces the need for additional error and uncertainty analysis and mitigation.

With this in mind, the work here will present and discuss a novel upscaling strategy using the method of Random Forests regression [11]. This algorithm is capable of modeling complex non-linear systems and is able to simultaneously handle continuous data (e.g., temperature, elevation) and categorical data (e.g., landcover type). This is the first demonstration and application of the Random Forests algorithm to soil moisture upscaling.

In Section II we present an overview of the study sites along with their spatial domain and *in situ* network characteristics. The Random Forests regression scheme and methodology is outlined in Section III, including an examination of the ingestion of closely related geophysical data layers along with microwave remote sensing data. Upscaled soil moisture results at the study sites are then presented and discussed in Section IV, along with independent validation of the methodology.

II. Study Sites

A. Tonzi Ranch, California, United States

1) Site Characteristics: The greater Tonzi Ranch domain, located in north-central California (centered at 30.3°N, 120.9°W), is a 36 km $\times$ 36 km domain. Predominantly a grassland/herbaceous region mixed with large patches of forests (dominated by blue oak) and shrubs. The climate is classified as mild Mediterranean with hot and dry summers (Csa) and a Woody Savanna International Geosphere-Biosphere Program (IGBP) classification. The annual mean temperature is 16°C with 560 mm mean annual precipitation. Generally, grass and herb growth cycles are limited to the rainy seasons (October to May). Surface soil texture throughout the site is approximately 20 % clay and 40 % sand. Since 2001, the primary Tonzi ranch site has been part of the AmeriFlux network, and includes many long-term ecological and climatological monitoring sensors. Figure 1 shows the contrast in understory growth between hot and dry summer months and wetter winter periods.

Tonzi Ranch also includes multiple independently managed soil moisture stations. Of these, three stations operated by the Oregon State University (OSU) are part of the NASA AirMOSS mission calibration and validation sites and measure soil moisture at depths of 5 – 40 cm within the soil column. Soil moisture data from these stations are used as independent validation points of the proposed up-scaling scheme in this work and discussed in Sections III and IV.

(a) (b)

Fig. 1. Nodes installed at the Tonzi ranch site showing conditions for a) winter and b) summer.

TABLE I
SUMMARY OF SOILSCAPE SENSOR NETWORK IN CALIFORNIA

Site Name	Coordinates	Land Cover	Number of Nodes	Node Density (ha^{-1})	Maximum Distance from LC (m)
Tonzi Ranch	(38.43, -120.96)	Woody Savanna	19	1.5	340
BLM* I	(38.39, -120.90)	Woody Savanna	17	4.7	207
BLM II	(38.38, -120.90)	Woody Savanna	16	5.0	145
BLM III	(38.46, -120.99)	Woody Savanna	3	3.0	50
New Hogan I	(38.17, -120.80)	Shrubs	14	1.7	257
New Hogan II	(38.17, -120.80)	Shrubs/Grassland	18	1.0	346
Terra d'Oro	(38.50, -120.8)	Vineyard	27	1.5	317

* BLM: Bureau of Land Management

2) SoilSCAPE Network: The Soil Moisture Sensing Controller and oPtimal Estimator (SoilSCAPE) project was initiated to provide *in situ* measurements of soil moisture in support of validation of spaceborne and airborne products with a focus on novel wireless sensor network technologies [12], [13]. A main objective of SoilSCAPE is to demonstrate that by utilizing smarter network technologies and optimally placing sensors, representative measurements of soil moisture could be obtained, using a small number of sensors at a range of spatial scales, and with lower set up and maintenance costs than traditional networks.

To this end, SoilSCAPE has deployed multiple sub-networks of wireless soil moisture nodes throughout the greater Tonzi ranch site in North-central California (Figure 2). Each sub-network, consists of a cluster of sensors nodes (10 – 30) which are also referred to as End Devices (EDs). Each node incorporates 3 – 4 soil moisture sensors nominally installed at 5, 20 and 40 cm depths within the soil column. EDs wirelessly communicate with a Local Coordinator (LC) on a nominal 20 min sampling cycle but also have the capability for variable adaptive scheduling. The LC from each network uploads the soil moisture measurements to a data server. The server performs data quality control and web publishing in near-real time. As of September 2016, SoilSCAPE includes seven networks clusters and 114 EDs, in North-central California. A summary of the network features is given in Table I. Placement of sensors for the entire SoilSCAPE network was primarily driven by two factors (a) SMAP soil moisture product validation efforts and (b) within-network sensor placement considerations. The SMAP mission was intended to provide soil moisture estimates at 3 (radar-only), 9 (radar-radiometer) and 36 km (radiometer-only) spatial resolutions. Each SoilSCAPE sub-network (Table I) was initially deployed to sample 3 or 9 km region. Within each network, sensor placement was driven by the need to sample representative vegetation and soil types. Additional constraints such as range of wireless communication, land use permissions, dense vegetation and topography were also considered. With the SMAP radar ceasing operation as of 7^{th} July 2015 and high resolution (3 or 9 km) soil moisture products no being produced. The greater Tonzi Ranch site (Tonzi Ranch, BLM I, II and III) sites have collectively been utilized within the proposed up-scaling scheme to address the SMAP radiometer-only (36 km) soil moisture product validation.

B. Winnipeg, Manitoba, Canada

1) Site Characteristics: Located in southern Manitoba, Canada, this site was the focus of the SMAP Validation Experiment 2012 (SMAPVEX12; [14]). The study domain covers an area of approximately 13 km $\times$ 70 km (Figure 3) and is within the Canadian Red River Watershed. The region is predominantly agricultural, with some permanent wetlands and mixed forests. Dominant crops include soybean, canola, corn, and cereal grains. The primary tree types in the forested regions are Aspen and Bur oak with a dense understory. Seeding typically occurs in the months of April or May with harvesting in August or September. Across the domain, soil texture, especially clay content varies significantly with very high clay fractions on the eastern edge with a sharp transition to loamy soils towards the western edge. The region any lacks significant topography.

