A Rule Transference Algorithm for Obtaining High-Resolution Soil Moisture Surface in Arid and Semi-Arid Regions

Michael G. Lewis, Andmorgan Fisher, Clint Smith, John J. Qu, Paul Houser

Yearbook of the Association of Pacific Coast Geographers, Volume 82, 2020, pp. 92-114 (Article)

Published by University of Hawai'i Press
DOI: *https://doi.org/10.1353/pcg.2020.0005*

A Rule Transference Algorithm for Obtaining High-Resolution Soil Moisture Surface in Arid and Semi-Arid Regions

Michael G. Lewis, Andmorgan Fisher, Clint Smith, John J. Qu, Paul Houser
George Mason University

ABSTRACT

Soil moisture is vital to understanding many natural systems such as hydrology, climate and weather, erosion, and biology. Current remote sensing provides soil moisture data with a resolution on the scale of tens of kilometers, due to the current constraints of microwave antennae technology. In this study, we present a machine-learning technique based on rule transference that allows us to use a low-resolution but high-accuracy product, obtained through multiple proxies, to produce a high-resolution model of Earth's soil moisture. The low-resolution, high-accuracy microwave product is utilized as a dependent variable in rule-building only. This algorithm is simple, utilizes public data, and overcomes many local issues inherent in other techniques, such as topographic, biographic, temporal, and climatic variations. The final result demonstrates close parity with high-resolution airborne L-band radiometric data.

Keywords: *Downscaling, soil moisture, random forests*

Introduction

Soil moisture is a measure of the hydrological component within a finite amount of soil. The variability of soil moisture is highly dependent on the soil properties (Cosby et al. 1984), the local geologic conditions (Weizu and Freer 1995), vegetation density and draw (Denmead and Shaw 1962), and antecedent conditions. Soil moisture is also very important for understanding the physical conditions of the Earth for many systems. Agriculture relies on soil moisture for plant vigor and growth, providing root-zone moisture and a direct relationship to CO_2 respiration through soil microbial activity (Orchard and Cook 1983). The moisture itself acts as a primary nutrient for growing crop and plant life, carrying with it the organic and inorganic

trace nutrients necessary for plant growth. Weather and climate models are strongly coupled to land–atmosphere interactions (Koster et al. 2004). Flood control is highly dependent on antecedent soil moisture conditions (De Michele and Salvadori 2002). Slope failure and mass movement are linked to soil plasticity induced by elevated moisture content, creating hazards for local communities and travel (Crosta 1998). Soil moisture, in short, affects our communities in countless ways. In order to quantify the impact, we need to be able to measure soil moisture in a reliable manner.

The Soil Moisture Active Passive (SMAP) mission, launched in 2015, attempted to develop a higher-resolution product, one with a resolution better than 10 km. This was accomplished through the integration of the highly accurate but low-resolution passive radiometer with the higher-resolution active microwave, which is sensitive to vegetation and surface roughness effects (Wu et al. 2015). Unfortunately, SMAP suffered an irrecoverable system failure in the active microwave portion of the sensor, leading to the passive radiometer being the only usable sensor part of the instrument. SMAP has a very coarse pixel size that is problematic at the local scale, giving a functional received product with a thirty-six kilometer resolution. Although some algorithms linking the SMAP sensor to ground-condition soil moisture have been proposed, the resolution issue remains.

Downscaling this data to field scale is useful for local, actionable intelligence on soil moisture, and can be accommodated by several methods. Peng et al. (2017) provide a comprehensive, in-depth review of these methods. We will briefly describe what governed our choice in downscaling in terms of the types of soil moisture downscaling methods lists.

Our initial concern was the end user. Our focus is on the underserved communities that may not have resources to utilize advanced methods of downscaling to support their civil works and agricultural communities.

Downscaling methodologies can fit broadly into three classes: satellite-based methods, geoinformation methods, and model-based methods. Geoinformation methods tend to be highly localized and, while they have great potential for amending the other two downscaling types, cannot as of yet be relied on for downscaling soil moisture at regional scale.

Among satellite-based fusion methods, Active-Passive fusion was the initial plan for SMAP, but given the failure of the active portion, it is not an option, although SMAP passive has been successfully fused with other active systems. The optical/thermal and microwave fusion methods have the same general input as our model, relying on the vegetation properties and

surface temperature. However, these tend to be a polynomial fitting function and therefore are not as capable in applying multiple sub-routines over a region. Instead, we felt a model that is able to differentiate regime zones and establish models best suited to that zone was desirable. Over such a large region within a single Landsat scene, it seemed an algorithm that could in effect have multiple Soil-Vegetation-Atmosphere triangle (SVAT, discussed below) clusters instead of one could provide a more reliable response.

Model-based methods are separated between land-surface integrated models and statistical models. We wanted to avoid land-surface models due to our envisioned end user. Furthermore, the statistical approach also increases in complexity, especially when considering the fractal nature of soil moisture. While we believe a statistical approach would be a fit in amending our method's shortcomings in addressing increasing variability at larger scale, we will demonstrate our method is effective at obtaining soil moisture at field scale while being relatively simple.

Here, we propose a method, Rule Transference Algorithm (RTA), which improves the resolution of soil-moisture imagery products, using an inferred-learning soil-moisture algorithm. The intent is to provide a simple, repeatable methodology to downscale low-resolution soil-moisture data to field scale, in a manner that can be performed with relatively low skill on public data, using open-source tools. This methodology shares many similarities with optical/thermal and microwave fusion methods of soil-moisture downscaling, as well as polynomial fitting approaches (Peng et al. 2016). However, this model is adaptable to heterogeneous conditions within a single scene, unlike traditional polynomial fitting models, and has fixed variables, unlike many learning models. While we are using SMAP, there is no reason other sensors such as SMOS, ASCAT, or AMSR-E could not supplant SMAP in the algorithm, with minor adjustments. The inferred learning approach appears to be non-conservative, meaning that the aggregated high-resolution soil moisture is not necessarily equal to the coarse soil-moisture resolution. Furthermore, the conceptual validity of the method is drawn from the Soil-Vegetation-Atmosphere Transfer (SVAT) triangle model. While temperature and the Normalized Difference Vegetation Index (NDVI) are both accounted for in the variables used by this method, the albedo effect, as explained by Zhan et al. (2002) and Chauchan et al. (2009), is replaced by two direct-sensing soil-moisture indices, in addition to a new variable that appears to describe well the lack of moisture. The polynomial methodology is not incorporated into the inductive random forest algorithm, which

addresses the anomalous features or transitions in climatic, topographic, or biotic schemes, because a fitted solution would be biased against such shifts. Nevertheless, our method enjoys the same advantages of polynomial fitting methods, in that it does not require *in-situ* measurements (Peng et al. 2016) but does require clear atmospheric conditions and has not been tested extensively in dense vegetation regions. This methodology, at present, is limited in its application to arid and semi-arid regions, due to the need for soil temperature estimation.

