 doi:10.20944/preprints201902.0046.v1

Article

Localized Learning of Downscaled Soil Moisture

Michael Lewis* [1,3], Andmorgan Fisher [1,2], Clint Smith [1,2], John J. Qu [3] and Paul Houser [3]

[1] US Army Corps of Engineers, ERDC—Geospatial Research Lab., Alexandria, VA 22315, USA; miglewis@gmail.com (M.L.); andmorgan.r.fisher@usace.army.mil (A.F.); clint.b.smith@usace.army.mil (C.S.)

[2] Department of Environmental Science & Policy, George Mason University, Fairfax, VA 22030, USA

[3] Department of Geography & Geoinformation Science, George Mason University, Fairfax, VA 22030, USA; jqu@gmu.edu (J.J.Q.); phouser@gmu.edu (P.H.)

* Correspondence: miglewis@gmail.com

Abstract: If given the correct remotely sensed information, machine learning can accurately describe soil moisture conditions in a heterogeneous region at the large scale based on soil moisture readings at the small scale through rule transference across scale. This paper reviews an approach to increase soil moisture resolution over a sample region over Australia using the Soil Moisture Active Passive (SMAP) sensor and Landsat 8. This approach uses an inductive localized approach, replacing the need to obtain a deterministic model in favor of a learning model. This model is adaptable to heterogeneous conditions within a single scene unlike traditional polynomial fitting models and has fixed variables unlike most learning models. For the purposes of this analysis, the SMAP 36 km soil moisture product is considered fully valid and accurate. Landsat bands coinciding in collection date with a SMAP capture are down sampled to match the resolution of the SMAP product. A series of indices describing the Soil-Vegetation-Atmosphere Triangle (SVAT) relationship are then produced, including two novel variables, using the down sampled Landsat bands. These indices are then related to the local coincident SMAP values to identify a series of rules or trees to identify the local rules defining the relationship between soil moisture and the indices. The defined rules are then applied to the Landsat image in the native Landsat resolution to determine local soil moisture. Ground truth comparison is done via a series of grids using point soil moisture samples and airborne L-band Multibeam Radiometer (PLMR) observations done under the SMAPEx-5 campaign. This paper uses a random forest due to its highly accurate learning against local ground truth data yet easily understandable rules. The predictive power of the inferred learning soil moisture algorithm did well with a mean absolute error of 0.054 over an airborne L-band retrieved surface over the same region.

Keywords: Soil Moisture; Remote Sensing; Landsat; SMAP; Random Forest; Machine Learning; Downscaling; Microwave

1. Introduction

Soil moisture is a measurement of the hydrological component within a finite amount of soil. The variability of soil moisture is highly dependent on the soil properties [1], the local geologic situation [2], vegetation density and draw [3], and antecedent conditions. Soil moisture is a vital key to understanding the physiologic condition of the earth for many systems. Agriculture relies on soil moisture for plant vigor and growth, and weather and climate models are strongly coupled to land-atmosphere interactions [4]. Flood control is also highly dependent on antecedent soil moisture conditions [5].

 doi:10.20944/preprints201902.0046.v1

1.1. The Self-Affine Fractal Property of Soil Moisture

In 1995 Ignacio Rodriguez-Iturbe et al. explicitly stated that "spatial correlation remains unchanged with the scale of observation and follows a power law decay." This means that variation in the soil moisture field is related to observation area in a log relationship, decreasing with increasing observation size. Each larger aggregate observation is a conservative process and equal to the mean of the component observations. This in concert with the unchanged spatial correlation suggest a multifractal relationship in the soil moisture field, one that is dependent on scale of observation. This multifractal attribute demonstrates a smoothing in lower resolution but shows no significant change in variation, described by Pelletier et al. [6] as a self-affine fractal, a fractal with self-similarity but scales differently.

Fortunately, many other systems also demonstrate a multifractal characteristic. Mirás-Avalos et al. [7] characterized this lack of homogeneity in soil moisture storage as a complex function of multiple nested processes "related to topographical, geological, edaphic, and vegetation factors." As soil moisture distribution is linked strongly to these independent variables, it can be hypothesized that these attributes can also represent the self-affine fractal qualities of moisture distribution in their own spatial organization. Many deterministic models of soil moisture acknowledge the strong binding of these factors to spatial moisture distribution and so the use of variables describing these factors should promote a solution resulting in a self-affine fractal result increasing variation with decreasing observation areas.

1.2. Optical

The reflectance properties of soil moisture are affected by drier soils having a higher overall albedo due to the absorbing properties of water in pores filled with soil [8]. This action, described by Ångstom in 1925 as arising from the property of radiation being absorbed while it is transmitted through soil water films before and after reflectance [9], was confirmed by Planet in 1970 [10] and again by Grasser and Van Bavel in 1982 [11]. Ångstom described the reflective property of the wetting of natural soils, assuming no change to reflectance of the surface material, by a liquid with the index of refraction (n) where the reflectance when wet (ϱ') and that when dry (ϱ) as:

$$\rho' = \frac{\rho}{n^2(1-\rho)+\rho}. \quad (1)$$

Testing by Planet showed this relationship to be effectively valid though some effects did cause some deviation between expected and measured reflectance. Some minerals can change the refractive properties of the liquid as the water becomes a solution of the constituent components. This effect does not disqualify the above statements but merely changes the variable of refraction. Likewise, the surface reflectance may be altered via the Christiansen effect. This effect occurs when the refractive indices of the liquid and solid are nearly identical.

Twomey et al., in 1986, concluded Ångstom's argument for geometric mechanism for the darkening of a surface would hold true if other solvents were not included but that it is in actuality the forward scattering of photons arising from the increasing refractive index in the soil-liquid boundary as opposed to the soil-air boundary in dry surfaces [12]. This increase in forward scattering thus increases absorption and had Ångstom, and later Planet, evaluated other solutions with differing refractive properties they may have concluded the same. While the mechanism is accurately described by Twomey et.al, in standard practice the relationship described by Planet and Ångstom holds true and is valid in remote sensing practice.

While reflectance and absorption have both been contemplated as potential sources of the darkening effect of soils an examination of the absorption and reflectance of water compared to the spectral signatures of wetted soils reveals something obvious. The same soil, when wetted, does not significantly change its spectral signature but instead shifts downward (see Figures 1 and 2). Given the rather smooth reflectance curve and the lack of significant alteration of the spectra of wetting soils in concordance with the absorption spectra of water. It appears that the darkening is largely

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

independent of absorption effects of surface water [13] and only affected by the strong absorption bands of water in the pore water of soil, this in effect suggests that both

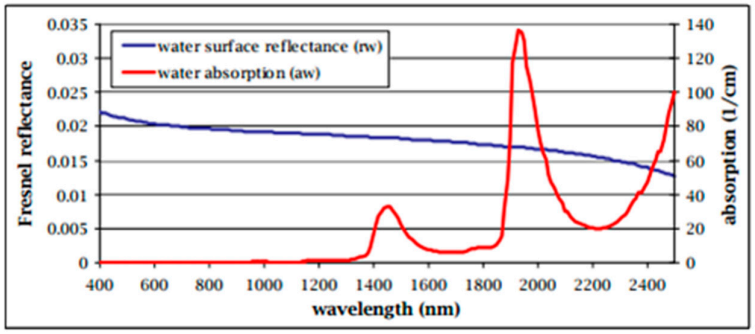


Figure 1. Spectral reflectance and absorption of water [9].

Ångstom and Twomey were correct. The surficial water affecting spectra strongest in reflectance while the weaker effect of forward scattered radiation into the pore water is more strongly affected by absorption spectra of water.

In practical terms for remote sensing in the visible and infrared wavelengths, those from 0.43 to 2.295 μm, this is codified in a standard decrease in reflectivity identifiable through the ratioing of imagery bands. Testing of this by Musick and Pelletier in 1988, utilizing Landsat Thematic Mapper bands TM5/TM7 found a correlation of $r^2 = 0.88–0.99$ for moisture and spectral decrease in the ratio within any one of ten different soil types, yet overall, as a population, the correlation was much lower at $r^2 = 0.23$ [14]. Much the same result was found in a later study by Muller & Décamps in 2000 wherein they stated, "Models are more efficient when computed by soil category, but the efficacy remained when soils are partly grouped excluding silt soils: soils moisture can be estimated with a mean error of 3.3% [14]." Furthermore, Muller & Décamps found that the impact upon reflectance of soil moisture is a greater factor then that of soil category [15].

The standard decrease in soil moisture is simulated by Wang and Qu in 2009 modeling MODIS response to soil moisture levels [16]. This simplified model, while accurate, does not account for variation based on soil type or properties. Soil spectra of various soil types demonstrate wide variability but similarity in their non-linear rate of change in spectra with respect to change in soil moisture [17] (see Figure 2). Using this relationship, a ratio relationship may be able to be established to approximately account for soil moisture content regardless of soil characteristics. This is in effect the point of the model put forth by Wang and Qu.

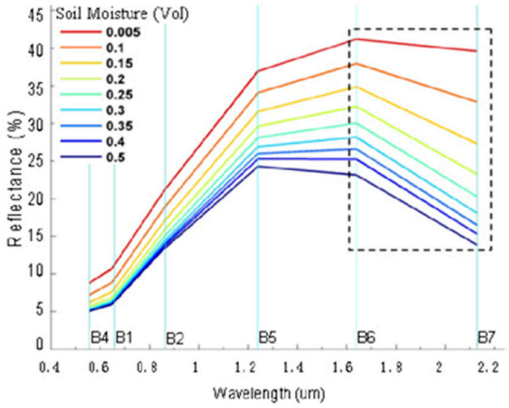


Figure 2. Soil Moisture spectra modeled by Wang and Qu with MODIS Bands 1–2, 4–7.

Variation among soils was further described by Liu et al. stating the variation among 10 soil types chosen specifically for their variation showed a general decrease of reflectance across all bands

 doi:10.20944/preprints201902.0046.v1

for an increase in moisture with a more pronounced variance for longer wavelengths, particularly in water absorption bands from 1450 to 1940 nm [18]. Using a linear forward stepwise regression, Liu et al. found seven bands to be consistent across soil types in relation to explaining the residual with any other band. Those are, in order: 1400, 986, 1998, 574, 2189, 1672, and 450 nm. These markers can be utilized and standardized markers in descending order of stability in the creation of a ratio standard to eliminate soil spectra differences arising from properties inherent to individual samples. Whiting et al. regarded many of these bands as unusable for soil moisture modeling through remote sensing due to the proximity or co-location with primary bandwidths experiencing significant atmospheric attenuation due to water vapor and CO_2 absorption in the atmosphere.

Working around the regions of atmospheric attenuation, several groups have developed novel approaches to account for Soil Moisture using Imaging Spectroscopy (IS). Following the work of Whiting et al. in 2004 [19] in determining a soil moisture Gaussian model (SMGM) based on the area of a convex hull, where spectra are normalized by dividing the reflectance at each band by the maximum reflectance centered at the water absorption center at 2.8 μm. This work was successful at estimating water content between two disparate regions to a coefficient of determination of 0.94 with an RMSE of 1.7 to 2.5% [20].