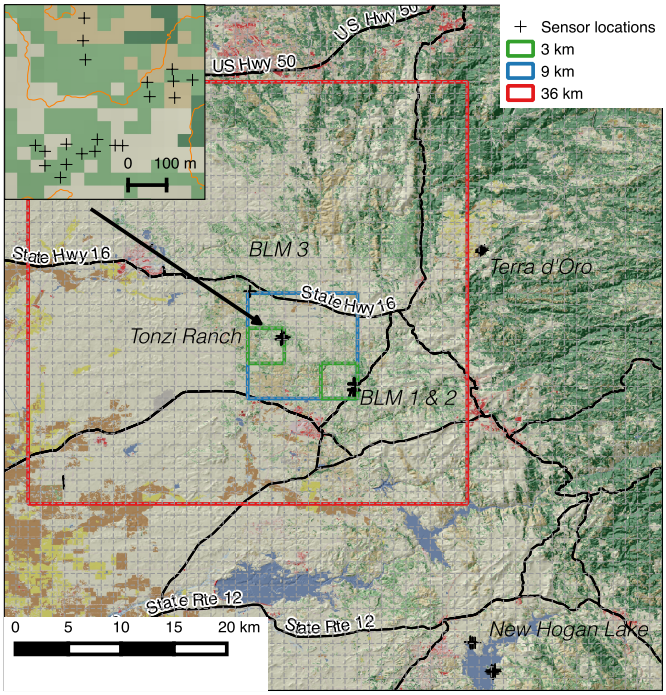

Fig. 2. Sensors deployed as part of SoilSCAPE network in central California with footprints for SMAP soil moisture retrievals. Insert shows placement of individual sensors within Tonzi Ranch site. Background is NLCD landcover classification.

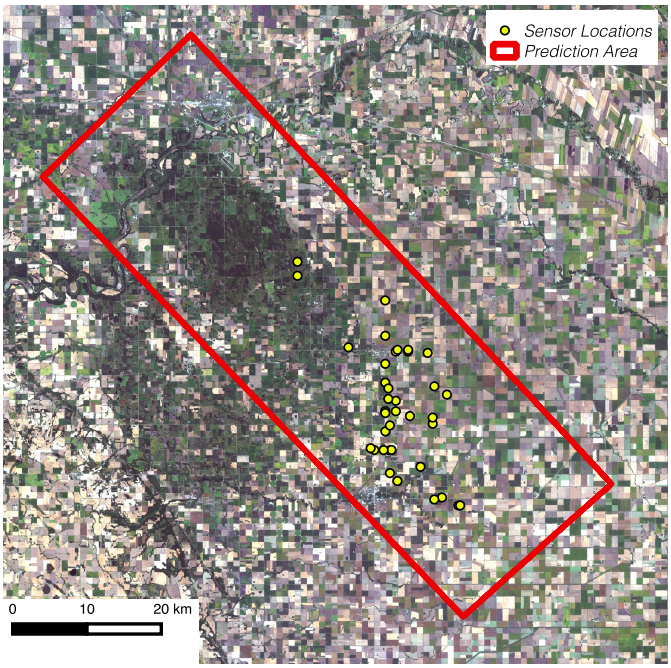

Fig. 3. Location of USDA sensors within the SMAPVEX12 site near Winnipeg, Canada. Shown over Landsat 8 true-color data.

2) SMAPVEX12 Campaign: The SMAPVEX12 campaign was carried out to support pre-launch SMAP algorithm development activities, with special focus on agricultural regions. The campaign lasted over 40 days from 7^{th} June – 19^{th} July 2012. Throughout the SMAPVEX12, campaign two airborne NASA microwave remote sensing instruments, the Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR) and the Passive Active L- and S-band Sensor (PALS), were flown over the region. L-band radar (1.26 GHz) and L-band radiometer (1.4 GHz) observations were collected on 17 different days during this time period as proxies to the SMAP observatory. Concurrent with airborne flights multiple ground crews collected extensive soil gravimetric and bulk density samples, soil temperature, and roughness measurements across the entire domain. Hand-held sensor probes were also used to sample soil moisture and complex permittivity. Significant effort was also made on vegetation structure and geometry characterization both within the agricultural sites and forested regions. A comprehensive description of SMAPVEX12 activates is documented in [14]. Temporary soil moisture stations (31) were also installed during SMAPVEX12 and managed by the US Department of Agriculture (USDA). Each station included profile soil moisture measurements at 5 and 10 cm depths. Multiple long-term soil moisture networks managed by Agriculture and Agri-Food Canada (AAFC) also exists within the domain.

C. Texas

1) Site Characteristics: The Texas Soil Moisture Observation Network (TxSON, [15]) covers a 36 km $\times$ 36 km area (centered at 30.3°N, 98.8°W) near Fredericksburg, Texas along the Pedernales River and within the middle reaches of the Colorado River. Located on the Edwards Plateau, the karst terrain is a combination of rugged limestone hills with thin soils and lowland alluvial deposits with elevations ranging from 320 m to 670 m. Soils are generally clay loams on hilltops with loamy materials in valley bottoms. Clay content ranges from as low as 3 % in the low-lying areas up to 50 % in other locations. The landcover in the area is a combination grassland/herbaceous (live oak savanna) and evergreen woodland (ashe juniper) primary used for grazing livestock. The region has a semiarid climate, with hot summers and a generally mild winters. Temperatures range from 35°C in the summer to 2°C during winter and mean annual precipitation is around 700 mm.

2) TxSON Network: The TxSON network consists of 40 stations, distributed throughout the 36 km cell (Figure 4), measuring situ soil moisture, soil temperature, and precipitation. The stations were installed in a nested design to provide mean soil moisture at 3, 9 and 36 km scales in support of the SMAP mission.