Finally, this algorithm is similar to downscaling but is not actually a standard method of downscaling. Instead, it performs rule transference across spatial resolutions, and is best described as a hybrid model between statistical and optical/thermal fusion. Typical downscaling is a granularity-increasing exercise, whereas our approach ignores and discards the original SMAP data after rule learning is complete, and the SMAP data is not assimilated into the final product. However, for convenience in this text, we continue to refer to our method as a downscaling product. This method of downscaling is linked to the feature attribute selection, described in the following section. The attribute selection was done naively, without reference to current methods of downscaling, and its similarity to other methods, such as the triangle method described below, arose from evaluation of the highest-performing attribute selections.

Materials and Methods

This study was an exploration in the effectiveness of a rule transference algorithm is downscaling soil moisture. However, organizationally it is a smaller study of feature selection within the larger downscaling study. While the feature selection is of importance to understand the downscaling method, it is not being directly evaluated outside the wholistic nature of the algorithm. Therefore, the feature selection testing is outside the scope of this report but is included in the following section, to familiarize the reader with the underlying methodology and importance of the variables chosen. Following a discussion of feature selection, we will discuss the rule-transference algorithm. This method of statistical downscaling using a random forest was selected in order to accommodate various biome and topographic regimes within the large sampling space. Finally, this study utilizes data from the SMAPEx-5 campaign performed in the area of Yanco, Australia (Ye et al. 2017), for validation. This field study utilized several remote-sensing instruments and in-situ sensors, but it utilizes the airborne Polarimetric L-band

Multibeam Radiometer (PLMR) and intensive near-surface measurement of near-surface soil moisture, using capacitance sensors over a regular sampling pattern of 250m spacing across three 3km x 3km focus fields.

Feature Attribute Selection

Feature attribute selection is the act of pairing down variables from a full set of variables to those most able to increase performance and which contain the least amount of redundancy. Models using many variables can be unnecessarily complicated, but should generally be kept as simple as possible while using as few initialization parameters as possible to attain the desired outcome, unless there is an explicit reason for doing otherwise. Reducing initial variables helps minimize generalization errors (Cannon & Whitfield 2002). Furthermore, for computational efficiency, the training times on lower-dimension models are significantly faster and many times more accurate than higher-dimension algorithms, a situation known as the curse of dimensionality.

In this study, initial variable exploration used field-collected point time-domain reflectometry (TDR), point COsmic-ray Soil Moisture Observing System (COSMOS) data, and field-collected volumetric soil-moisture data from multiple sites, including Nevada, New South Wales, and Arizona, in the spring, summer, and fall of the years 2014–15. These collections were made for multiple projects run by the Army Corps of Engineers but were reused in this project for feature selection. From the initial collection of soil moisture point readings, we used those that also coincided with Landsat imagery within a day of the reading. This created a training set with a varied biome and soil type spanning 12,555 Landsat pixels. This distribution of training sites would diminish seasonal and soil specifics to elevate more-universal relationships in the testing phase. Landsat 8 level-1 digital number (DN) values were extracted for each pixel over all bands of a sample set, to test the relationships between the variables and pair them. As we are interested only in evaluating ratios, the DN do not need to be radiometrically corrected, although we did perform dark-object subtraction. This was also a conscious choice, with the goal of achieving an easily copied method using open-source data and processing so that surface soil-moisture maps could be generated by anyone with a computer. Since these are arid to semi-arid regions, there will be inherently less atmospheric intrusion in the signal as well. In 2001, Song et al. tested a variety of atmospheric correction techniques using Landsat Thematic Mapper data and concluded that atmospheric correction is only

necessary when comparing across images with differing spatiotemporal characteristics. While indices such as NDVI may shift upward by perhaps twenty percent without correction (Nigam, et al. 2012), if such shifts are accommodated in the training then they are accommodated in the testing phase as well. Variables are generated through the iterative ratio of each band with every other band. In the following iteration, the ratio is calculated with all previous bands, continuing through four iterations. Working exclusively with ratios largely nullifies atmospheric effects because each band value is affected relatively, and all values in a single set of scenes take over a similar region on the same day.

This created a very large dataset of Landsat variables to test against. Beginning with 11 Landsat bands, the ratios of each band presented a rapid expansion of variables according to the triangle number of starting variables, so that the initial 11 Landsat band set generated 66 variables in the first iteration of ratios and 2,211 in the second one. We should note that the panchromatic resolution, band 8, was aggregated to 30 m, whereas bands 10 and 11 were assigned the value at the spatial pixel centerpoint of bands 1–7; this is also how they are dealt with in the final processing. Due to the massive number of variables and a need for multiple evaluations, Jonathon Nunez, a master's student from the University of Puerto Rico, assisted by running a parallel evaluation to confirm or deny variable effectiveness. Because such a large dataset would be computationally difficult, the initial pairing was conducted through correlation feature selection in MS Excel. All the attributes were tested against field-sampled soil-moisture surfaces, collected using TDR or FDR (Frequency Domain Reflectometry), with Pearson's correlation coefficient. To conform with the computing requirements and the column limits of the spreadsheet program, only the variables showing the highest positive or negative correlations to the measured moisture were brought forward to the next two iterations. Those attributes scoring above the absolute value of 0.65, a natural cut-off in the data below which most variables fell, were brought forward for additional testing using a wrapper method with a random forest classifier using Weka, an open-source platform developed by the University of Waikato in New Zealand. Although lower-correlation attributes were occasionally brought forward if they showed a higher ranking in information gain, a secondary test was used for marginal variables. The 20 highest performing variables were compared in terms of correlation-based feature selection, information gain, redundancy, principal component analysis, and learner-based feature selection using the random

forest algorithm. The eight selected final variables represented the highest-performing variables across the selection criteria. Interestingly, Variable 16 was included in these due to its consistency and information gain, even though it rarely scored in the top three-to-five attributes. In the final analysis, the algorithm performed better with this variable than without it. We also included many initial pre-existing VIS/IR ratio variables in addition to the NDVI, from the full list of vegetation-, soil-, and hydrology-related indices of the Index DataBase (Verena 2012). These variables were excluded from the initial correlation feature selection but were included in the wrapper selection phase. Test of prediction skill, multicollinearity, and leave-one-out variable testing provided a base set of eight variables.