While this work extended outside the range of most common sensors, later work by others began to move into commonly used bands. Using a novel Normalized Soil Moisture Index (NSMI), Haubrock et al. utilized a ratio of the 1.8 μm–2.119 μm/1.8 μm + 2.119 μm to quantify soil moisture achieving a coefficient of determination of 0.61–0.71 given certain influencing factors such as vegetation cover and soil type [20,21].

1.3. Microwave

Remote sensing of soil moisture is most sensitive in the electromagnetic spectrum in the range of 1 to 5 GHz (6 cm–30 cm wavelength) due to the maximum stretching of the dielectric effect of dry soil to water. In the range of 1.4–1.427 GHz (21.0 cm–21.4 cm wavelength), the L-band is not only the signal most sensitive to soil moisture, but the atmosphere is nearly transparent, and vegetation is semi-transparent allowing nearly universal viewing at the soil surface [22]. The negative to using such frequencies is the requirement for a large antenna for even very low resolution; using a classical orbiting arrangement without any novel approach, a 50 km resolution would require a 20 m antenna. Further confounding interpretation of microwave based remote sensing is the response to variations of surface roughness. The response however is not constant, but variable based on moisture content and soil type [23], with clay soils demonstrating noticeably more aberrant response than from other types [22].

This has been the primary area of investigation into soil moisture due to the soil moisture sensitivity in the L-band however, due to the wavelength size, resolution will remain coarse. The Soil Moisture and Ocean Salinity (SMOS) mission was the first soil moisture dedicated mission. Launched in 2009, the mission has a return period of 3 days and a pixel size of roughly 40 km [24], based on latitude. The SMAP mission launched in 2015 attempted to produce a higher resolution product (with a resolution better than 10 km), through the integration of the highly accurate and low-resolution passive radiometer with higher resolution active microwave, which is prone to vegetation and surface roughness effects [25]. 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 upon the instrument. This analysis used a subsample of SMAP and airborne L-band (PLMR) data with a center frequency of 1.413 GHz, and a band width of 24 MHz, acquired on 13 September 2015 as part of the SMAPEx-5 campaign over southern New South Wales, Australia.

1.4. Integration between Microwave and VIS/NIR

In 2014, Song, Jia, and Menenti [26] proposed a hybrid methodology for obtaining soil moisture at a 1 km resolution in which microwave soil moisture brightness data is downscaled through the use of land surface temperature and NDVI data. While the method does appear to hold some validity, additional research should be conducted with increased ground truth to support and ensure the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

results are not spurious, owing to the collinear relationship between NDVI and soil moisture. However, it is a worthy alternative due to the relative lack of subjectivity over SVAT methodologies. The key to this method is the downscaling of microwave brightness using the exponential relationship between the microwave polarization difference index (MPDI) and NDVI:

Equation (2) Microwave Polarization Difference Index (MPDI) and Normalized Difference Vegetation Index (NDVI) relation

$$\text{NDVI} = \text{E0} + \text{E1}*\exp(\text{E2}*\text{MPDI}), \quad (2)$$

where E_0, E_1, and E_2 are constants. The necessity for downscaling microwave data is due to the resolution being limited by the antennae length of onboard sensors. This is due to the dependent system parameter of beam width which is determined by the wavelength over the antennae length. Microwave satellite sensors would be far too large to gather high resolution data using current technology or budgetary constraints.

Combining this equation with the MPDI, a standard difference index between the vertical and horizontal polarizations of either 18.7 or 36.5 GHz to mitigate the influence of soil moisture in the temperature analysis, brings the equation:

Equation (2) Microwave Polarity relation with NDVI [26]

$$T_{18.7H} = \frac{1 - F_1 * \ln(F_2 * (NDVI - F_0))}{1 + F_1 * \ln(F_2 * (NDVI - F_0))} * T_{18.7V} \quad (3)$$

In which the constants can be solved when upscaling the NDVI to the same resolution as the original microwave data. Once the constants and NDVI are known, the downscaled horizontal polarization can be determined. Although not explicitly explained, it appears this can be done as follows:

If:

Equation (3) First constant relation

$$\text{B} = \frac{1 - F_1 * \ln(F_2 * (NDVI - F_0))}{1 + F_1 * \ln(F_2 * (NDVI - F_0))} = \frac{T_{18.7H}}{T_{18.7V}} \quad (4)$$

Then

Equation (4) Second constant relation

$$A_i = \frac{\text{B}_\text{i} - \text{D}_0 + \text{B}_\text{i}\text{D}_1\text{Ts}}{2A_1 - B_i D_2}; \quad (5)$$

Equation (5) derived temperature polarization relation [26]

$$C = D_0 + D_1 T_s + D_2 Ai \quad (6)$$

where D_0, D_1, D_2, and F_0, F_1, and F_2 are constants that can be determined, although it appears to do so, at least three pixels would need to have a known surface soil moisture. Song, Jia, and Menenti showed a distinct link between an earth observing sensor (EO) and a microwave L-band sensor, demonstrating the EO sensors can be successfully used as a dependent variable set against an independent variable soil moisture set to identify a higher resolution result. This establishes a direct relation between NDVI and the microwave polarization difference allowing us to use the NDVI to bypass the need for microwave in a visual wavelength learning algorithm.

However, the use of machine learning can also bypass much of the localized uncertainty due to seasonal and optical vegetation depth as well as reduce computation time significantly by establishing a series of variables that describe the relationship between imagery data and the SMAP soil moisture values and apply those rules over a single scene or local set of scenes in a single collect. Given enough training sets, the constants are deduced implicitly and inductively within the algorithm. Furthermore, machine learning can incorporate and identify soil moisture via optical and infrared easily in the same algorithm, not only using the visible and infrared earth observing sensors as a proxy for microwave but gaining valuable additional skill in determining true soil moisture distribution.

 doi:10.20944/preprints201902.0046.v1

2. Methods

2.1. Feature Attribute Selection

Feature selection is the act of pairing down variables from a full set of variables to those most able to increase performance and contain the least amount of redundancy. Models using many variables are simply inelegant. While machine learning can partially overcome inelegant solutions, the fundamental goal of machine learning in research is not the successful application of the algorithm but also to understand the reasoning behind the success of such an algorithm. High dimensional algorithms are often overly complex to understand for the researcher. Furthermore, for computational efficiency, the training times on lower dimension models are significantly faster and many times, more accurate than larger dimension algorithms, a situation known as the curse of dimensionality.

Initial variable exploration used field collected data from multiple sites including Nevada, New South Wales, and Arizona, over a span of two years in the spring, summer, and fall in the respective regions. 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 digital number (DN) values were extracted for each pixel over all bands over a sample set over the area of interest to conduct relationship testing and pairing of variables.

As the algorithm evaluates all variables are in terms of ratios within a single temporal-space setting, the DN values are appropriate in an uncorrected form. Variables are generated through the iterative ratio of each band with every other band. In the following iteration, the ratio is conducted with the initial bands as well as the new ratios, continuing through four iterations. Working exclusively with ratios largely nullifies atmospheric effects due to each band value being affected in a comparative fashion and all values being in a single set of scenes taking over a similar region on the same day. Since ratios look at total response in comparison to another total response, no other manipulation (such as atmospheric correction) is required. The sampled values are then compared via a simple correlation and those ratios performing best are brought forward for further scrutiny. This was done completely naively using multiple high-resolution collected field soil moisture surfaces. This data was then used to inductively learn the highest information variables in accommodating the soil moisture surface. The final model then should be able to be bound to a deterministic model describing the mechanics of soil moisture.

All the attributes were tested against field sampled soil moisture surfaces, collected using TDR or FDR (capacitance), for Pearson's correlation coefficient. To keep within the computing requirements and the column limits of the spreadsheet program, only the variables showing the highest positive or negative correlations to measured moisture were brought forward to the next two iterations. Those attributes scoring above an absolute 0.65 were brought forward for additional testing using a wrapper method with a random forest classifier. Although occasionally lower correlation attributes were brought forward if they showed a higher ranking in information gain, a secondary test used for marginal variables. With the highest performing variables, a collection of about 20, each was compared in terms of a 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 three-selection criterion with the exception of *Variable 16*, which was brought forward due to its consistency, and information gain, even though it rarely scored in the top three to five attributes. In final analysis it continued to demonstrate increased performance in the algorithm over exclusion.

Many initial pre-existing VIS/IR ratio variables were chosen in addition to the NDVI, using a full list of vegetation, soil, and hydrology related indices from the Index DataBase [27]. These variables were excluded from the initial correlation feature selection and immediately brought forward to the wrapper selection phase. Test of prediction skill, multicollinearity, and leave one out variable testing provided a base set of six variables total after five initially variables were discarded due to high multicollinearity with existing indices brought in at this stage.

 doi:10.20944/preprints201902.0046.v1

The majority of machine learning is, at its core, a classification exercise, often of complex relationships with incomplete information. The applicability of machine learning is in the ability to make reasonably accurate classifications of expected outcomes without attempting to encapsulate and explain each outlying error. Instead of being the end goal, the machine learned model is the initialization of the exploration of causality. When a sufficiently strong predictive power or classification is discovered, then the researcher has a point from which to begin describing the interaction and proxy effects causing the relationship. As such, the model should be kept as simple as possible using as few initialization parameters as possible to attain the desired outcome, which will help minimize generalization errors [28].

Because the final goal is not the algorithm itself but to identify the why behind the success of the algorithm, the variables were evaluated for how they were effective in identifying soil moisture. Each variable's neighborhood association was evaluated to understand what key factors appeared to be the best predicting variable types. Neighborhood 1, being clearly grounded in the thermal range is attempting to identify moisture via temperature variation in the soil. Neighborhood 2 appears to be tightly bound directly to detecting moisture while Neighborhood 3 is more aligned with using a vegetation proxy to identify moisture and therefore is associated with plant growth. These neighborhoods of variables are representative proxies for the key components of the Soil-Vegetation-Atmosphere Transfer (SVAT) "triangle" model.

2.2. The Final Variable Set

The final set consists of six variables. Three, NDVI, NMSI, and MNDWI, are pre-defined variables. One, *Variable 2*, is a simple ratio representing the thermal return. The final two, *Variable 10*, and *Variable 16*, are unique to this algorithm. The numbers of the variables are simply the defacto names used while testing the final set of 20 variables.

The final variable set for Landsat 8 is:

Neighborhood 1

Variable 10: (Band8/Band11)/(Band9/Band10)
Variable 2: Band1/Band10

Neighborhood 2

NMSI: (Band7-Band6)/(Band7 + Band6)
MNDWI: (Band3-Band6)/(Band3 + Band6)

Neighborhood 3

Variable 16: (Band4*Band7^3)/(Band5*Band6^2)
NDVI: (Band5-Band4)/(Band5 + Band4)

In the TDR measured subsamples of the entire population, *Variable 10* is poorly correlated with NDVI at −0.53, NMSI at −0.48, and MNDWI at −0.21 but well correlated with L-band soil moisture readings at 0.89. *Variable 2*, which can be described as an image adjusted thermal band is less correlated to soil moisture at 0.73 while the next highest correlation to *Variable 2* is *Variable 10* at 0.34. *Variable 16* is highly inversely correlated to soil moisture, NDVI, and NMSI, at −0.82, −0.89, and −0.92 respectively. Broadening to the entire population of L-band measured samples, Correlations between variables drop considerably.