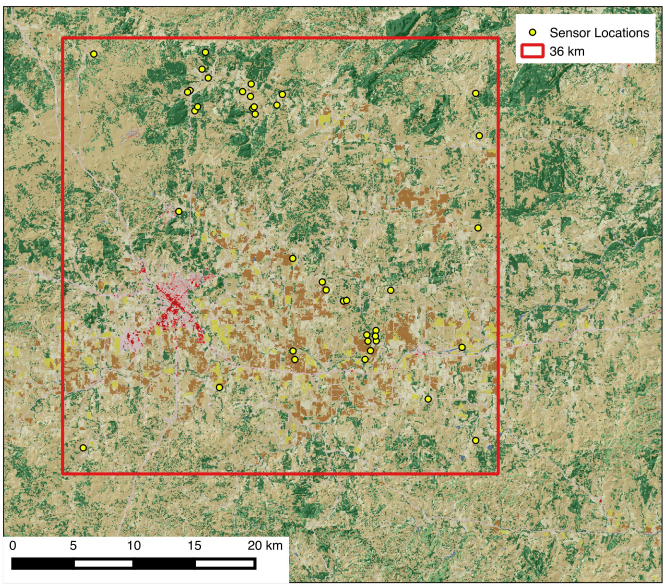

Fig. 4. Location of sensors within the TxSON 36 km grid. Background is NLCD landcover classification.

III. METHOD

The upscaling function relies on building a model expressing soil moisture, for a given date, as a function of a number of geophysical parameters or data layers. The model is trained using measurements from *in situ* sensors within a given time period. This model is then applied to produce gridded estimates of soil moisture for cells (pixels), with an approximate spatial resolution of 100 m. Up scaled soil moisture values are then provided by simple averaging of all high resolution estimates within the domain. An overview of the process is shown in Figure 5. When applying the method to a time series of data, a separate model is built and applied for each time-step.

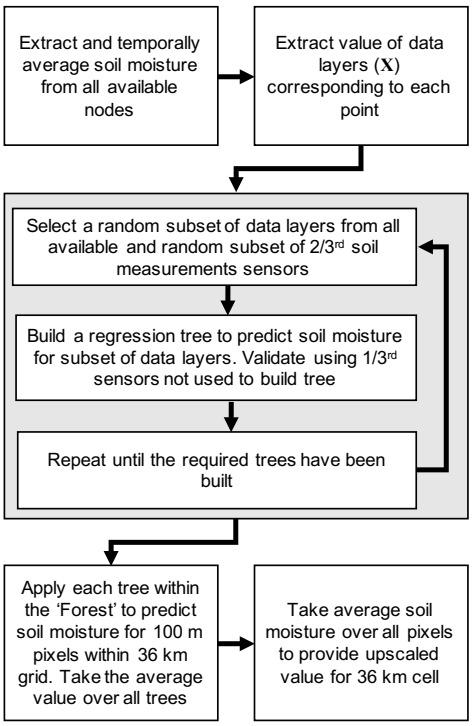

Fig. 5. Random Forests upscaling procedure for a single time period. The grey box shows training of the Random Forests algorithm by building a large number of trees using a subset of data layers and *in situ* sensors. Temporal averaging is optional but can reduce noise in *in situ* data.

Due to the complex relationship between soil moisture and the data layers used, the Random Forests algorithm [11] was used to build the model. Random Forests was selected is able to take continuous and categorical data as input and has been applied in a number of areas to solve classification and non-linear regression problems (e.g., [16]). Random Forests is an extension of the Classification and Regression Trees algorithm (CART; [17]) and utilizes multiple decision trees. Each decision tree is constructed by randomly selecting N samples (where N is the total number of samples), with replacement (bootstrap aggregation; bagging) so as not to alter the characteristics of the pool as selections are made [18]. Approximately 2/3 of the samples are selected to form each tree; the remaining samples (out-of-bag samples) are used to validate the tree. The combination of out-of-bag results across all trees can be used to evaluate overall regression accuracy on the absence of independent validation data. When applying Random Forests, the average soil moisture predicted over all trees is taken as the value for that cell.

A number of layers were used as input to Random Forests; these were chosen as they are known to drive the spatial and temporal variations of soil moisture. Many of these layers are similar to those used to parameterize distributed land surface models (e.g., [9]). The following data layers providing static (i.e., invariant over the period of the study) information:

- Land Cover – *best available for site*
- Elevation – *highest available resolution for site*
- Slope – *derived from the DEM*
- Aspect – *derived from the DEM*
- Flow accumulation – *derived from the DEM*
- Soil clay fraction – *from a regional database*

For both the SoilSCAPE site in California and the TxSON site in Texas, the 1/3 arc second ($\sim$10 m) resolution National Elevation Dataset (NED; [19], [20]) was used to provide elevation information and derived products. For the SMAPVEX12 site the elevation data from the Shuttle Radar Topographic Mission (SRTM) were used at 3 arcsec resolution [21]. SRTM was used for SMAPVEX since SRTM provides the elevation data used in the SMAP soil moisture retrieval algorithm for latitudes in the range 60°N to 56°S [22], and we sought to maintain consistency with SMAP where possible without impacting performance.

To provide information on landcover, the National Land Cover Dataset (NLCD; [23]), derived from Landsat data at a resolution of 30 m, was used for both the SoilSCAPE site and the TxSON site. As the NLCD is only available within the United States, the MODIS landcover product (MCD12Q1; [24]), with a spatial resolution of approximately 500 m, was used for the SMAPVEX12 site.

For both SoilSCAPE and TxSON, soil clay fraction was taken from the USDA's SSURGO soils database ([25]). For SMAPVEX12, soil clay fraction was taken from the Canadian Soil Information Service's National Soil DataBase (CanSIS

NSDB, [26]).