Evaluation of similarity on the remaining variables seemed to suggest they fit well into three clusters, or variable neighborhoods. Neighborhood 1 seemed to determine moisture from the temperature variations in the soil, while Neighborhood 2 attempted to directly detect water. Neighborhood 3 uses a vegetation proxy to identify moisture and, therefore, is associated with plant vigor. We quickly noticed that these neighborhoods of variables are also representative proxies for the key components of the SVAT "triangle" model. The triangle method is independent of readings from a site outside those that can be obtained from the red, NIR, and thermal bandwidths. However, it requires multiple data points to establish the triangle of the method. The triangle is a scatterplot of NDVI against thermal surface data. The assumption is that, given a sufficiently large sampling, the pixels exhibit a sufficiently large variance to plot the full range of possible scenarios from wet to dry and from bare to a fully vegetated Earth (Carlson 2007). This scatterplot forms a triangle that delineates a shape "warm edge" where surface temperature is higher than in other similarly vegetated pixels, indicating a lack of water (Figure 1). The assumption made is that vegetation temperature does not vary spatially and that surface temperature alone accounts for the variation seen, at least within the detection error margin (Carlson 2007). This is somewhat in disagreement with the work done earlier by Thomas Jackson of the USDA on the moisture availability of foliage and temperature (Jackson et al. 1981). However, the leaf area-sourced radiation is minor in comparison to the ground temperature effect. The "cool edge," somewhat less defined, represents pixels with adequate or excess moisture for cooling. The lower surface trails off into a low-density point dispersion, which can be accounted for by surficial anomalies, such as standing bodies of water, and clouds, and should be removed to create a cluster in a tight arrangement.

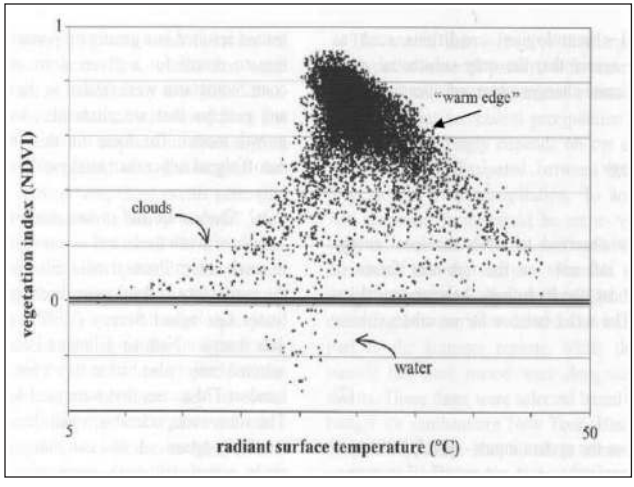

Figure 1.—Triangle method scatterplot (Carlson 2007).

This model requires some degree of subjectivity in defining the triangle edges, which poses a problem for cross-site and cross-experiment comparisons. A sufficiently large number of instances is also required, often hundreds to thousands of points, assuming that soil characteristics remain rather uniform overall. Despite the apparent subjectivity of the method, Carlson (2007) found that the error in estimating the evapotranspiration function, and thus the available soil moisture, from such a method is typically ± 0.1–0.2 (Carlson 2007), near the theoretical minimum.

Replacing our created Neighborhood 2 ratio variables with established variables such as the Modified Normalized Difference Water Index (MNDWI) and the Normalized Soil Moisture Index (NMSI) demonstrated better performance in test/train scenarios using our original training dataset. Neighborhood 1 (thermal) was best served by the two new ratio variables, referred to as Variable 2 and Variable 10 in our testing. Neighborhood 3 is well served by the NDVI alone. A third variable, Variable 16, has also shown promise in discriminating soil moisture in vegetated areas, and was brought forward due to repeatedly high performance and high independence from other variables. Notice that we have two variables, each describing the three aspects, or neighborhoods, of soil moisture detection: surface heat, vegetation vigor, and spectral water detection. It is important that both variables have low collinearity with each other, yet describe well the underlying aspect.

The final set consists of six variables. Three of them, NDVI, NMSI, and MNDWI, are predefined variables. Variable 2 is a simple ratio representing the thermal return, whereas the other two, Variable 10 and Variable 16, are unique to this algorithm. The numbers of the variables are simply the names used while testing the final set of twenty variables.

The final variable set (using Landsat 8, level 1 data) is:

Var10: (Pan/TIRS2)/(Cirrus/TIRS1)
Var2: Coastal/TIRS1
Var16: (Red × SWIR2^3)/(NIR × SWIR1^2)
NDVI: (NIR − Red)/(NIR + Red)
NMSI: (SWIR2 − SWIR1)/(SWIR2 + SWIR1)
MNDWI: (Green − SWIR1)/(Green + SWIR1)

The Learning Algorithm

After the variables were chosen, it was necessary to choose the dependent variable, which was not obvious, since precise data on the time and location related to soil moisture is often unavailable. Here, the SMAP values provide a local training set and the independent variable to the machine learning algorithm. Our test site is the SMAPex-5 field campaign in support of the SMAP validation project in the Murumbidgee River Valley, NSW Australia (Ye et al. 2017; Panciera 2013).

The model tree iteratively splits a training set of variables in order to minimize the regression error, until the deviation is only a small fraction of the standard deviation of the original instance. At that point, the splitting process ceases, and the model is pruned back by one node. At the new tree branch termination, called the leaf, a regression model is constructed to describe the data reaching that leaf. This construct creates a non-linear, piecewise function that describes the data. Model trees deliver better compactness and prediction accuracy in comparison to classical regression trees (Deepa et al. 2010). However, a fault often seen in model trees is a tendency toward overfitting (Hastie et al. 2008). In order to avoid such problems, an ensemble method of generalization is used. The random forest corrects for this by assimilating multiple decision trees based on selections of data points, building multiple decision trees, and using a voting model based on the aggregated results, a method similar to bootstrap aggregating. Machine learning algorithms can be inherently unstable, with small changes in the training data, producing vastly different models. Bootstrap aggregation works by resampling with replacement on the training data multiple times

and then averaging the mean of the member models (Cannon and Whitfield 2002). This averaging process effectively controls model variance, preventing errors due to instability of the model and/or limited training data, without increasing the overall bias (Breiman 1996). In this lies the strength of the random-forest algorithm. A noted limitation of random forests is the inability to predict regression beyond the range of the training data. Due to the multifractal nature of soil moisture, this may pose a problem in that as we increase soil moisture resolution, we also expect to increase the maximum value in our population. While a higher-resolution sampling method may help alleviate this problem, it is a known limitation within this study.

This downscaling algorithm (Figure 2) is a form of transfer learning in which rules are learned at the native SMAP resolution using coarsened Landsat data and then applied at the native Landsat resolution to achieve a higher-resolution product. The Landsat values must be aggregated up to the SMAP pixel size to determine the localized ruleset. We used ESRI's raster aggregation using a median strategy to match the SMAP pixel size; however, the process can be repeated in QGIS or other open-source systems. We then used a random forest in Weka, due to its superior performance over other algorithms in our local testing.