2.3. Variable 10 & Variable 16

While correlation and the originating bands suggest the basic mode of operation of these two new variables it seems appropriate that a further exploration of their method of action be pursued here. *Variable 10* is relatively poorly correlated with any other variable except *Variable 16*, which is only moderate, but shows considerable ability to discriminate soil moisture while using bands which aren't meant for these purposes. *Variable 16* on the other hand is highly correlated with both NMSI and NDVI but its performance against field tested soil moisture is considerably better than either one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

of those. It appears somewhat similar to a hybrid of NDVI and a cellulose index and may be a total biomass indicator.

Variable 10 is an intriguing index, stated as: (Band8/Band11)/(Band9/Band10). The resolution of this ratio is multi-scale with Band 8 being 15 m, Band 9 is at 30 m, and Bands 10 & 11 both at 100 m. All data is resampled to 30 m. Furthermore, Band 8, ranging from 0.503 μm to 0.676 μm covers a large portion of the visible spectrum from mid-blue through red.

Band 9 is just as interesting as it is shortwave infrared (1.363 μm–1.384 μm) and is usually absorbed by the atmosphere. However, because this algorithm is primarily for use in arid or semi-arid environments and the air is relatively free of moisture.

To further understand this variable, it could be rewritten as: (Band8*Band10)/(Band9*Band11). Both thermal bands have a correlation of 0.997 and could be negated on either side of the ratio, leaving only Band8/Band9. This substitute variable has a 0.994 correlation with the original variable but demonstrates information loss over the original, otherwise the Band8/Band9 version of the variable would have scored higher in original testing and been brought forward over the current *Variable 10*. Therefore, the thermal bands are adding information and returning to the original form of the equation, it appears to be evaluating the slope of the Planck function over the Shortwave to Thermal by the larger Visible to Higher Thermal.

A qualitative view of the variable may be useful in understanding exactly what *Variable 10* is mapping. Initial review shows a confirmation of considerable overlap with NDVI, although the relationship is inverse. To compare accurately, both NDVI and *Variable 10* are mapped over the same area and stretched to a standardized 0–1 range, *Variable 10* is inverted to correspond with NDVI values (see Figure 3). A key difference seems to be that NDVI values lower over natural vegetation in the region whereas *Variable 10* shows no significant decline in these regions. While they look like they are similar superficially, there are considerable differences throughout the image. The correlation with NDVI and thermal cooling appears to be somewhat spurious in that both photosynthesis and evaporative cooling both are common in highly vegetative regions (see Figure 4). Instead we may need to look at a tree algorithm output to understand the actuality of how the variable is being used. It appears from a sample tree that *Variable 10* is a branch end discriminator making fine adjustments to the soil moisture estimation. *Variable 10* seems to be a function of total brightness returned in the visible spectrum, a measure of Ångstom's reflective property of wetting, that property described in remote sensing terms by Wang and Qu in 2009 [16].

2.4. Variable 16

Variable 16 is a new variable that despite its relatively weak correlation with soil moisture, is utilized intensively in nearly all decision trees in determining soil moisture via machine learning. The Landsat band composition of *Variable 16* is (Band4*Band7^3)/(Band5*Band6^2) or Red*SWIR2 cubed/NIR*SWIR1 squared. The use of the Red and near-infrared immediately suggest it is a variant of the NDVI. The inclusion of SWIR indicates it may also account for cellulose and lignin, non-photosynthesizing vegetation as well. A suspicion that is borne out immediately upon a qualitative view of the variable in comparison to NDVI (see Figure 5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

Figure 3. Stretched NDVI (**a**) and a Stretched and inverted *Variable 10* (**b**) and a surface of their difference (**c**) illustrates a contextual difference between the two alluding to *Variable 10*'s use of thermal bands. Their similarity may be bound in the propensity of vegetation being in cooler temperatures due to evaporative cooling.

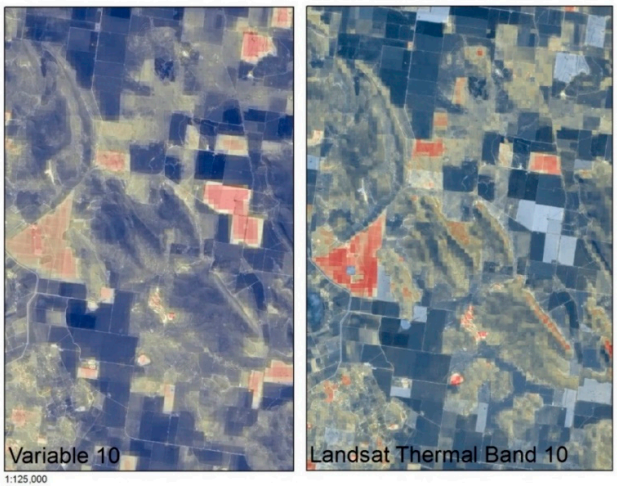


Figure 4. Comparison between *Variable 10* (**a**) and the Thermal band (**b**) shows much similarity in the cool region but also key points of disagreement on the higher temperature side.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

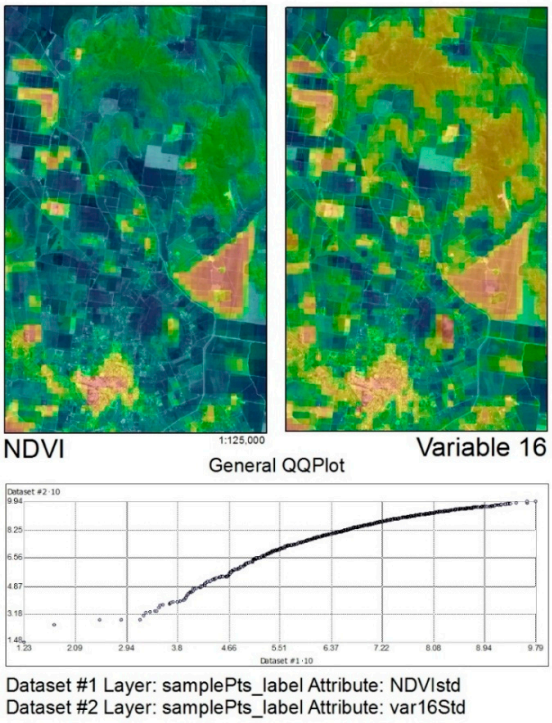


Figure 5. Stretched NDVI comparison (**a**) to a stretched and inverted *Variable 16* (**b**) with QQ plot (**c**) shows the similarity between the two datasets.

The value the algorithm seems to have with a modified NDVI is that the feature space increases in dimensionality permitting better differentiation between otherwise very similar data points. NDVI appears to be highly valued in this soil moisture algorithm and the use of this alternative metric to describe it permits the random forest to better allocate values to lower moisture points.

2.5. The Learning Algorithm

Now that the variables are identified, the obvious issue is the dependent variable since ample ground truth data is rarely available for the date, time, and location required. Here the SMAP values (Figure 6) can provide a local training set, the independent variable to the machine learning algorithm.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

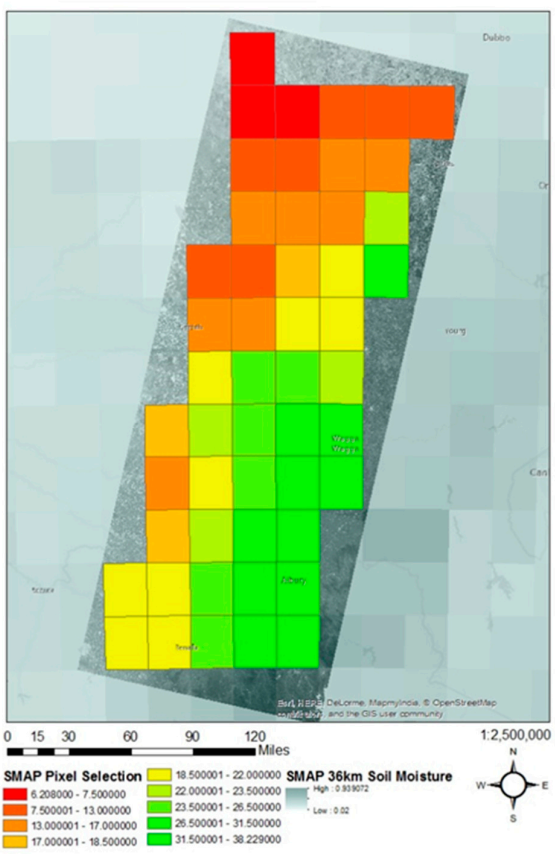


Figure 6. Selected SMAP pixels over the three joined Landsat scenes used to train the algorithm.

While a multilayer perceptron may be appropriate to provide an ongoing predictive application, the researcher should continue beyond such an implementation to discover the causal relationships driving those connections. Some models provide a considerably more intuitive construct from which to begin constructing further research understanding the underlying principles. A more intuitive model that performs often as well, if not better than an artificial neural network (ANN) such as the multilayer perceptron, is the tree regression model. These models are newer then the artificial neural network and unlike the black box response seen from ANN, the method of decomposition is easily understood in the model formulation [29]. The tree regression model is clearly read and denotes by its very structure, the import of variables and their inherent weights. An exploration of variable importance down the length of the tree, exploring possible proxy and spurious relationships should foster a movement from vague approximation of the import of variables to the full numerical model describing the system (i.e., data to knowledge).

The model tree iteratively splits a training set of variables in order to minimize regression error until the deviation is only a small fraction of a standard deviation of the original instance, at which point the splitting process ceases and the model is pruned back one node. At the new tree branch termination, 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 [30]. However, a fault often seen in model trees is the tendency to over fit. 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 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 [28]. This averaging process

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

effectively controls model variance, error due to instability of the model and limited training data, without increasing overall bias [31]. In this lies the strength of the random forest algorithm. Noted limitations of random forests are the inability of regression prediction significantly beyond the range of the training data. In this usage we might expect some results in the tails of the distribution to be slightly exaggerated due to prediction beyond the training range.

This algorithm (see Figure 7) 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 rule set, in this case using a random forest due to its superior performance in local testing. However, regression, an ANN, or other inductive method may be applied as well. This is an agnostic step and the choice of random forest was purely based on performance results and the afore mentioned benefits to avoid over fitting. Random Forest are excellent at classification tasks (see Figure 8) which essentially is what downscaling is, due to the sample being in the same space and time as the requested output however its continuous nature is described in terms of a regression. Doing so allows the user to identify the specific rule set, in this case a set of decision trees, applicable to the region of inquiry. Those rules are then utilized at the Landsat native resolution of 30 m to provide the down-sampled soil moisture surface. The rules are variable across regions and not transferable across Landsat scene sets due to changes 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 its previous and following scene in a single path. However, the training data provides a strong localized approximation of the local soil moisture condition.

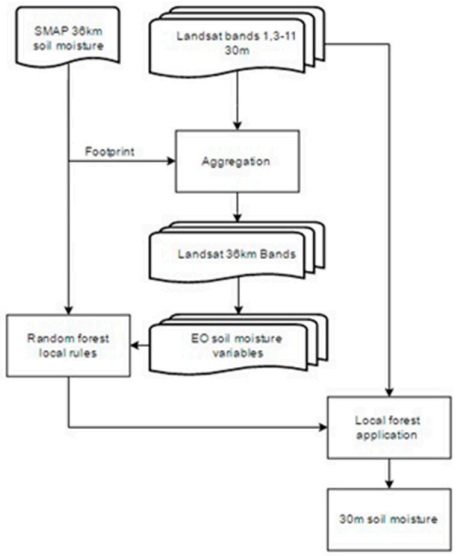


Figure 7. Process flowchart of localized downscaling process via inferred learning of soil moisture.