The landcover, topography, and soil clay fraction datasets were used to guide sensor placement at each site within the SoilSCAPE network to ensure the maximum possible number of combinations were sampled.

In addition to the static layers, dynamic layers were used, chosen based on the date of the *in situ* estimates, to provide information on or be related to the temporal dynamics of soil moisture. These included:

- Daily precipitation.
- Daily temperature.
- SAR Backscatter – *HH, VV and HV polarisation, closest date scene selected.*

Daily precipitation and mean temperature at 4 km resolution from the PRISM Climate Group [27] were used for both the SoilSCAPE site and the TxSON site. As precipitation is infrequent for the SoilSCAPE site, temperature was also included as a driver of evapotransipiration. For SMAPVEX12, where PRISM data was not available, precipitation data from the European Centre for Medium-Range Weather Forecasts (ECMWF ERA-Interim, [28]) were used; these had a spatial resolution of 16 km. These were chosen as they are also being used as part of the SMAP mission product generation.

Multiple acquisitions of airborne SAR were available for use in the upscaling for the SoilSCAPE site and the SMAPVEX12 site, but not for the TxSON site. For the SoilSCAPE site, P-band data from the AirMOSS mission were available from a total of 12 dates in November 2013, February 2014, September 2014, February 2015 and May 2015 [29]. For each date upscaled results were generated, the AirMOSS acquisition closest to that of the date being modeled was selected. Similarly, for the SMAPVEX12 site, L-band UAVSAR data were available for 10 dates between June and July 2012, with the UAVSAR acquisition timestamped closest to the date being modeled used in the regression. Most of these UAVSAR flights were coincident with SMAPVEX12 field sampling dates, although UAVSAR data collection did not start until the fourth day of field sampling. As airborne SAR data are only available over a small number of sites, all runs were performed including SAR data and then repeated without SAR data to evaluate any effects on accuracy.

To construct each tree within the 'Random Forest', a random chosen subset of all available data layers was used. Using only a subset of available data layers is part of the Random Forests algorithm, and is designed to reduce the correlation between trees. Three layers per tree were chosen as a good compromise between creating strong individual trees and reducing correlation between trees. Random Forests also works best when a large number of trees are used. In this case, the number of trees was chosen to be 300, as trial runs had shown that the number of errors stopped decreasing after $\sim$ 200 trees and we wanted to allow for run-to-run variability. The implementation of Random Forests regression available through the Python library scikit-learn [30] was used.

IV. Results

The Random Forests upscaling was first applied to the SMAPVEX12 site since extensive field measurements and samples were available to compare up scaled soil moisture estimates. The method was then applied to the TxSON site where a large number of spatially distributed sensors allowed for splitting into training and validation datasets. Lastly, to address and assess soil moisture upscaling activates in support of SMAP, the Random Forests upscaling was applied to the SoilSCAPE network in California.

A. SMAPVEX12

The upscaling method was applied to 13 days of *in situ* data from a temporary network between June and July 2012, coincident with extensive field sampling. The importance of each of the layers when UAVSAR data were included, computed by Random Forests and averaged over all days, is shown in Figure 6. As all sensors fell within the same MODIS class (croplands), the landcover had no importance in the classification. The most important class was elevation, next most important were UAVSAR HV and the clay fraction. Elevation and clay fraction are likely to be been ranked so important because the SMAPVEX12 study area comprised two sharply distinct regions: a high elevation/low clay region covering the majority of the study area and a low elevation/high clay region in the southeastern corner of the study area. The latter region seems to have retained more water than the former, and exhibited consistently higher levels of soil moisture. This is also where the majority of the in-situ sensor stations were installed. The least important layer was precipitation, likely because there was only precipitation for eight of the 13 days. When only days with precipitation were considered, precipitation was the second most important layer.

Results from the proposed Random Forests upscaling method were evaluated against the mean of samples collected during the SMAPVEX12 field campaign and compared to six existing methods for upscaling (Table II and Figure 7). Five of the existing methods used only sensor data, multiple linear regression using the same layers as the Random Forests approach was also included. The results indicated that the lowest RMSE and bias (mean difference) relative to the field means were obtained using our proposed method with UAVSAR data (RMSE 0.025 m^3m^{-3}; Table II). Excluding UAVSAR slightly increased the error, but both runs using Random Forests performed better than the other methods tested. It is important to note that the *in situ* stations were installed primarily in the lower right part of the study domain, where there is a high clay fraction and higher soil moisture than the rest of the domain. This inherently biases all of the methods which use only the sensor data. The

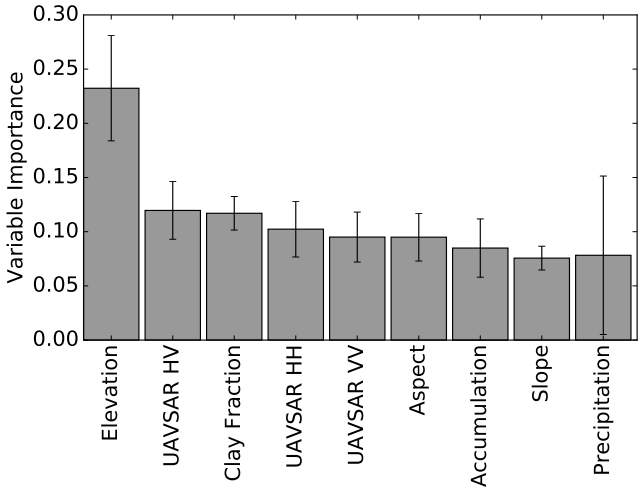

Fig. 6. Average importance of each data layer within Random Forests, over all dates, for the SMAPVEX12 site. Bars indicate $\pm$ 1 standard deviation on the mean.

Random Forests and linear regression methods both included clay fraction as input and were both able to capture the spatial variation of soil moisture with respect to clay fraction. However, in contrast to Random Forests, the values produced when applying the linear regression model were unrealistic, including many negative values which skewed the upscaled value.