The rules from the aggregated Landsat data are then utilized at the Landsat native resolution of 30m to provide the downsampled soil moisture surface. The rules are variable across regions and not transferable across Landsat scene sets, due to the differences in the land surface model and the ability of the spectral and thermal data to fully encompass the full range of periodic and localized events. The term "scene set" refers to a single scene as well as the previous and following scene in a single path. While the algorithm rule set must be recreated for each scene set, the overall algorithm remains the same. In a continental-scale implementation of this process, a moving window approach would be appropriate in which each Landsat scene is processed against the corresponding SMAP granule as well as the granule above and below in the path.

Essentially, this is a simple process for determining soil moisture in relatively arid and semi-arid environments. Landsat data is aggregated to the SMAP scale, and the six aforementioned variables are created to train against the SMAP soil-moisture estimate, which is the dependent variable here. The trained random-forest algorithm is then kept and applied at the native Landsat resolution to create much higher-resolution soil-moisture estimation. This algorithm must be rerun for each new Landsat scene, as

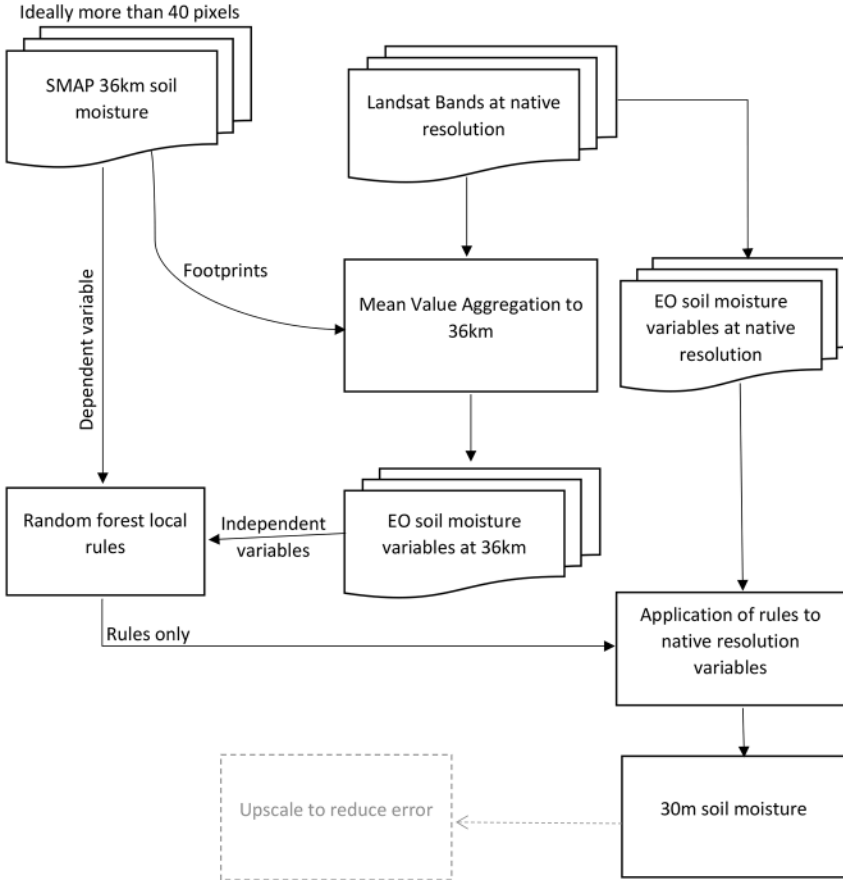

Figure 2.—Process flowchart of localized downscaling process via inferred learning of soil moisture. See Figure 3 for a map of multiple SMAP grid pixels captured in three granules from a single pass. Variable selection is not a repeated process and is therefore not part of the process; the variables are predefined above, and we suggest maintaining those relationships.

local variations in insolation, humidity, soil drying, and topography change the rule set. We propose that three adjacent along-track Landsat scenes should be used to provide ample training data; indeed, using a larger area should be avoided due to biome changes. Furthermore, the area of each Landsat scene is most likely the maximum area that can be covered in one run of this algorithm. The training data should be an along-track moving window encompassing both adjacent scenes, if possible. Aggregating the

final product may be required for minimal error, but this study was run at the native Landsat resolution of 30m.

Validation Data

Ground truth over the area is available via an excellent dataset of capacitance or frequency domain reflectometry (FDR), which will be discussed below. However, the coverage of this dataset is minimal compared with the entire range of the tested region. An airborne L-band radiometer (PLMR) also obtained readings on the same day and is the primary source used to validate the localized downscaling method. However, a field within the test region, referred to as field C in this study (SMAPEx Y7), was purposely set aside in feature selection to not contaminate data when testing. In-situ soil moisture will be compared to both the PLMR data and the rule transference algorithm. Ideally, the rule transference algorithm will perform as well as the PLMR, an already accepted method of soil moisture estimation, over field C.

Results

Airborne L-band radiometry is well accepted as a reliable, remote-sensing soil-moisture measuring technique. Overall, the predictive power of the downscaled soil moisture is satisfactory, although just outside the NASA standard of 0.04 for the SMAP validation program (NASA 2020). The dataset includes the full population of 3,097 pixels by 2,257 pixels, for a total of 6,989,929 samples, with an absolute error of 0.054 over the L-band retrieved surface (Tables 1 and 2; Figures 3 and 4).

Table 1. The mean of the proposed algorithm is 0.26, and the mean of the L-band retrieval is 0.28; the mean absolute error (MAE) is 0.05.

	RT Algorithm	L-Band	Error	MAE
Mean	0.2586	0.2756	0.017	0.0535
StdDev	0.0662	0.0285	0.0628	0.037
Skew	-0.4595	-1.4099	0.3754	0.7348

The comparison between the Landsat solution and the original SMAP data is admittedly imperfect, because the Landsat-inferred solution does not cover all the SMAP pixels involved. Therefore, variance increases with the observation resolution. This provides some support to the self-affine fractal distribution of soil moisture. The Landsat solution has a standard deviation of 0.066, while the six SMAP pixels overlapping the Landsat investigation