A total of four datasets were utilized in the training, testing, and validation of this method. The SMAP radiometer data was used as the source of soil moisture at low resolution; it is the source we wish to downscale. Landsat 8 imagery was acquired the same day and is the source of the variables in which the SMAP data is to be downscaled. The TDR, or capacitance soil monitor over Australia, and aircraft acquired PLMR are the validation ground truth measurements used to compare our results to high fidelity external measurements (see Figure 9).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

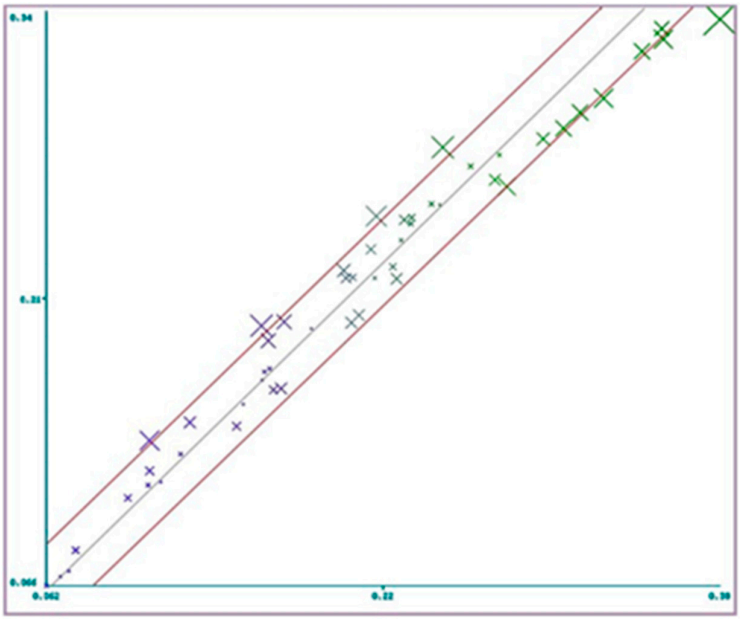
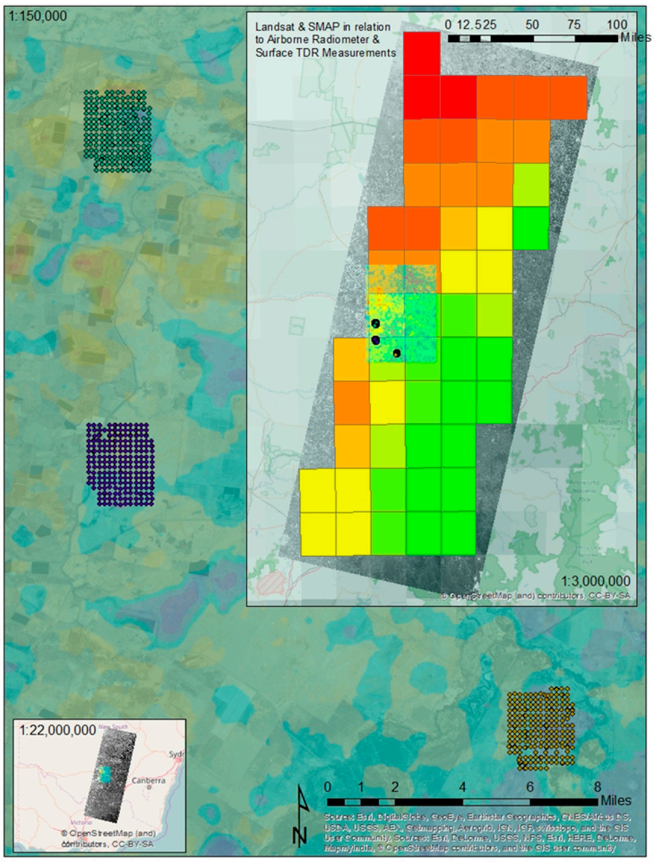


Figure 8. Scatter plot of training effectiveness using the SMAP values as the independent variable and the down sampled Landsat imagery input into the aforementioned variables as dependent variables, the red lines demarcate a 0.04 variance off of a perfect prediction.

The final output of the algorithm can be seen in Figure 10. Noticeable is the wetting of agricultural fields while natural woodlands and shrub regions remain dryer. This is likely due to the reliance on irrigation in the region but analysis against the PLMR and ground truth point soil moisture datasets will provide accuracy measurements.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

Figure 9. 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) 14 August 2015. The north and southeast points were included in original training set and are excluded from testing.

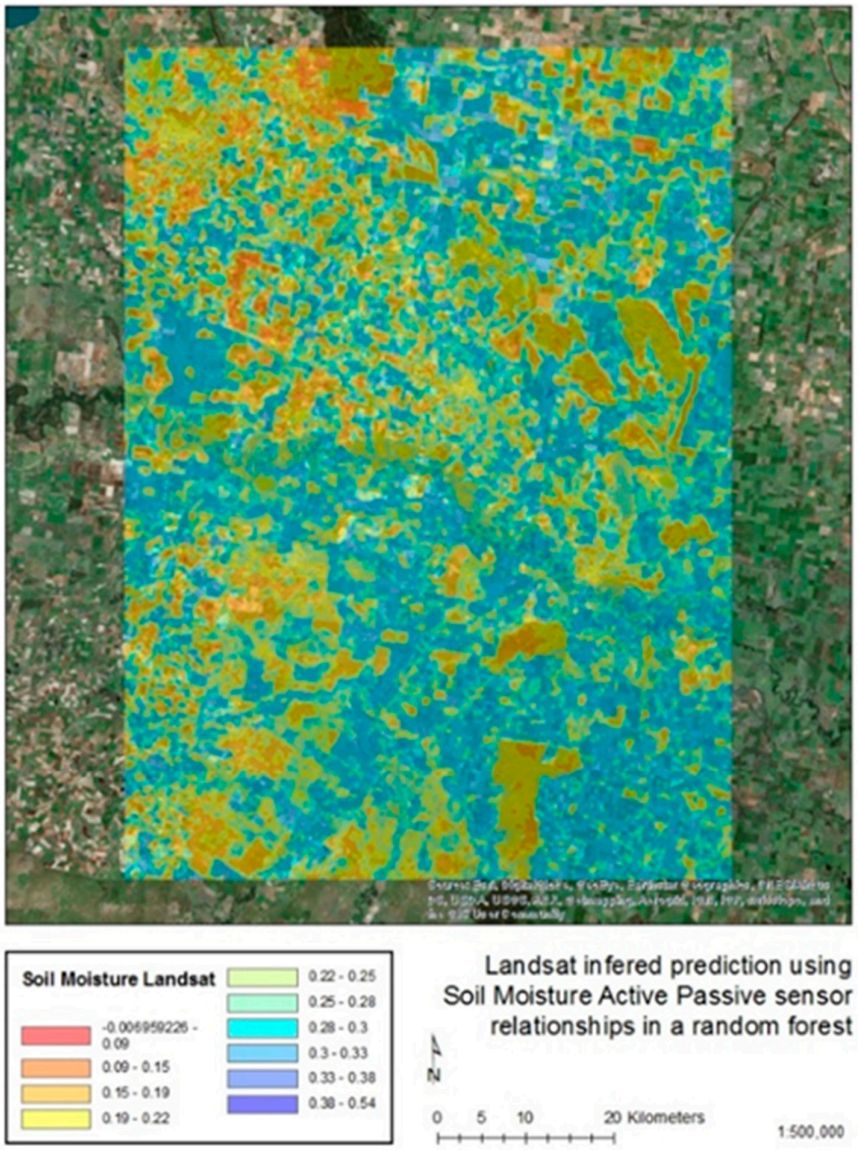


Figure 10. The soil moisture output from the localized soil moisture downscaling process 14 August 2015.

3. Results

Ground truth over the area was available via an excellent dataset of frequency domain reflectometry; however, the coverage of this dataset was minimal compared to the entire range of the tested region. An airborne L-band radiometer (PLMR) also took readings on the same day and was the primary source used to validate the localized downscaling method. Utilizing the PLMR microwave radiometer flown over the site required conversion of the radiometric brightness to the dielectric constant. This was be done via the Fresnel reflection equations [32] using the formulas:

Equation (6) Horizontal Fresnel Reflection Equation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

$$R_0^H = \left| \frac{cos\theta - (\varepsilon_r - sin^2\theta)^{\frac{1}{2}}}{cos\theta + (\varepsilon_r - sin^2\theta)^{\frac{1}{2}}} \right|^2 ; \quad (7)$$

Equation (7) Vertical Fresnel Reflection Equation

$$R_0^V = \left| \frac{\varepsilon_r cos\theta - (\varepsilon_r - \sin^2\theta)^{\frac{1}{2}}}{\varepsilon_r cos\theta + (\varepsilon_r - \sin^2\theta)^{\frac{1}{2}}} \right|^2 \quad (8)$$

However, because a relationship is known to exist via the microwave polarization difference index between the polarizations (H and V), and NDVI via the Microwave Polarization Index:

Equation (8) Microwave Polarization Difference Index

$$\text{MPDI} = (\text{TV} - \text{TH})/(\text{TV} + \text{TH}) \quad (9)$$

Equation (9) NDVI MPDI relation

$$\text{NDVI} = E_0 + E_1 * \exp(E_2 * \text{MPDI}) \quad (10)$$

when combined with surface temperature, and all data can be reported to the same resolution negating the need for rescaling, soil moisture can easily be identified using reported values in localized areas for training via the function:

Equation (10) soil moisture function

$$f(SM) = = (\text{TV}, \text{TH}, \text{TS}, \text{NDVI}) \quad (11)$$

To find land surface temperature, the DN value of Band 10 must be converted to spectral radiance. Band 10 is used over band 11 due to an ongoing investigation into stray light boosting reported temperatures from band 11 by up to 8 kelvin, while band 10 has reported an increase of half that value or less.

Equation (11) Spectral radiance

$$L_\lambda = M_L * Q_{cal} + A_L \quad (12)$$

where:

L_λ = Spectral radiance (W/(m2*sr*μm))

M_L = Radiance multiplicative scaling factor for the band

A_L = Radiance additive scaling factor for the band

Q_{cal} = L1 pixel value in DN. (USGS 2016)

Equation (12) Spectral radiance

$$L\lambda = 3.342 \times 10^{-04} * Q_{cal} + 0.10000 \quad (13)$$

This can then be transferred to the Top of Atmosphere Brightness Temperature via the formula:

Equation (13) Top of Atmosphere Brightness Temperature

$$T = \frac{K2}{\ln(\frac{K1}{L\lambda} + 1)} \quad (14)$$

where:

T = TOA Brightness Temperature, in Kelvin.