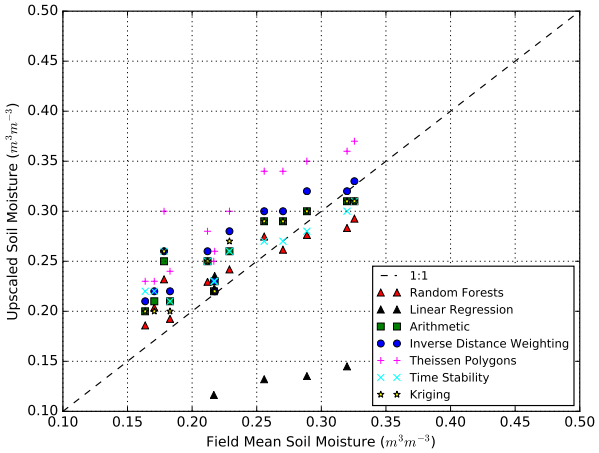

Fig. 7. Comparison of SoilSCAPE Random Forest method and other upscaling methods for SMAPVEX12 site against mean from field measurements.

B. TxSON

For the TxSON site, upscaling was applied to daily measurements of soil moisture covering a one year time series between December 2014 – December 2015. As no field or independent sensor measurements were available 20 of the 40 available TxSON nodes were set aside during each upscaling run to use for validation. It was possible to do this at the TxSON site due to a dense and evenly-spaced distribution of sensors within the grid, allowing an representative value of the mean upscaled soil moisture to be obtained using an arithmetic mean taken over a subset of nodes. The validation nodes were randomly selected for each run, with multiple runs used to reduce variation due to the choice of nodes. To evaluate the sensitivity of the Random Forests upscaling to number of nodes, tests were performed using between 5 – 20 of the remaining non-validation nodes. For each test an arithmetic mean was taken over all training nodes to compare this method of upscaling with the results from Random Forests.

The results are presented in Table III. The lowest RMSE (0.027 m^3m^{-3}) was obtained when all 20 nodes were used with an arithmetic mean, performing slightly better than the Random Forests upscaling with the same number of nodes (RMSE

TABLE II
COMPARISON OF PROPOSED UPSCALING METHOD AND EXISTING TECHNIQUES TO FIELD SAMPLE MEAN (m^3m^{-3}) FOR THE SMAPVEX12 SITE.

Method	RMSE	Bias	uRMSE
Random Forests	0.025	0.007	0.024
Random Forests (without UAVSAR)	0.029	0.018	0.023
Linear Regression	0.142	-0.129	0.060
Linear Regression (without UAVSAR)	0.107	-0.095	0.049
Arithmetic	0.032	0.023	0.022
Inverse Distance Weighting	0.041	0.034	0.023
Theissen Polygons	0.067	0.063	0.022
Time Stability	0.036	0.021	0.029
Kriging	0.033	0.022	0.025

TABLE III
COMPARISON OF UPSCALING RESULTS WITH A VARYING NUMBER OF SENSORS FOR TXSON SITE COMPARED TO AVERAGE OF 20 SENSORS HELD BACK FOR VALIDATION (m^3m^{-3}). AVERAGE OVER A TIME SERIES AND USING 5 RANDOM PERMUTATIONS OF TRAINING / TESTING SENSORS. RESULTS FROM RANDOM FORESTS UPSCALING ARE COMPARED WITH A SIMPLE ARITHMETIC MEAN OF TRAINING PIXELS.

No. Sensors	Method	RMSE	Bias	uRMSE
5	Arithmetic	0.038	0.027	0.027
	Random Forests	0.033	0.016	0.028
10	Arithmetic	0.03	-0.002	0.03
	Random Forests	0.03	0.005	0.03
15	Arithmetic	0.036	-0.007	0.035
	Random Forests	0.034	-0.004	0.034
20	Arithmetic	0.027	0.006	0.027
	Random Forests	0.029	0.013	0.026

0.029 m^3m^{-3}). When only 5 and 15 nodes were used Random Forests performed better than an arithmetic mean and the same when 10 nodes were used (RMSE 0.03 m^3m^{-3}).

C. SoilSCAPE

The upscaling was applied to data over a two year period between 2014 – 2016 [31]. Figure 8 shows average over all *in situ* soil moisture sensors and upscaled soil moisture for the entire time series. The upscaled estimates of soil moisture were similar to the average over all sensors, differing by 0 – 0.08 m^3m^{-3} and having a lower standard deviation. The number of sensors used for each date ranged from 15 to 49 with an average of 31. Over all runs the average out-of-bag error was 0.046 m^3m^{-3}. The importance of each of the data layers is shown in Figure 9. The most important layer was accumulation, followed by slope then temperature (a driver of evapotranspiration). When considered over all dates, precipitation had the lowest importance as it was only non-zero for 21 % of dates used. When considering only dates where precipitation was non-zero, it was the fourth most important layer after slope, accumulation and temperature. For both instances, AirMOSS data were more important than landcover and clay fraction, with HV having the highest importance of all polarizations when all dates were included and HH when only dates with precipitation were included.

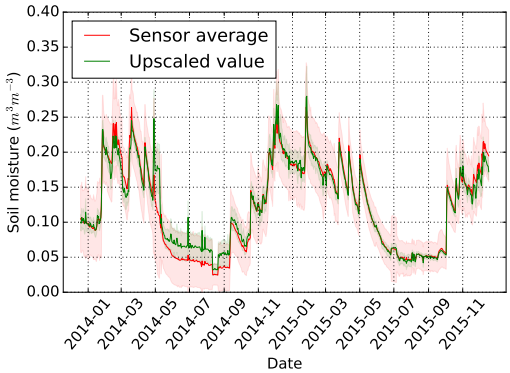

Fig. 8. Average soil moisture, taken over all available sensors (red) and all 100 m sub-cells within the 36 × 36 km sampling grid (green) for the SoilSCAPE site. Shaded areas show ± 1 standard deviation around the mean.