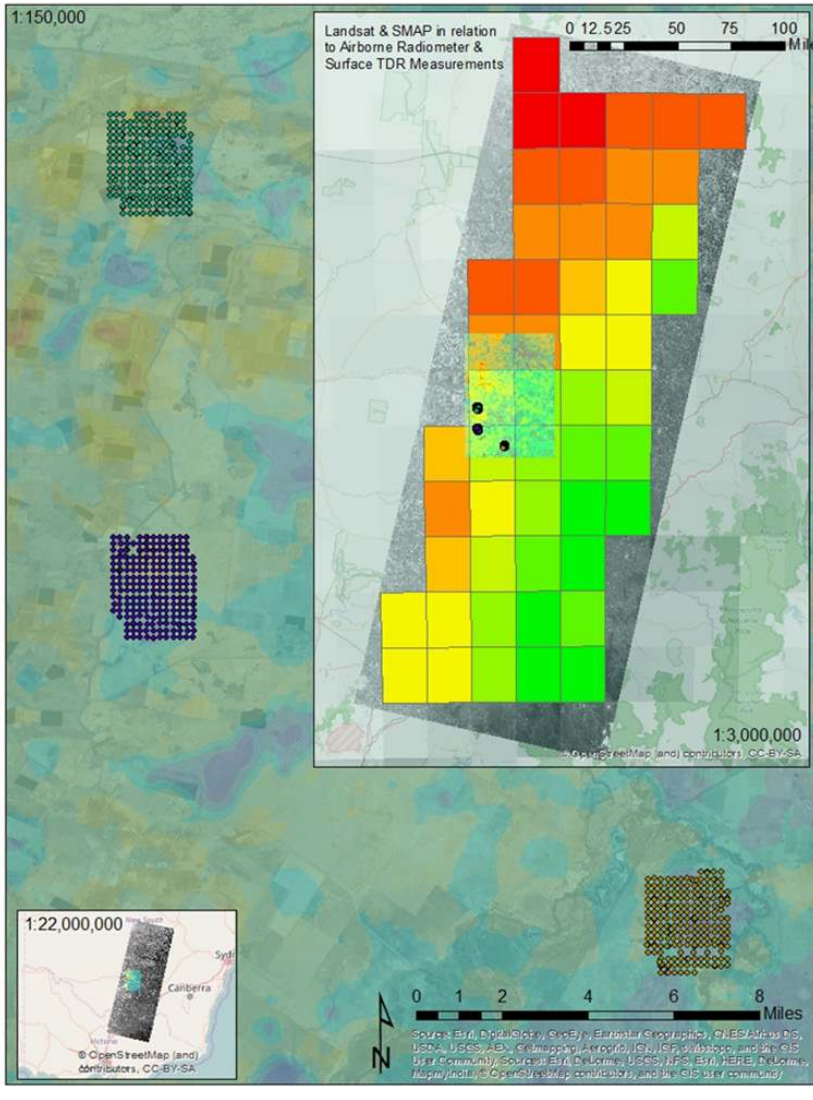

Figure 3.—Field soil-moisture measurement grids in which capacitance measurements were conducted in triplicate within the field of the airborne radiometers collection. The radiometer is within the SMAP sampling and Landsat aggregation region (see upper right inset) on 14 August 2015. The north and southeast points were included in the original training set and excluded from the testing set.

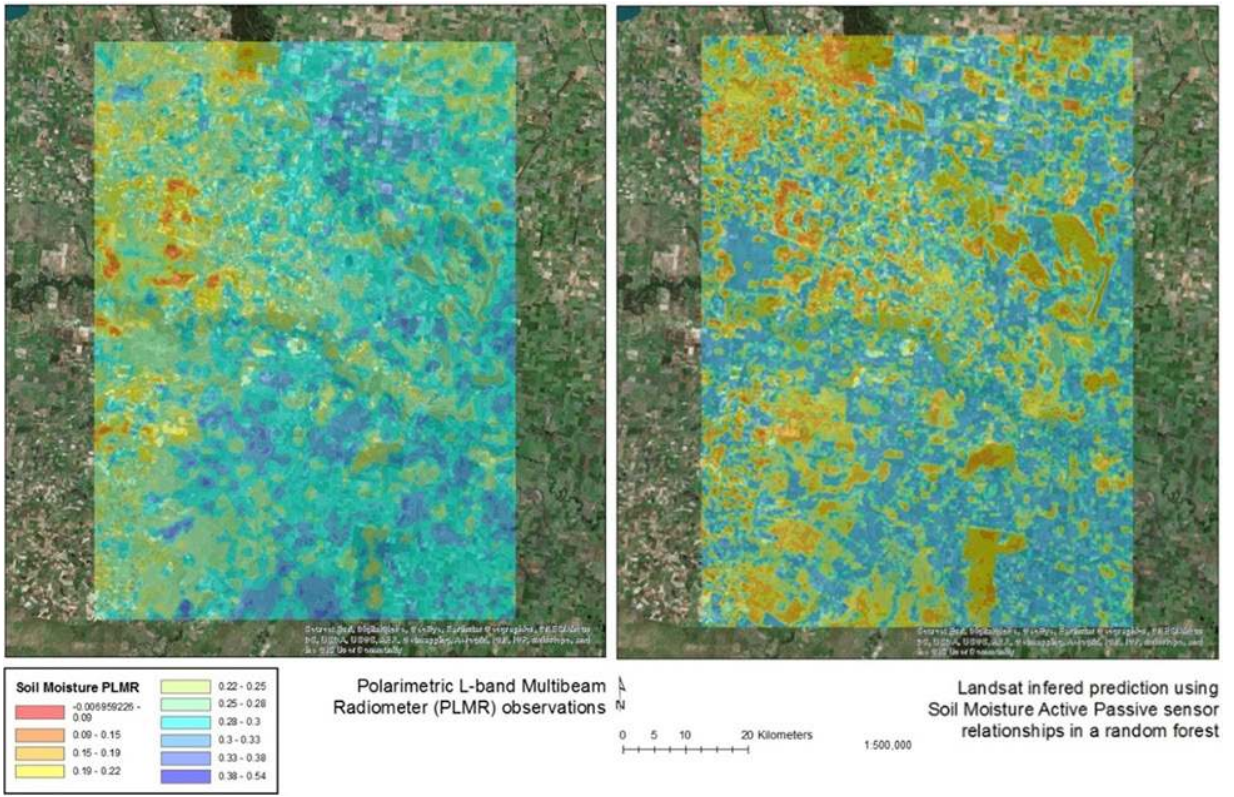

Figure 4.—A visual comparison of the L-band radiometer and the Landsat-inferred soil moisture shows generally good agreement overall, with areas of notable exception, such as the southern edge, just west of the center.

region have a standard deviation of 0.038. As the standard deviation is the square root of the variance, the variance of the Landsat solution is 0.0044, and the variance of the SMAP distribution is 0.0015. This increased variability in the Landsat solution suggests the learning algorithm was able to determine increased variability in the data at higher resolution, as would be expected in a self-affine multi-fractal system. The total mean of the Landsat solution is quite high compared with the SMAP mean of 0.21; however, again, the SMAP pixels extend beyond the Landsat investigation region into drier regions to the north, and the Landsat mean is still lower than the L-band PLMR solution. Kurtosis and skewness in both datasets remain remarkably similar, with the SMAP skewness at −0.59 and the Landsat solution at −0.48, while kurtosis is −1.68 and −1.25, respectively.