L_λ = Spectral radiance (Watts/(m2*sr*μm))

K_1 = Thermal conversion constant for the band

K_2 = Thermal conversion constant for the band

Land surface temperature can be determined via the single window algorithm via the formula

Equation (14) Land surface temperature

$$T/1 + w*(T/p)*\ln(e) \quad (15)$$

where:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

- T = Top of atmosphere brightness for Band 10
- w = wavelength of emitted radiance (11.5 μm)
- p = h*c/s (1.438 × 10^{-2} mK)
- h = Plank's constant (6.626 × 10^{-34} Js)
- s = Boltzmann constant (1.38 × 10^{-23} J/K)
- c = velocity of light 2.998 × 10^{8} m/s)
- e = land surface emissivity

and where:

Equation (15) Proportion of vegetation

$$P_v = (NDVI - NDVImin/NDVImax-NDVImin)^2 \ \& \ e = 0.004P_v + 0.986 \quad (16)$$

When combined with NDVI, all variables are now available that are a function of soil moisture, $f(SM) = (T_V, T_H, T_S, \text{NDVI})$. Using the measured values in the high-resolution grids, the site-specific soil moisture equation should be easily determinable from the polarized microwave radiometer datasets.

Evaluating for the f-channel radiometer was conducted via an empirical Bayesian kriging as it provides a better fit to the data over standard kriging methods with a mean error of 0.03, a Root Mean Square (RMS) of 3.4, and an average standard error of 3.6532. Initial analysis of the radiometer data shows considerable banding in the direction of the flight path, while a filter function may be appropriate, the first step is to standardize all the data from the various beams (Figure 11).

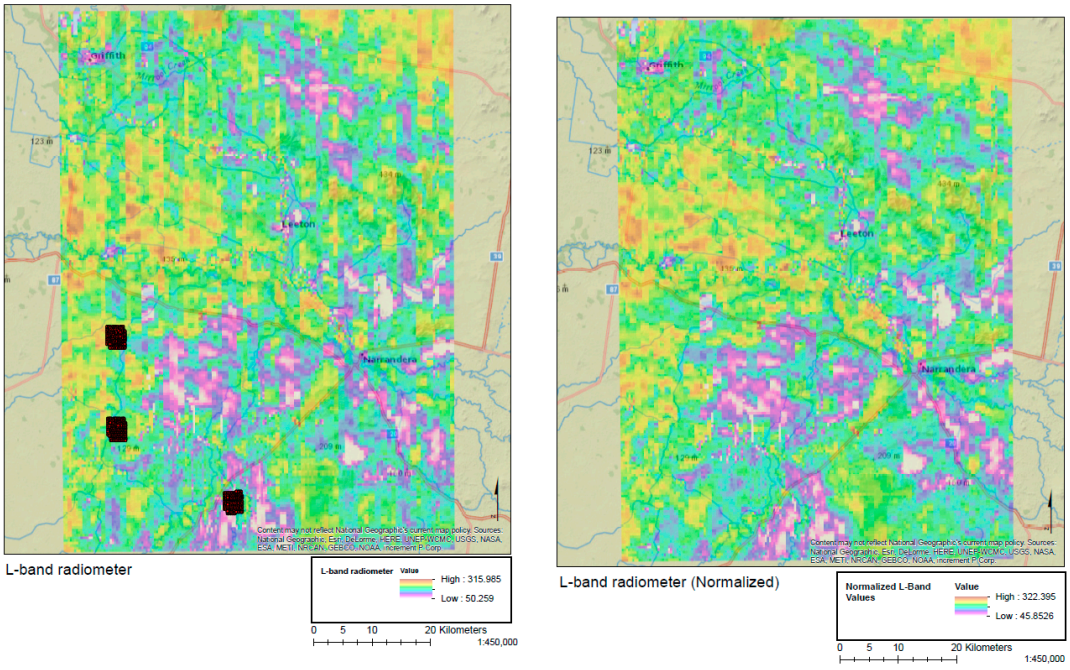


Figure 11. Initial radiometer data (**a**) showed a strong banding artifact in the direction of the flight path for the 14 August 2015 data. Normalizing the L-band return adjusts the variation in the individual beams along flight path (**b**).

As the beams cover a relatively equal area and are over a wide enough region, the assumption can be made that all six should be from the same population and share the same distribution of values. Therefore, all six individual beams are normalized to the population of the entire dataset (Table 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

Table 1. Mean and Standard Deviation of each individual flight path and the full population.

	Beam 1	Beam 2	Beam 3	Beam 4	Beam 5	Beam 6	All
Mean	231.9140	236.8200	245.7915	242.4373	241.3469	236.4652	239.1292
Std. Deviation	27.0966	27.1964	25.9045	27.8285	27.9406	26.4818	27.4653

Normalizing each band to that of the total population reduces the mean error of prediction in the kriged surface to 0.027 and the Average Standard Error to 3.48. The final output of the normalized L-band data shows the majority of banded effects removed, which can be used to determine volumetric water content in concert with the vegetation optical depth via the NDVI, and ground surface temperature.

Application of Land Surface Temperature, NDVI, and L-band radiometric brightness temperature was averaged over an area of 0.5 km pixels and trained against the aggregation of several hundred capacitance measured soil moisture values in the same resolution. The training and application algorithm utilized was a random forest. The ground truth metric upon which evaluation of the Inferred Learning Soil Moisture Algorithm (ILSMA) will be measured is the normalized L-band which derived values from the aggregated point soil moisture measurements. The radiometer derived soil moisture had a correlation with the radiometer raw return values of 0.9674 with a mean absolute error of 0.0207. Variation is due primarily to input of ground thermal data and vegetation density. Numerical values were then transferred to a surface via an empirical Bayesian krig with an error of 1.6×10^{-5} (Figure 12).

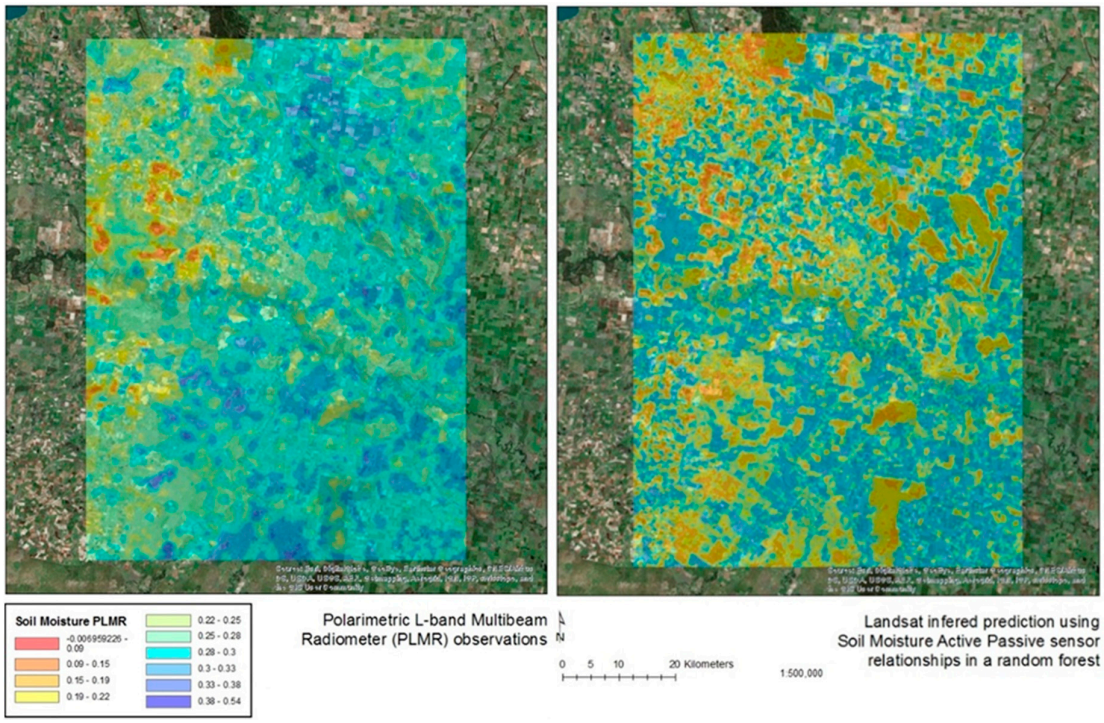


Figure 12. A visual comparison of the L-band radiometer (**a**) and the Landsat inferred soil moisture (**b**) shows good agreement overall with areas of notable exception, such as the southern edge, just west of center.

Overall the predictive power of the downscaled soil moisture did quite well, although just outside the NASA standard of 0.04 for the SMAP validation program. The full population of 3097 pixels by 2257 pixels, for a total of 6,989,929 samples with an absolute error of 0.054 over the L-band retrieved surface (see Table 2, Figure 12).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

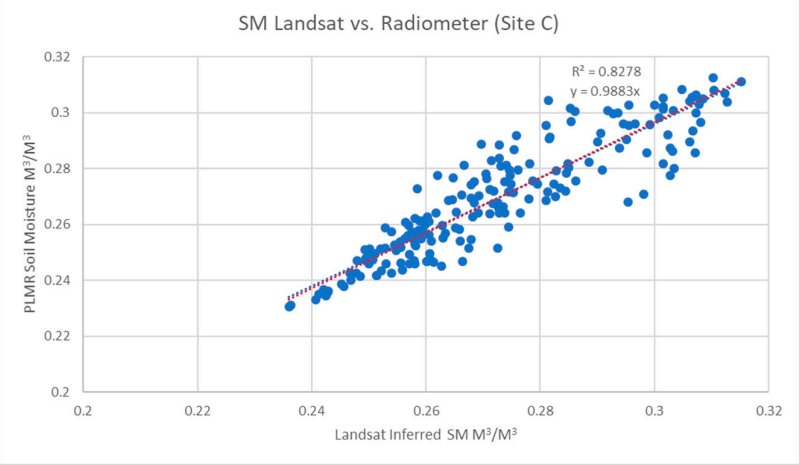
Table 2. The mean of the proposed algorithm is 0.26 while the mean of the L-band retrieval is 0.28; the mean absolute error (MAE) is 0.05.

	ILSMA	L-Band	Error	Abs. Error
Mean	0.2586	0.2756	0.017	0.0535
Std. Dev	0.0662	0.0285	0.0628	0.037
Skew	-0.4595	-1.4099	0.3754	0.7347

When comparing the Landsat solution back to the original SMAP data, an admittedly imperfect comparison due to the lack of complete coverage of the Landsat inferred solution of the totality of the SMAP pixels involved, we see that variance does increase with an increase of observation resolution. This provides some support to the self-affine fractal distribution of soil moisture. The Landsat solution, as shown in Table 2, has a standard deviation of 0.066 while the six SMAP overlapping the Landsat investigation region has a standard deviation of 0.038. As the standard deviation is the square root of the variance we can place the variance of the Landsat solution at 0.0044 and the variance of the SMAP distribution at 0.0015. The total mean of the Landsat solution is quite high compared to the SMAP mean of 0.21 but 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 skew for both datasets remain remarkably similar however, with the SMAP skewness at −0.59 and the Landsat solution at -.048 while kurtosis is −1.68 and −1.25 respectively.