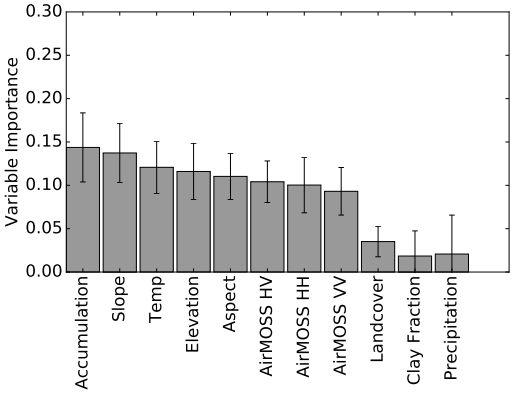

Fig. 9. Average importance of each data layer within Random Forests, over all dates, for the SoilSCAPE site. Bars indicate $\pm$ 1 standard deviation on the mean.

In the absence of extensive field samples to calculate the true upscaled soil moisture the separate network of sensors installed at Tonzi ranch by OSU were used to compare predicted soil moisture for the 100 m cells containing the sensors. The comparison, covering the period January 2014 – December 2015, is presented in Figure 10. The results show an RMSE of 0.041 m^3m^{-3} when all three OSU sensors are considered. When AirMOSS data were not included the RMSE remained at 0.041 m^3m^{-3}.

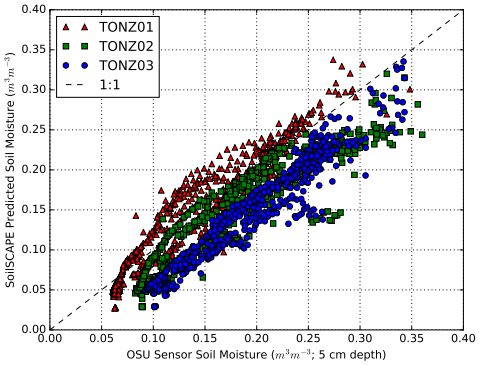

Fig. 10. Comparison of soil moisture for 100 m cells with co-located soil moisture sensors from a separate study, not included in the training data, for the SoilSCAPE site. RMSE = 0.041 m^3m^{-3}, bias = -0.028 m^3m^{-3}, ubRMSE= 0.03 m^3m^{-3}

V. Discussion

A. Approach to upscaling

To address validation of spaceborne soil moisture products, especially the NASA SMAP mission, a data driven method has been presented which upscales surface soil moisture measurement using *in situ* sensors as well as relevant geophysical data layers, e.g., precipitation, soil texture, etc. The technique makes use of the Random Forests regression and classification algorithm to form data-driven regressions trees linking surface soil moisture to a number of different, but interrelated, data layers.

Independent validation of the method was performed by application of Random Forest of data gathered during the SMAPVEX12 campaign and as well as analysis on *in situ* sensors in the TxSON site. Analysis of Random Forests for SMAPVEX12 show that the method outperforms traditional upscaling techniques using only sensor data: Arithmetic mean, Inverse Distance Weighting, Theissen Polygons, Time Stability and Kriging. It also performed better than multiple linear regression, using the same input layers. With respect to intensive ground truth sample, the Random Forests method had the lowest RMSE (0.025 m^3m^{-3}) and smallest bias (0.008 m^3m^{-3}), when compared to these other techniques. The poor performance of multiple linear regression is likely due to lack of a strong linear relationship between soil moisture and the data layers used as input for the SMAPVEX12 site.

For the TxSON site, the number and distribution of nodes were sufficient to capture the variation present so a simple average was sufficient to provide upscaled soil moisture. For this case there was no benefit in the Random Forests approach. Where reducing the number of nodes used for upscaling the performance of Random Forests improved, compared to a simple average.

For the SoilSCAPE network in California, high resolution Random Forest predictions strongly track an independent network of *in situ* sensors. Mean upscaled soil moisture is similar to mean sensors *in situ* values over the two-year study period, but with less spatial variation. Due to limited precipitation throughout the study, this data layer was minimally ranked, however, when considering precipitation-only days, albeit still small, it was the fourth highest ranked parameter.

Utilization and application of interrelated geophysical data layers, as drivers of spatial and temporal soil moisture dynamics, within the regression scheme proves advantageous and overcomes sensors-only methods shortcomings. The latter methods performance suffers significantly with inadequate number of sensors and poor placement. Information not captured by the sensors can be inferred from different data layers. Unlike distributed land surface models the importance of each data layer is driven by the available *in situ* measurements rather than being pre-determined and can site or even date specific. The importance of each layer can be analyzed after each run.

The upscaling method is computationally efficient and can automatically be applied to a long time series of data. For example, running the upscaling for the 730 dates data from the SoilSCAPE network were available took 90 minutes on a high end workstation (Intel i7). The upscaling can easily scale to higher resolutions and larger areas, if suitable input layers and *in situ* data are available without significantly increasing run time as applying Random Forests to an image is quick and can be applied on a per-pixel basis requiring very little memory. In contrast, resolution and domain size has a large impact on run time for distributed land surface models

B. Role of SAR data

As airborne SAR data were available over both the SMAPVEX12 (UAVSAR; L-band) and SoilSCPAE (AirMOSS; P-band) sites they were included as data layers in the upscaling. These datasets provided high resolution information related to landcover and moisture content of both soil and vegetation. While the landcover maps were able to provide discrete classes the SAR data could provide more granular information with short dense blocks of trees exhibiting a different response to tall sparse blocks. Analysis of Random Forest outputs demonstrated both P-band SAR from AirMOSS and L-band SAR from UAVSAR could play an important role in the upscaling function, based on the importance of these variables in the classification. Further tests without the inclusion of SAR data revealed little difference in overall accuracy compared to when the SAR data were included, indeed for the SMAPVEX12 site, including UAVSAR had a slightly negative effect on accuracy. It is likely that for the two examples presented, the information provided from the SAR data was already better captured in other layers (e.g., landcover classification). For other sites including SAR data may have a greater impact. One of the advantages of the Random Forests approach proposed is that it can take airborne or spaceborne data from a range of sources as input rather than requiring derived products.