Table 2. Magnitude of error within the algorithm, based on the L-band surface. Currently, 56.44 percent of the results meet the NASA specifications of 0.04, indicating that the L-band retrieval was accurate.

Pct. of population within each error bin	
< -0.06	6.56%
-0.06 : -0.04	11.82%
-0. 4 : 0.00	33.15%
0.00 : 0.04	11.47%
0.04 : 0.06	6.67%
> 0.06	30.34%

While qualitatively the two surfaces derived from PLMR and Landsat provide similar solutions to volumetric soil moisture, a more quantitative review is required. Three high-density soil-moisture FDR point fields were obtained on the same test day. The central field (Site C) was left out of any portion of the training and testing on variable selection to provide a completely independent dataset. This is likely not a necessary precaution, given the wide range of other datasets included, but it is generally a good practice to test on truly independent data that have not influenced testing in any way. We evaluated the inferred solution against the radiometer and the point-collected data, although the FDR point data is for informational purposes only at sites N and SE.

First, an evaluation of the Landsat-inferred rule-transfer soil moisture versus the radiometer measurement from the PLMR shows a very good agreement of the field site, with an R2 of 0.8278 (Figure 5). This is in stark contrast to the point data that has a much lower agreement, with an R2 of

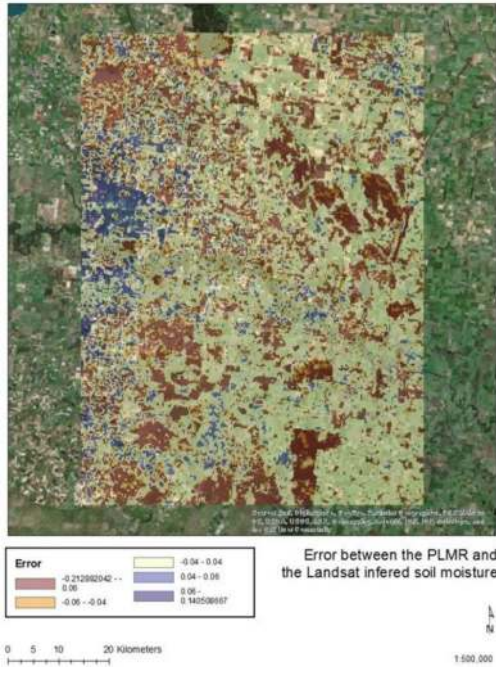

Figure 5.—Qualitative review of error between the L-band retrieval and the Landsat–SMAP downscaling algorithm.

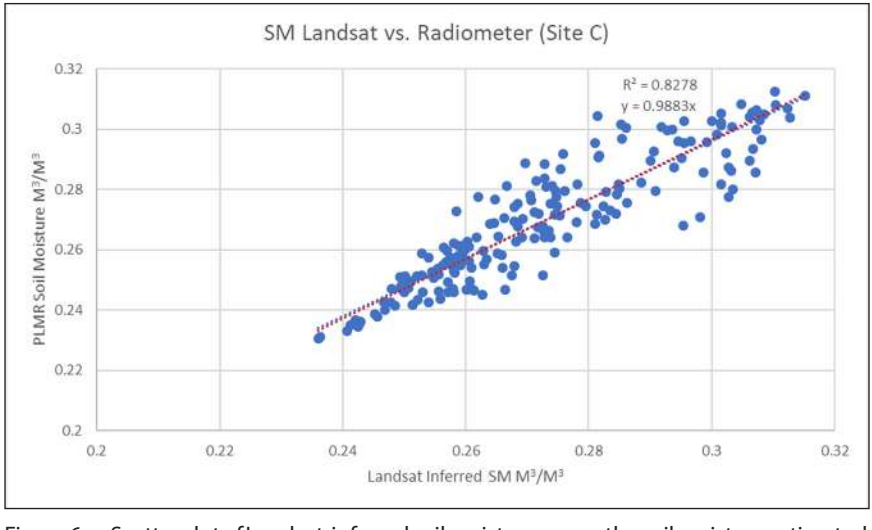

Figure 6.—Scatterplot of Landsat-inferred soil moisture versus the soil moisture estimated from the PLMR at Site C.

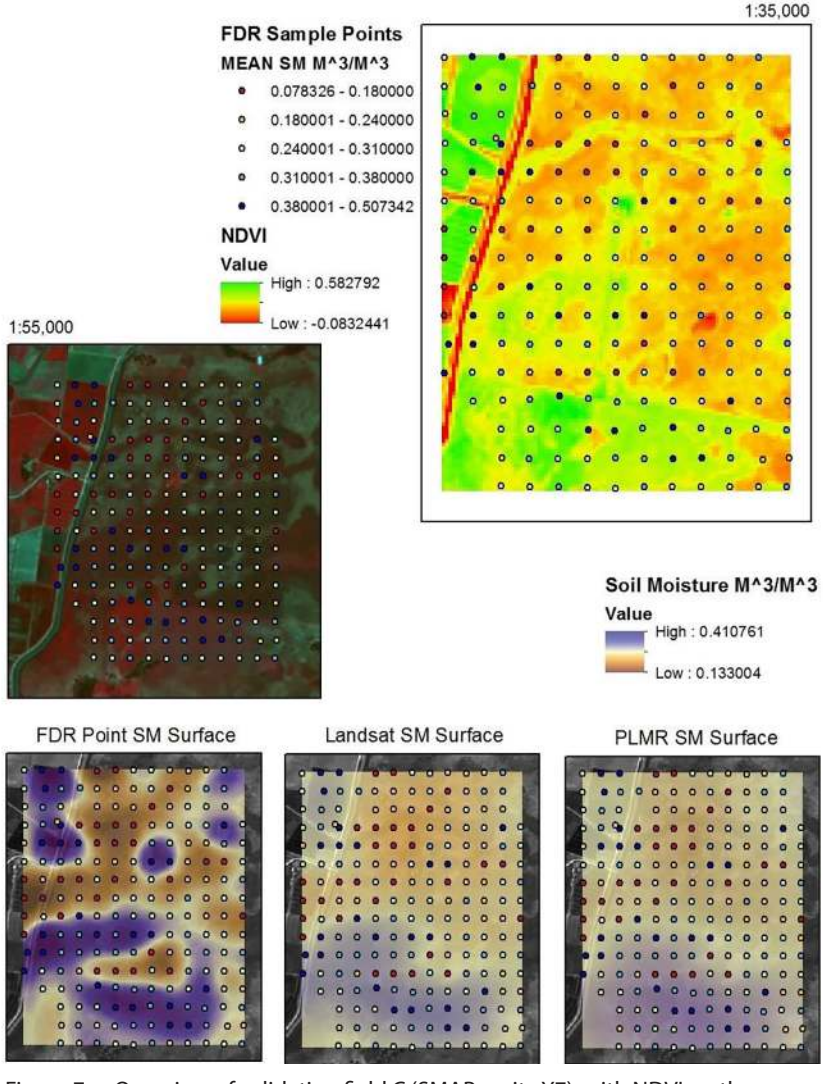

Figure 7.—Overview of validation field C (SMAPex site Y7), with NDVI on the upper right with overlain field-collection points. On the center left is a false-color image showing the vegetated field on the west side of the road, while the east is primarily ranchland. The lower three images are kriged soil-moisture surfaces from point data (left), the Landsat-inferred method (center), and PLMR soil moisture data (right).