While qualitatively the two surfaces, PLMR and Landsat inferred method, provide similar solutions to volumetric soil moisture a more quantitative review is required. Three high density soil moisture FDR point field 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 naïve dataset. This was likely not a necessary precaution given the wide range of other datasets included but it is generally a good practice to always test on truly independent data that has not influenced testing in any way. These sub-sites, sites C (central), N (north), and SE (southeast), provided a good arrangement of diverse environments for testing. We evaluated the inferred solution against the radiometer as well as 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 soil moisture versus the radiometer measurement from the PLMR shows a very good agreement of the field site C with an R^2 of 0.8278 (see Figure 14). This is in stark contrast with the point data that has a much lower agreement with an R^2 of 0.2985 however, given the R^2 of the PLMR versus the soil moisture in the sample field of 0.2026, this seems a reasonable, albeit low, value (see Figures 15, 16).



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

Figure 14. Scatterplot of Landsat inferred soil moisture versus the soil moisture estimates from the PLMR at Site C.

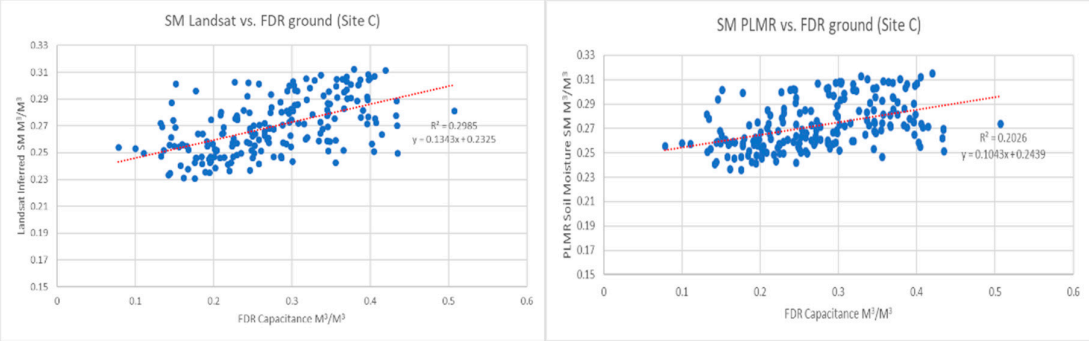


Figure 15. Landsat inferred soil moisture plotted against the point measurement data (**a**) showing a low R^2 value at site C. The Landsat inferred data does have a slight trend associated but the variability in the data is significantly lower than the FDR soil moisture values for the sample field leading to a lowered R^2 score. The PLMR estimated soil moisture (**b**) also shows a poor R^2 score at site C when compared to the point retrieved soil moisture data. Again, the variability of the PLMR data is much lower than the field collected data.

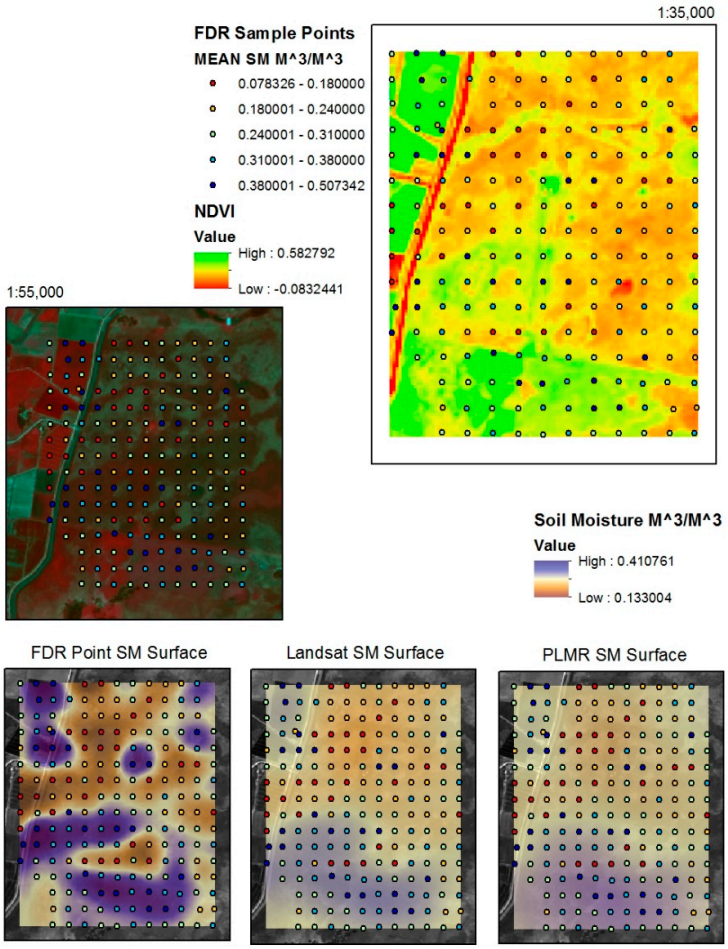


Figure 16. Overview of validation field C (SMAPex site Y7) with NDVI on the upper right with overlain field collection points. On the center left is a false color image showing the vegetated field on the west side of the road while the east is primarily ranchland. The lower three images are kriged soil

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

moisture surfaces from point data (left), Landsat inferred method (center), and PLMR soil moisture data (right).

It is our suspicion that vegetation may be interfering with some of the soil moisture observation and that some additional error may be introduced by the difficulty in describing soil moisture using point data. To test the effect of vegetation on the error between the field collected soil moisture and the Landsat inferred values we ordered each point in ascending order by NDVI and calculated an average error over five variables creating a moving window. A scatterplot of the averaged error against NDVI does appear to confirm our suspicion that as NDVI increases, so too does the error between field collected data and the Landsat inferred soil moisture. The linear R2 of 0.2934 is not proof but does suggest a trend exists. Even more interesting is a polynomial trend line with an R2 of 0.3921 in which error decreases above an apogee around an NDVI value of 0.46 (see Figure 17). Additional data sets will need to be reviewed to understand if this is just an overfit line or truly a pattern in the data.

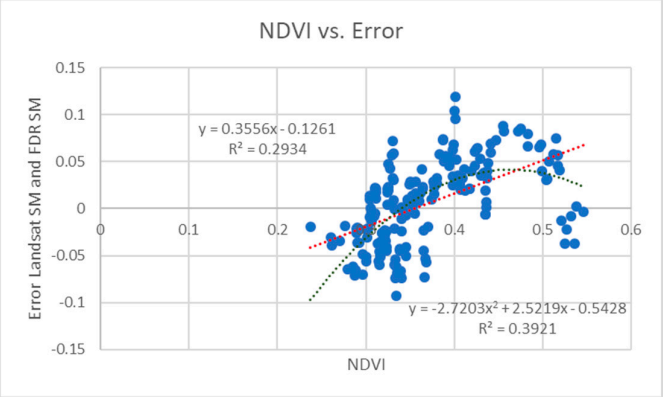


Figure 17. NDVI compared to a moving average error of soil moisture measured between the Landsat inferred method and the field collected capacitance measurements at site C. Upper left equation is for the linear solution while the lower right is for the polynomial solution.

The two other sites can be evaluated by comparing the PLMR calculated soil moisture and the Landsat inferred solution. Again, point collected data is for information purposes only, although the data was trained against the SMAP radiometer data, the N site and SE site were part of the dataset used for variable identification and there is a possible over-reporting of performance due to an unusually high ability of the variables to parse those particular sites. The variable testing may also not have any affect and we may see a similar or lower R^2 score over those areas. The uncertainty though cautions me against the use of such a region for evaluation purposes.

Site N (see Figure 18) showed a similar result to site C although all measures were marginally better. The Landsat inferred solution had an R^2 of 0.933 with the radiometer data (see Figure 19) and the informational R^2 of both the Landsat inferred and PLMR data against the point values also improved (see Figure 20). This is most likely due to the cultivated landforms being uniform and controlled monocultures without considerable variation, this provides a bit more uniformity to the signal over a larger area.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

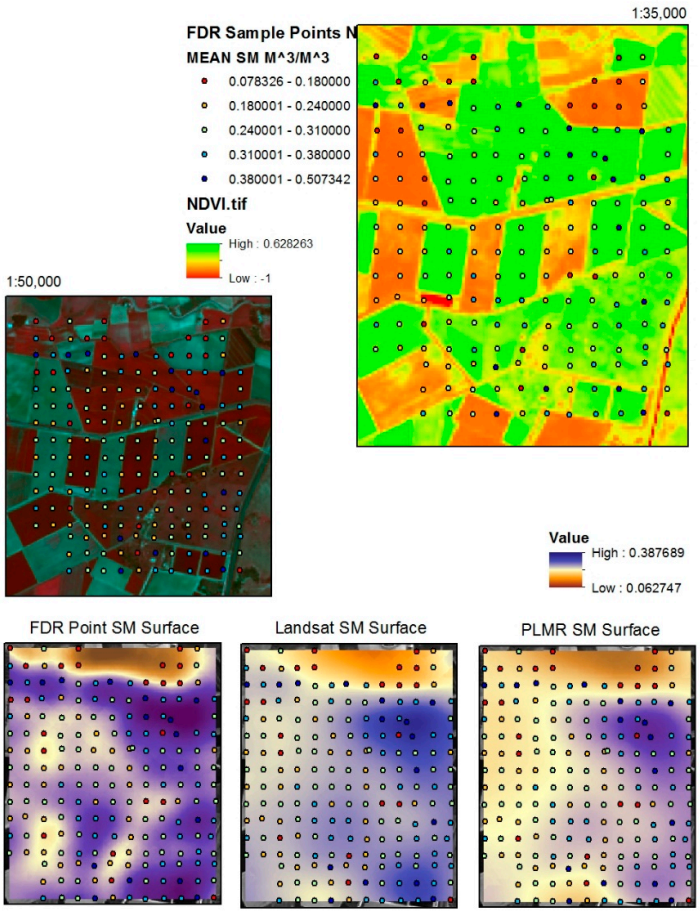


Figure 18. Site N (SMAPex site Y4) is dominated by irrigated agricultural cropland. Both PLMR and the Landsat methodology seemed to underestimate variability within the site. This may be use to crustal drying with moisture below the soil O horizon where the capacitance tines measure the higher moisture values.