C. Network and Sampling Design

One of the aims of the SoilSCAPE project was to improve the design of *in situ* soil moisture networks through the use of the latest technology and algorithms, including investigating alternative approaches to upscaling, in order to provide validation data at a range of scales. As part of this, landcover, topography and soil properties layers were used to guide sensor placement rather than only considering only spatial distribution. This design meant the network was poorly suited for traditional upscaling approaches but well suited for the method presented here which used the same data layers as those used when planning the network.

In contrast to the SoilSCAPE network, the SMAPVEX12 and TxSON networks were designed around standard upscaling approaches with sensors placed to maximize spatial coverage. The lower errors when applying our proposed Random Forests method to upscaling compared to standard upscaling approaches applied to SMAPVEX12 demonstrates that even for networks with a good nominal spatial distribution our method performs better than existing methods, provided that there are *in situ* observations over representative strata of the domain properties (soil texture, landcover, topography, etc.). Conversely, even if there are a large number of sensors used, if they not placed with landscape representation, they cannot be used to calculate an accurate mean value.

The SoilSCAPE network design, of deploying clusters of sensors centered around a main controller node significantly reduces the cost of network installation and maintenance as more expensive components such as data loggers and modems are shared between a large number of nodes rather than being required for each node. The design also allows the network to operate at a range of spatial scales.

D. Ongoing Validation of Spaceborne Products

One of the advantages of long-term soil moisture networks is that they can be used to provide validation data over the duration of a spaceborne mission, compared to field campaigns which are only able to provide data for a short period of time (typically weeks). The upscaling approach presented here has demonstrated the ability to be applied to such long-term *in situ* networks. For the SMAP mission another consideration is that the upscaled results can be provided to the calibration and validation team in a timely manner and that they can be incorporated within the existing science framework. The upscaling

method presented here has exhibited stable performance under the typical annual range of moisture conditions experienced within the SoilSCAPE network's 36 km $\times$ 36 km upscaling cell as well as an ability to cope with missing sensors. This has been demonstrated by applying our method to two years of data for the SoilSCAPE site, during which soil moisture ranged from $<$0.05 – 0.3 m^3m^{-3}. The number of *in situ* measurements available varied from 15 - 49, with the variation due to sensors failing and being repaired over the course of the run and additional sensors being installed.

E. Future work

In addition to the use of two independent field sites (SMAPVEX12 and TXSoN) for validation of the presented Random Forests approach, validating the 36 km upscaled results for the SoilSCAPE network in California using extensive field sampling could provide greater confidence in application of the method. However, given the size of the area, which covers a large fraction of privately owned land planning and executing such a campaign is not practical or feasible. Using a smaller number of samples and using these to validate the 100 m scale predictions, as was done with the sensors from OSU, is feasible. Combining sampling with site visits will allow a larger validation dataset to be gradually built up.

The method is not specific to a single site and only needs appropriate data layers as input. Applying to different sites, where field campaigns have been carried out, will provide another means of validating the method. To facilitate this the code will be made available from https://github.com/mixil, where our group has made a variety of other codes available.

VI. Conclusions

A method has been presented for upscaling soil moisture measurements from *in situ* sensors over large areas, such that the resulting upscaled values can be used to validate products from spaceborne instruments and in particular from SMAP. The approach makes use of Random Forests to form a data-driven model relating soil moisture to a number of data layers providing information on soil type, landcover and topographic conditions. The method has been validated against field samples and outperformed other common methods used for upscaling.

Acknowledgments

The SoilSCAPE project was funded through the NASA Earth Science Technology Office (ESTO) Advanced Information System Technology (AIST) program. Deployment of the SoilSCAPE network was possible thanks to the permission and assistance of a number of land holders including the Tonzi family, the US Army Corps of Engineers at New Hogan, the Bureau of Land Management, and the staff at Terra d'Oro vineyards. The SoilSCAPE sensors were installed and are maintained by a number of staff from the Microwave Systems Sensors and Imaging Lab (MiXIL) at USC. The SMAPVEX12 experiment was supported in part with grants from the Canadian Space Agency.

References

[1] D. Entekhabi, E. Njoku, P. O'Neill, K. Kellogg, W. Crow, W. Edelstein, J. Entin, S. Goodman, T. Jackson, and J. Johnson, "The Soil Moisture Active Passive (SMAP) Mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[2] Y. H. Kerr, P. Waldteufel, J. P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M. J. Escorihuela, J. Font, N. Reul, C. Gruhier, S. E. Juglea, M. R. Drinkwater, A. Hahne, M. Martin-Neira, and S. Mecklenburg, "The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 666–687, May 2010.

[3] T. Jackson, M. Cosh, R. Bindlish, P. Starks, D. Bosch, M. Seyfried, D. Goodrich, M. Moran, and J. Du, "Validation of Advanced Microwave Scanning Radiometer Soil Moisture Products," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 12, pp. 4256–4272, 2010.

[4] W. T. Crow, A. A. Berg, M. H. Cosh, A. Loew, B. P. Mohanty, R. Panciera, P. de Rosnay, D. Ryu, and J. P. Walker, "Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products," *Reviews of Geophysics*, vol. 50, no. 2, p. RG2002, Apr. 2012.

[5] K. Y. Vinnikov, A. Robock, S. Qiu, and J. K. Entin, "Optimal design of surface networks for observation of soil moisture," *Journal of Geophysical Research*, vol. 104, no. D16, pp. 19 743–19 749, Aug. 1999.