0.2985. Nevertheless, given the R2 of the PLMR versus the soil moisture in the sample field of 0.2026, this seems a reasonable, albeit low, value (Figure 6).

To confirm that the population from the Landsat-inferred method is as valid as that from the L-band PLMR method, we conducted an equivalence test, which is a two one-sided t-test (TOST). This test is useful for proving equivalence in the same manner as the standard t or F tests are used to prove dissimilarity. The TOST results for the control group 1 (PLMR) and the test group 2 (Landsat-inferred method) compared over the entirety of the three sample fields are significant to $p = 0$ based on both Cohen's d and raw scores, regardless of an assumption of equal variance.

Discussion

Our method shows considerable fitness in describing the soil moisture distribution at a significantly higher resolution than other currently available methods using SMAP alone. The method utilizes the refined SMAP algorithm, ingraining in the results the description of the soil-moisture relationship to the spectral conditions on the ground. While the method does not provide accuracy within the 0.04 range required by the NASA SMAP mission in this experiment, this may be achievable with additional variable refinement or additional validation. Capacitance soil monitors, like all point-based *in-situ* soil measurements, observe ground conditions in a highly localized area, no more than 1.2in beyond the waveguides (Munoz–Carpena et al. 2004). Since such a small sample of the pixel space is tested, the point soil-moisture values assigned to a pixel in the original interpretation of the L-band may not be accurate in describing the true value of the pixel, but instead may fall within the tail of the distribution of the soil-moisture values within a pixel. This can be demonstrated by evaluating the number of moisture sample points within a 30-m Landsat pixel and the error of that averaged point moisture value with the downscaled Landsat-inferred algorithm results. As the number of point measurements increases within a single cell, the error from both the inferred and the PLMR data to that measured aggregated soil-moisture value is expected to decrease. Over the study region, cells with an average of one to four measurements have an average error of $0.1 m^3/m^3$, which decreases to about $0.08 m^3/m^3$, with five readings per cell, and falls below $0.05 m^3/m^3$ error with seven readings. In short, over a large validation field, point samples are insufficient for validation at lower counts, as their coverage is far below the required one in adequately defining the pixel value. As this study was designed for lower-resolution validation

of the SMAP instrument, it may not serve to compare the inferred solution to soil-moisture FDR points. This is an ongoing problem in the study of soil moisture. Direct measurement is required for ground truth, but this is highly laborious and spatially covers a small area.

While the sheer number of observations aids the alleviation of some of the training errors, a large number of them can affect the training algorithm that allocates the transition variables from a raw L-band return to soil-moisture content. Among the three sampling regions of soil-moisture point collection, two were used for scaling the L-band radiometer data while the third (Site C, see Figure 6) was used for testing. That region (Site C) experienced a 0.0628 Mean Average Error (MAE), not too dissimilar from the slightly lower 0.0606 MAE seen over the same test region for the Landsat-inferred data and the point measurements. However, the Landsat-inferred data and the radiometer demonstrated a significantly higher coupling in their returns with a 0.0071 MAE, suggesting that the Landsat/SMAP-inferred algorithm produced a dataset with a stronger description of the L-band radiometer data than that of the point sampling, and thus is a viable option for downscaling SMAP into the local resolution (Table 3). The Root Mean Standard Error (RMSE) for the Landsat/SMAP rule transference algorithm to FDR measured values is 0.105 over the entirety of the scene. Given the heterogeneity of the scene, the MAE and RMSE values of the target evaluation sites are shown in Table 7.

Table 3. Mean Absolute Error (MAE) and Root Mean Standard Error (RMSE) over selected test sites (C, N, SE). RF is the Random Forest of the Landsat-inferred rule transference method, while FDR is the frequency domain site sampling. Micro refers to the PLMR.

n = 515

All	Mean Abs Err	RMSE
RF-Micro	0.007359559	0.009398089
RF-FDR	0.068428304	0.091237308
Micro-FDR	0.070209739	0.093150219

Continued

n = 197

Site C	Mean Abs Err	RMSE
RF-Micro	0.00710297	0.009136707
RF-FDR	0.060604748	0.074836289
Micro-FDR	0.062825105	0.077160298

n = 210

Site N	Mean Abs Err	RMSE
RF-Micro	0.008557376	0.010510188
RF-FDR	0.061731992	0.079604448
Micro-FDR	0.064656494	0.082503581

n = 208

Site SE	Mean Abs Err	RMSE
RF-Micro	0.006393245	0.00839998
RF-FDR	0.082598815	0.113624224
Micro-FDR	0.082810481	0.114648599

An obvious disadvantage of our method is that visual remote sensing cannot obtain imagery through cloud cover, as opposed to a microwave. Nevertheless, our algorithm does allow for widespread, moderately high-resolution soil-moisture capture over large regions. This could be a viable option for detecting moderately high-resolution soil moisture in arid and semi-arid regions, as well as for soil-moisture pattern-distribution analysis.

This easily accessible algorithm is highly reproducible and provides a framework for users to understand soil-moisture conditions and distributions in localized areas at an actionable resolution. This is especially useful for those who need such data but do not have extensive resources on hand, as this can be programmatically run using public data and without specialized training. The methodology is robust and learns site and seasonal variation on the fly, creating a unique solution for each instantiation of the algorithm. While other methodologies have created similar solutions, most notably the UCLA method of Jiang and Islam (2003), using temperature and vegetation indices as proxies for downscaling, few have done so in a machine-learning environment. Those that have produced similar solutions did not perform a thorough, variable pairing process to identify the key variables governing

the process, nor did they include multidimensional variables. This method is repeatable and programmable, and the learning algorithm changes for each iteration, but the processing never alters. No information is needed, other than the passive microwave soil-moisture and Landsat Earth Observation Remote Sensing data (although other optical systems may be used). The machine-learning algorithm also enables something none of the other algorithms address, which is differentiating soil moisture schemes existing within the same scene, such as a swath of drying soils in the same scene as saturated soils.

Two new variables have been identified from the exhaustive variable pairing process: Variable 16 and Variable 10. Variable 16 is a dimensional expansion of NDVI, an alternative methodology for quantifying photosynthesizing vegetation. Variable 10 is a complex variable corresponding to the total visible reflectivity but modified by a thermal response. These two variables, along with the other four identified as key to describing soil moisture, seem to be able to describe well the SVAT. This was hypothesized after a completely independent exploration of variables was completed. Instead of using deterministic models to guide the variable selection, we used inferred learning to identify the variables and then noticed they seemed to be describing the currently used model.