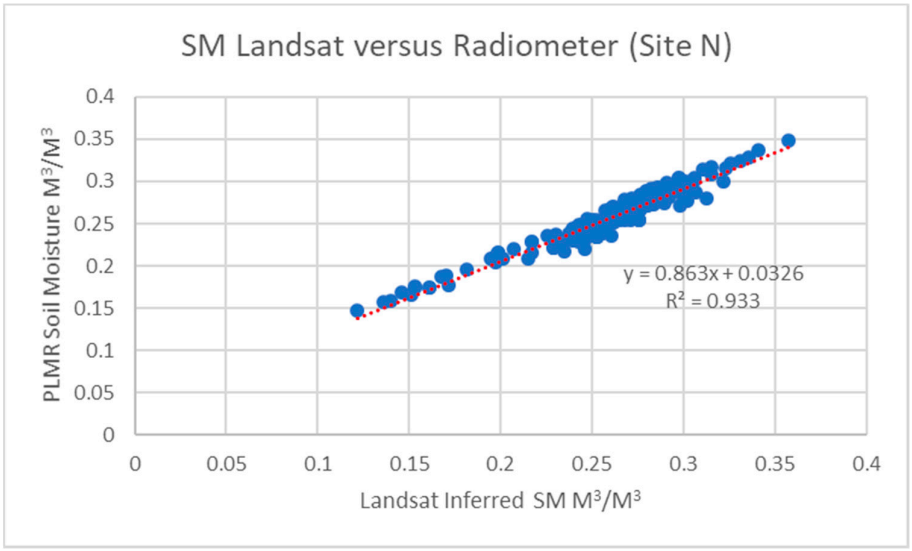


Figure 19. The R^2 of 0.933 between the radiometer and the Landsat inferred solution is quite remarkable with only an evident slight dry shift in lower Landsat inferred values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

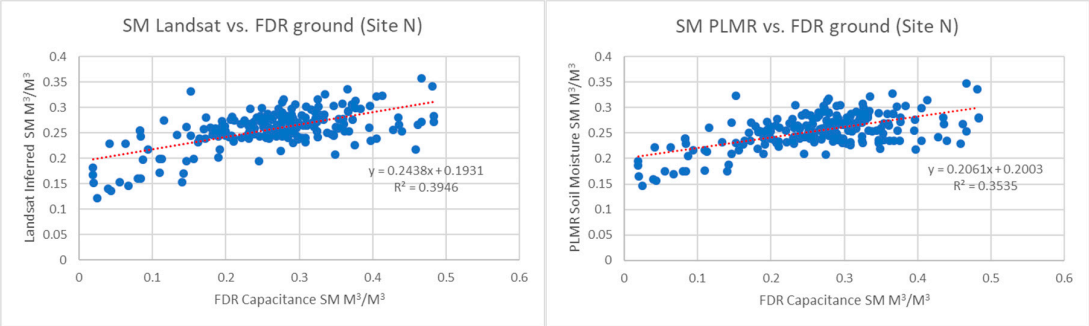


Figure 20. The informational Landsat to point data (**a**) does show an improvement in performance, this may be due to a uniform cultivated surface or may be a legacy of the variable training process, however unlikely. The higher radiometer R^2 (**b**) also suggest the higher R^2 scores are not an artifact of the variable training process but are linked to the surface uniformity.

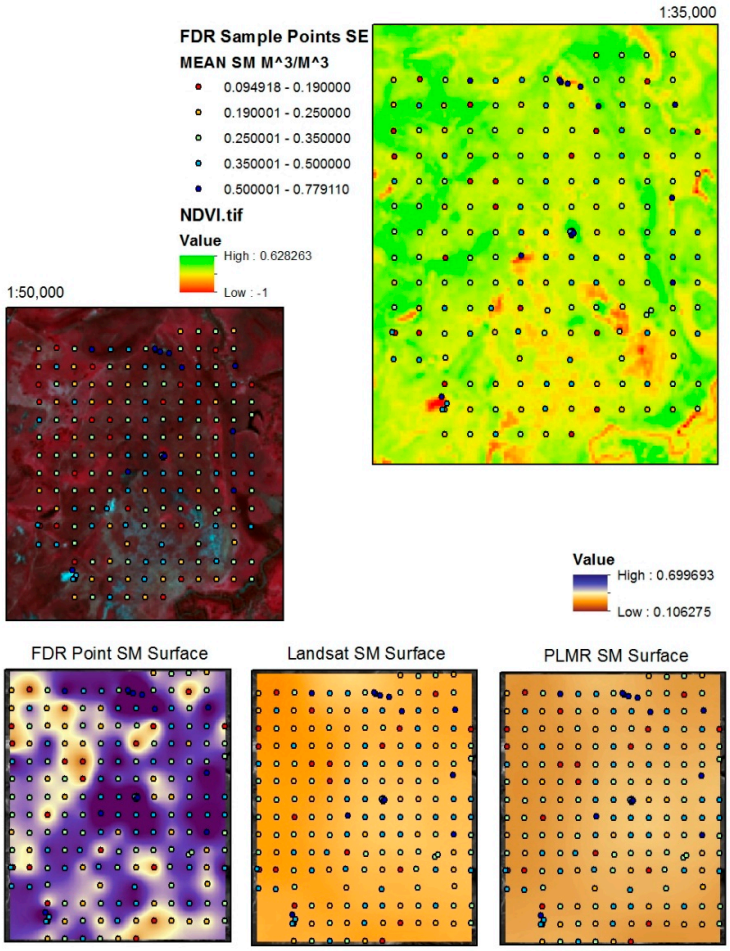


Figure 21. Site SE (SMAPex site Y10) is composed exclusively of uncultivated ranchland with poor drainage and isolated tree stands. This grassland site is prone to standing shallow pools of water.

While both the PLMR and the Landsat inferred solution were in good agreement at site SE (see Figure 21), neither was well in agreement with the ground data points with the Landsat inferred reporting an R^2 of 0.1394 and the PLMR an R^2 of 0.1144 (see Figures 22, 23). This ranchland site was very wet with many small pools of standing water in dense grass alternating with dry sparse grass regions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

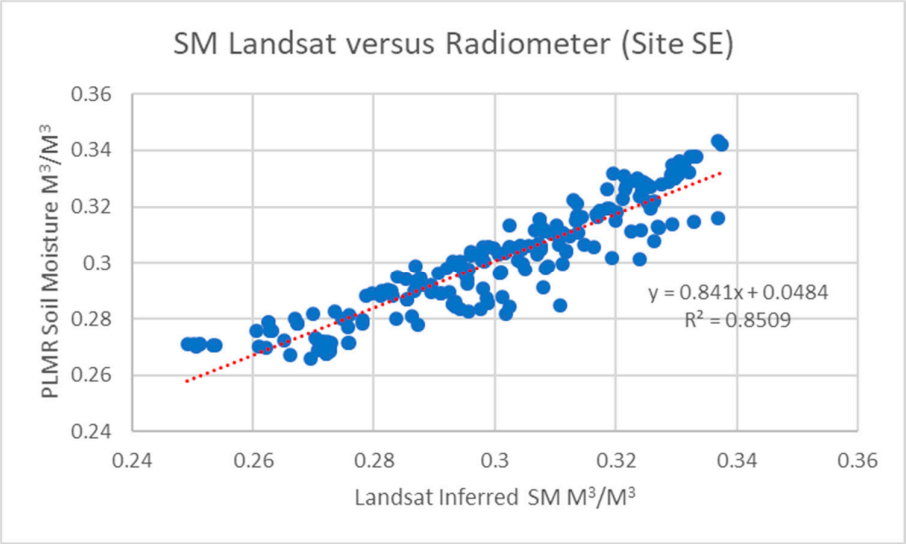


Figure 22. The Landsat inferred solution well described the PLMR solution for the soil moisture surface over the flat grasslands of site SE.

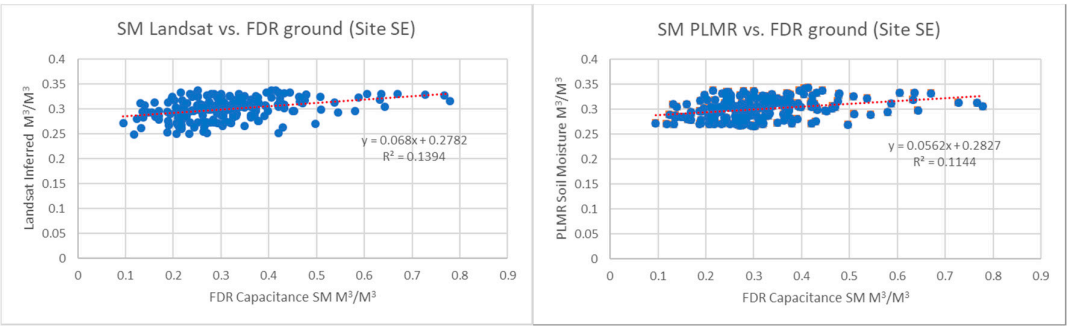


Figure 23. Neither the (**a**) Landsat nor the (**b**) PLMR solution poorly described the variation obtained by the FDR ground measurements.

Finally, I'd like to evaluate the full collection from all three sites. The PLMR and the Landsat inferred solution were remarkably similar to each other across all sites (see Figure 24). The Landsat inferred solution did a slightly better job of predicting the soil moisture readings of the capacitance ground readings across all sites in comparison to the PLMR solution (see Figure 25). The following two figures are to be used for informational purposes only as part of the data contained were previously used for variable selection and may influence the results in favor of the proposed solution.

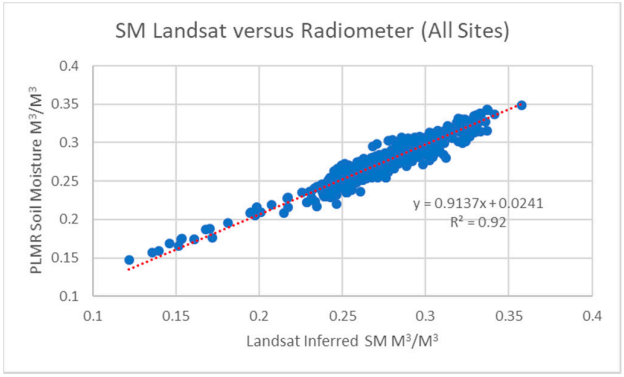


Figure 24. All three sites evaluated for the R^2 of the Landsat inferred solution versus the PLMR observed soil moisture surface shows considerable agreement.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

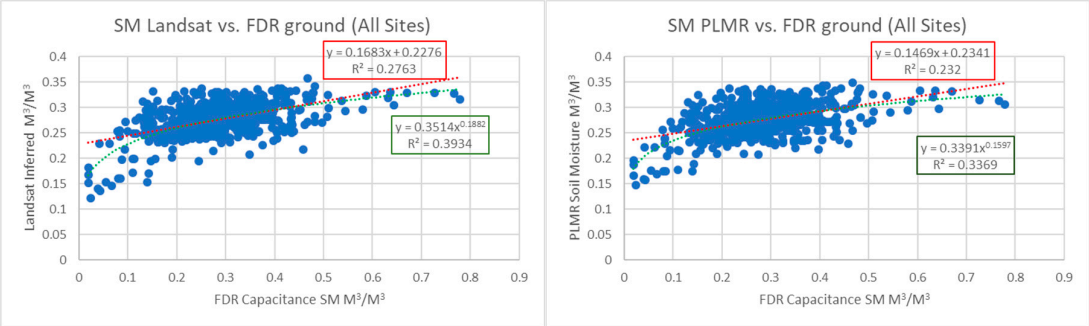


Figure 25. The Landsat inferred solution had only moderate success (red) in terms of a linear R^2 score against the ground readings however a simple power transformation (green) would improve the R^2 score to a 0.39, still only moderately successful but given the uncertainty of measuring moisture with point data, not an unreasonable solution (**a**). The PLMR solution (**b**) performed slightly worse across all sites (red) and a power transformation (green) could improve the overall performance to a 0.3396 R^2.