[6] G. Vachaud, A. Passerat de Silans, P. Balabanis, and M. Vauclin, "Temporal stability of spatially measured soil water probability density function," *Soil Science Society of America Journal*, vol. 49, no. 4, pp. 822–828, 1985.

[7] J. S. Famiglietti and E. F. Wood, "Multiscale modeling of spatially variable water and energy balance processes," *Water Resources Research*, vol. 30, no. 11, pp. 3061–3078, Nov. 1994.

[8] C. D. Peters-Lidard, M. S. Zion, and E. F. Wood, "A soil-vegetation-atmosphere transfer scheme for modeling spatially variable water and energy balance processes," *Journal of Geophysical Research*, vol. 102, no. D4, pp. 4303–4324, Feb. 1997.

[9] N. W. Chaney, J. K. Roundy, J. E. Herrera-Estrada, and E. F. Wood, "High-resolution modeling of the spatial heterogeneity of soil moisture: Applications in network design," *Water Resources Research*, vol. 51, no. 1, pp. 619–638, Jan. 2015.

[10] W. T. Crow, D. Ryu, and J. S. Famiglietti, "Upscaling of field-scale soil moisture measurements using distributed land surface modeling," *Advances in Water Resources*, vol. 28, no. 1, pp. 1–14, Jan. 2005.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] M. Moghaddam, D. Entekhabi, Y. Goykhman, K. Li, M. Liu, A. Mahajan, A. Nayyar, D. Shuman, and D. Teneketzis, "A Wireless Soil Moisture Smart Sensor Web Using Physics-Based Optimal Control: Concept and Initial Demonstrations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 3, no. 4, pp. 522–535, 2010.

[13] A. R. Silva, M. Moghaddam, and M. Liu, "Ripple-2: a non-collaborative, asynchronous, and open architecture for highly-scalable and low duty-cycle WSNs," *ACM Mobile Computing and Communications Review*, vol. 17, no. 1, pp. 55–60, Jan. 2013.

[14] H. McNairn, T. J. Jackson, G. Wiseman, S. Belair, A. Berg, P. Bullock, A. Colliander, M. H. Cosh, S.-b. Kim, R. Magagi, M. Moghaddam, E. G. Njoku, J. R. Adams, S. Homayouni, E. Ojo, T. Rowlandson, J. Shang, K. Goita, and M. Hosseini, "The Soil Moisture Active Passive Validation Experiment 2012 (SMAPVEX12): Prelaunch Calibration and Validation of the SMAP Soil Moisture Algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2784–2801, Jan. 2015.

[15] Caldwell, T and Young, M and Scanlon, B, "Texas Soil Observation Network (TxSON)," 2016. [Online]. Available: http://www.beg.utexas.edu/txson/
[16] J. Whitcomb, M. Moghaddam, K. McDonald, J. Kellndorfer, and E. Podest, "Mapping vegetated wetlands of Alaska using L-band radar satellite imagery," *Canadian Journal of Remote Sensing*, vol. 35, no. 1, pp. 54–72, 2009.
[17] L. Breiman, J. Friedman, O. R.A., and S. C.J., *Classification and Regression Trees*. New York: Chapman and Hall, 1993.
[18] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[19] D. Gesch, M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler, "The National Elevation Dataset," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 1, pp. 5–11, 2002.
[20] D. Gesch, *The National Elevation Dataset*, ser. Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd Edition. American Society for Photogrammetry and Remote Sensing, 2007, pp. 99–118.
[21] NASA Land Processes Distributed Active Archive Center (LP DAAC), "SRTMGL3: NASA Shuttle Radar Topography Mission Global 3 arc second V003." [Online]. Available: https://lpdaac.usgs.gov/dataset_discovery/measures/measures_products_table/srtmgl3_v003
[22] E. Podest and W. Crow, "Soil Moisture Active Passive Ancillary Data Report, Digital Elevation Model, Preliminary, v.1, JPL D-53056," NASA Jet Propulsion Laboratory, California Institute of Technology, Tech. Rep., 2013.
[23] C. Homer, J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, and J. Wickham, "Completion of the 2001 National Land Cover Database for the conterminous United States," *Photogrammetric engineering and remote sensing*, vol. 73, no. 4, pp. 337–341, Apr. 2007.
[24] NASA Land Processes Distributed Active Archive Center (LP DAAC), "Land Cover Type Yearly L3 Global 500 m SIN Grid." [Online]. Available: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd12q1
[25] USDA Natural Resources Conservation Service, "Description of SSURGO Database." [Online]. Available: http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils
[26] Agriculture and A.-F. Canada, "The National Soil DataBase (NSDB)." [Online]. Available: http://sis.agr.gc.ca/cansis/nsdb/index.html
[27] PRISM Climate Group, Oregon State University, "PRISM Climate Data." [Online]. Available: http://prism.oregonstate.edu
[28] European Centre for Medium-Range Weather Forecasts (ECMWF), "Reanalysis Datasets, ERA-Interim." [Online]. Available: http://ecmwf.int/en/research/climate-reanalysis/era-interim
[29] AirMOSS Science Team, "AirMOSS: L1 S-0 Polarimetric Data from AirMOSS P-band SAR, Tonzi Ranch, 2012-2015," 2016. [Online]. Available: https://doi.org/10.3334/ORNLDAAC/1414
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[31] M. Moghaddam, A. Silva, D. Clewley, R. Akbar, S. Hussaini, J. Whitcomb, R. Devarakonda, R. Shrestha, R. Cook, G. Prakash, S. Santhana Vannan, and A. Boyer, "Soil Moisture Profiles and Temperature Data from SoilSCAPE Sites, USA," 2016. [Online]. Available: https://doi.org/10.3334/ORNLDAAC/1339