This easy soil-moisture algorithm requires a clear atmosphere and is not yet demonstrated over non-arid or semi-arid environments. However, for field validation techniques, periodic snapshots of soil-moisture conditions for agriculture, traffic ability, health vectors, water management, and forecasting, this approach could provide an accessible methodology for high-resolution soil moisture to just about anyone with a computer and an Internet connection.

Literature Cited

Ahmed, M. Y., A. H. Nury, F. Islam, and M. J. Alam. 2012. Evaluation of geotechnical properties and structural strength enhancing road pavement failure along Sylhet-Sunamganj highway, Bangladesh. *Journal of Soil Science and Environmental Management* 3 (5):110–117. doi: 10.5897/JSSEM12.024.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24 (1996):123–140. doi:10.1007/BF00058655.

Cannon, A. J., and P. H. Whitfield. 2002. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. *Journal of Hydrology* 259 (1):136–151. doi: 10.1016/S0022-1694(01)00581-9.

Carlson, T. 2007. An overview of the "Triangle Method" for estimating surface evapotranspiration and soil moisture from satellite imagery. *Sensors* 7 (8):1612–1629.

Chauhan, N., S. Miller, and P. Ardanuy. 2003. Spaceborne soil moisture estimation at high resolution: A microwave-optical/IR synergistic approach. *International Journal of Remote Sensing* 24 (22):4599–4622. doi: 10.1080/0143116031000156837.

Cosby, B. J., G. M. Hornberger, R. B. Clapp, and T.R. Ginn. 1984. A statistical exploration of the relationship of soil moisture characteristics to the physical properties of soils. *Water Resources Research* 20 (1984):682–690. doi:10.1029/WR020i006p00682.

Crosta, G. 1998. Regionalization of rainfall thresholds: an aid to landslide hazard evaluation. *Environmental Geology* 35 (2-3):131–145. doi:10.1007/s002540050300.

De Michele, C., and G. Salvadori. 2002. On the derived flood frequency distribution: Analytical formulation and the influence of antecedent soil moisture conditions. *Journal of Hydrology* 262 (1-4):245–258. doi:10.1016/S0022-1694(02)00025-2.

Deepa, C., K. SathiyaKumari, and V. Prem Sudha. 2010. Prediction of the compressive strength of high-performance concrete mix using tree-based modeling. *International Journal of Computer Applications* 6 (5):18–24.

Denmead, O. T., and R. H. Shaw. 1962. Availability of soil moisture to plants as affected by soil moisture content and meteorological conditions. *Agronomy Journal* 54 (5):385–390.

Hastie, T., R. Tibshirani, and J. Friedman. 2008. *The Elements of Statistical Learning*, 2nd ed. New York: Springer.

Hunt, E. D., K. G. Hubbard, D. A. Wilhite, T. J. Arkebauer, and A. L. Dutcher. 2009. The development and evaluation of a soil moisture index. *International Journal of Climatology* 29 (5):747–759. doi:10.1002/joc.1749.

Jackson, R. D., S. B. Idso, R. J. Reginato, P.J. Printer, and J.L. Hatfield. 1981. Canopy temperature as a drought stress indicator. *Water Resources Research* 17 (4):1133-1138.

Jiang, L., and S. Islam. 2003. An intercomparison of regional latent heat flux estimation using remote sensing data. *International Journal of Remote Sensing* 24 (11):2221–2236. doi:10.1080/01431160210154821.

Koster, R. D., P. A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C. T. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C. H. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Oleson, A. Pitman, Y. C. Sud, C. M. Taylor, D. Verseghy, R. Vasic, Y. Xue, and T. Yamada. 2004. Regions of strong coupling between soil moisture and precipitation. *Science* 305 (5687):1138– 1140. doi: 10.1126/science.1100217.

Monerris, A., and T. Schmugge. 2009. Soil moisture estimation using L-band radiometry. In *Advances in Geoscience and Remote Sensing,* G. Jedlovec, ed. Shanghai: InTech.

Munoz-Carpena, R., S. Shukla, and K. Morgan. 2004. *Field devices for monitoring soil water content.* Southern Regional Water Program.

Onyari, E. K., and F. M. Ilunga. 2013. Application of MLP neural network and m5p model tree in predicting streamflow: A case study of luvuvhu catchment, South Africa. *International Journal of Innovation, Management and Technology* 4 (1):11–15. doi: 10.7763/IJIMT.2013.V4.347.

Orchard, V. A., and F. J. Cook. 1983. Relationship between soil respiration and soil moisture. *Soil Biology and Biochemistry* 15 (4):447–453. doi:10.1016/0038-0717(83)90010-X.

Panciera, R. W. 2013. The soil moisture active passive experiments (SMAPEx): Towards soil moisture retrieval from the SMAP mission. *IEEE Transactions on Geoscience and Remote Sensing* 52 (1):490–507. doi: 10.1109/TGRS.2013.2241774.

Panciera, R., J. P. Walker, and O. Merlin. 2009. Improved understanding of soil surface roughness parameterization for L-band passive microwave soil moisture retrieval. *IEEE Geoscience and Remote Sensing Letters* 6 (4):625–629. doi: 10.1109/LGRS.2009.2013369

Peng, J., A. Loew, O. Merlin, and N. E. C. Verhoest. 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews in Geophysics* 55 (2):341–366. doi:10.1002/2016RG000543.

Verena, H. G. K. 2012. *Index DataBase*. Retrieved from https://www.indexdatabase.de/

Weizu, G., and J., Freer. 1995. Patterns of surface and subsurface runoff generation. *Tracer Technologies for Hydrological Systems* 229:265–273. Boulder, CO: IAHS.

Wu, X., J. P. Walker, C. Rüdiger, and R. Panciera. 2015. Effect of land-cover type on the SMAP active/passive soil moisture downscaling algorithm performance. *IEEE Geoscience and Remote Sensing Letters* 12 (4):846–850.

N. Ye, J. P. Walker, X. Wu, R. de Jeu, Y. Gao, T. J. Jackson, F. Jonard, E. Kim, O. Merlin, V. Pauwels, L. Renzullo, C. Rudiger, S. Sabaghy, C. von Hebel, S. H. Yueh, and L. Zhu. 2017. The Soil Moisture Active Passive Experiments: Towards calibration and validation of the SMAP Mission. *Remote Sensing of Environment*. In Review.

Zhan, X., S. Miller, N. Chauhan, L. Di, and P. Ardanuy. 2002. *Soil moisture visible/infrared radiometer suite algorithm theoretical basis document*. Lanham, MD.