Returning to the comparison of the PLMR data comparison to the Landsat inferred solution over all three sites it is necessary to quantify the similarity or dissimilarity of the datasets. A statistical similarity of variance should signify a significant similarity in datasets and therefore one dataset well describes the alternate. Microwave L-band soil moisture is a well-accepted methodology of mapping moisture. If the Landsat inferred methodology can indeed well describe the PLMR data then by transference, it would well describe soil moisture. In this unusual instance, the standard null hypothesis case of no statistical variance is the alternative as a statistical similarity would indicate the Landsat inferred model is well modeling the same factors as the PLMR. The Analysis of Variance (ANOVA) is a useful tool for analyzing such variance differences. However, ANOVA can only disprove similarity; it cannot prove equivalence. the single factor ANOVA of the entirety of the sample site data between the PLMR and Landsat inferred data. Because the F critical, 3.849, is much larger than the F, 0.0329, we cannot reject the statistical null hypothesis of variance similarity. We cannot accept the statistical null hypothesis either at this point. An equivalence test such as two-one sided t-test (TOST) can provide a valid method for such problems.

To validate that the population from the Landsat inferred method is as valid as the L-band method we will be conducting a metric of equivalence test, the two-one sided t-test, or 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 evaluation [33] of the PLMR data and the Landsat inferred data provided a significant result to $p = 0$ that both datasets were equivalent albeit as samples devoid of their geographic nature.

4. Discussion

Applying the previously described methodology has shown considerable fitness in describing soil moisture distribution at significantly higher resolution than currently available using SMAP alone. The method utilizes the refinement of the SMAP algorithm, ingraining in the results the description of the soil moisture relationship to the spectral conditions on the ground. While the method did not prove an accuracy within the 0.04 range as required by the NASA SMAP mission in this experiment, it may be possible 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 1.2 inches beyond the waveguides [34]. 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 have always been accurate in describing the true value of the pixel and instead fell within the tail of the distribution of soil moisture values within a pixel. This can be demonstrated by evaluating the number of point moisture sample points within a 30 m Landsat

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 February 2019 doi:10.20944/preprints201902.0046.v1

pixel and the error of that averaged point moisture value with the downscaled Landsat inferred algorithm results (see Figure 26).

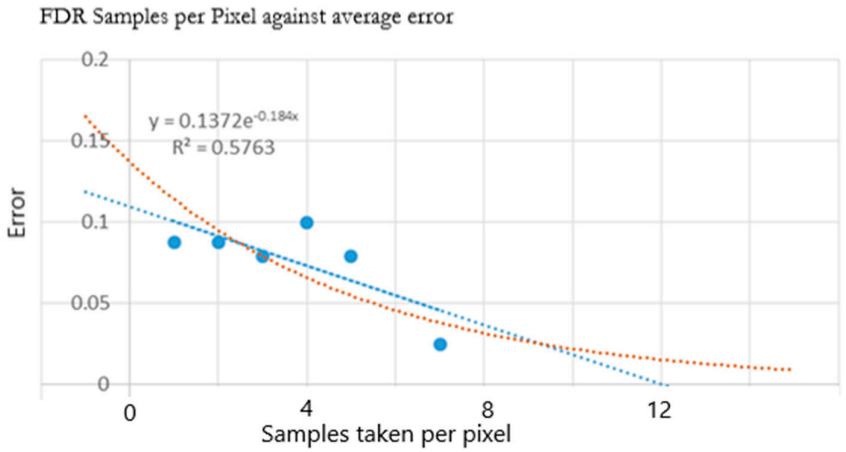


Figure 26. The number of point soil moisture samples on the x-axis compared with the error from the Landsat inferred soil moisture value suggests that point soil moisture values are prone to high error in terms of the average value of the pixel when less than nine samples are obtained over the 900 m² pixel area.

While the sheer number of observations aid the alleviation of some of the training errors, a large number of them can affect the training algorithm that allocated the transition variables from raw L-band return to soil moisture content. Among the three sampling regions of soil moisture point collection, two were used for training the L-band radiometer in transitioning return to soil moisture content, the 3rd was used for testing. That region (site C) experienced a 0.0628 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 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 a local resolution. The root mean standard error, RMSE, for the ILSMA to FDR measured values is 0.105 over the entirety of the scene. Given the heterogeneity of the scene though, the target evaluation sites have been isolated in Table 6 for MAE and RMSE.

Table 6. Mean Absolute Error (MAE) and RMSE over selected test sites (C, N, SE). RF is the Random Forest of the Landsat inferred method while FDR is the frequency domain site sampling.

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

n=197

Site C	MAE	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

5. Conclusions

The SMAPex-5 campaign [35] was an excellent testbed to attempt a validation of this methodology. While obvious weakness exists in that visual remote sensing cannot obtain imagery

through cloud cover as opposed to microwave, nor is this algorithm expected to perform well in heavily vegetated regions, this 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.

Author Contributions: Conceptualization, M.L., A.F., P.H.; Methodology, M.L., A.F., C.S., P.H., J.Q.; Software, M.L..; Validation, P.H., J.Q., A.F, C.S,; Formal Analysis, M.L.; Investigation, M.L., A.F., C.S.; Resources, A.F., P.H., C.S.; Data Curation, M.L.; Writing-Original Draft Preparation, M.L.; Writing-Review & Editing, A.F., C.S., P.H., J.Q., M.L.; Visualization, M.L.; Supervision, P.H., A.F.; Project Administration, M.L, A.F., P.H.; Funding Acquisition, A.F.

Funding: This study was conducted for the Engineer Research and Development Center's Center-Directed Research Program under "Army Terrestrial Environmental Modeling and Intelligence System (ARTEMIS)." The technical monitor was Dr. John Eylander, U.S. Army Engineer Research and Development Center, Cold Regions Research and Engineering Laboratory (ERDC-CRREL). The development of this report was led by the Data Signature and Analysis Branch (DSAB) of the ERDC-Geospatial Research Laboratory (GRL).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cosby, B.J.; Hornberger, G.M.; Clapp, R.B.; Ginn, T.R. A Statistical Exploration of the Relationship of Soil Moisture Characteristics to the Physical Properties of Soils. *Water Resour. Res.* **1984**, 682–690, doi:10.1029/WR020i006p00682.
2. Weizu, G.; Freer, J. Patterns of Surface and Subsurface Runoff Generation. *Tracer Technol. Hydrol. Syst.* **1995**, *229*, 265–273.
3. Denmead, O.T.; Shaw, R.H. Availability of Soil Moisture to Plants as Affected by Soil Moisture Content and Meteorological Conditions. *Agron. J.* **1962**, *54*, 385–390.
4. Koster, R.D.; Dirmeyer, P.A.; Guo, Z.; Bonan, G.; Chan, E.; Cox, P.; Gordon, C.T.; Kanae, S.; Kowalczyk, E.; Lawrence, D.; et al. Regions of strong coupling between soil moisture and precipitation. *Science* **2004**, *305*, 1138–1140, doi:10.1126/science.1100217.
5. De Michele, C.; Salvadori, G. On the derived flood frequency distribution: Analytical formulation and the influence of antecedent soil moisture conditions. *J. Hydrol.* **2002**, *262*, 245–258.
6. Pelletier, J.D.; Malamud, B.D.; Blodgett, T.; Turcotte, D.L. Scale-invariance of soil moisture variability and its implications for the frequency-size distribution of landslides. *Eng. Geol.* **1997**, *48*, 255–268.
7. Miras-Avalos, J.M.; Trigo-Corboba, E.; da Silva-Dias, R.; Varela-Vila, I.; Garcia-Tomillo, A. Multifractal behaviour of the soil water content of a vineyard in northwest Spain during two growing seasons. *Nonlinear Process. Geophys.* **2016**, *23*, 205–213.
8. Bryant, R.; Thoma, D.; Moran, S.; Holifield, C.; Goodrich, D.; Keefer, T.; Paige, G.; Williams, D.; Skirvin, S. *Evaluation of Hyperspectral, Infrared Temperature and RADAR Measurements for Monitoring Surface Soil Moisture*; United States Department of Agriculture: Tucson, AZ, USA, 2003.
9. Ångstom, A. The albedo of various surfaces of ground. *Geografiska Annaler* **1925**, *7*, 323–342.
10. Planet, W.G. Some Comments on Reflectance Measurements of Wet Soil. *Remote Sens. Environ.* **1970**, *1*, 127–129.
11. Grasser, E.A.; Van Bavel, C.H.M; The effect of soil moisture upon soil albido. Agricultural Meteorology **1982**, 27:17-26.
12. Twomey, S.A.; Bohren, C.F.; Mergenthaler, J.L.; Reflectance and albedo differences between wet and dry surfaces. *Appl. Opt.* **1986**, *25*, 431–437.
13. Philpot, W.; Spectral Reflectance of Wetted Soil; Cornell University: Ithica, NY, USA, 2011.
14. Musick, H.B.; Pelletier, R.E., Response to soil moisture of spectral indices derived from bidirectional reflectance in Thematic Mapper wavebands. **1988**; Remote Sens. Environ. 25:167-184.
15. Muller, E.; Decamps, H. Modeling soil moisture-reflectance. *Remote Sens. Environ.* **2000**, *76*, 173–180.
16. Wang, L.; Qu, J. *Multiband Drought Index Enhances Soil and Vegetation Moisture Monitoring*; SPIE: Washington, DC, USA, 2009.

17. Lobell, D.B.; Asner, G.P. Moisture Effects on Soil Reflectance. *Soil Soc. Am. J.* **2002**, *66*, 722–727.
18. Liu, W.; Baret, F.; Gu, X.; Tong, Q.; Zheng, L.; Zhang, B. Relating soil surface moisture to reflectance. *Remote Sens. Environ.* **2002**, *81*, 238–246.
19. Whiting, M.L.; Li, L.; Ustin, S.L. Prediciting water content using Gaussian model on soil spectra. *Remote Sens. Environ.* **2004**, *89*, 535–552.
20. Ben-Dor, E.; Chabrillat, S.; Dematte, J.; Taylor, G.R.; Hill, J.; Whiting, M.L.; Sommer, S. Using Imaging Spectroscopy to study soil properties. *Remote Sens. Environ.* **2009**, *113*, s35–s55.
21. Haubrock, S.N.; Chabrillat, S.; Lemmnitz, C.; Kaufmann, H.; Surface soil moisture quantification models from reflectance data under field conditions. *Inter'l J. Remote Sens.* **2008**, Vol. 29.
22. Monerris, A.; Schmugge, T. Soil moisture estimation using L-band radiometry. In *Advances in Geoscience and Remote Sensing*; Jedlovec, G., Ed.; InTech: Shanghai, China, 2009.
23. Panciera, R.; Walker, J.P.; Merlin, O. Improved Understanding of Soil Surface Roughness Parameterization for L-Band Passive Microwave Soil Moisture Retrieval. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 625–629.
24. Merlin, O.; Walker, J.P.; Chehbouni, A.; Kerr, Y. Towards deterministic downscaling of SOM soil Moisture using MODIS derived soil evaporative efficiency. *Remote Sens. Environ.* **2008**, *112*, 3935–3946.
25. Wu, X.; Walker, J.P.; Rüdiger, C.; Panciera, R. Effect of Land-Cover Type on the SMAP Active/Passive Soil Moisture Downscaling Algorithm Performance. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 846–850.
26. Song, C.; Jia, L.; Menenti, M. Retrieving High-Resolution Surface Soil Moisture by Downscaling AMSR-E Brightness Temperature Using MODIS LST and NDVI Dat. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 935–942.
27. Henrich, V.; Krauss, G.; Götze, C.; Sandow, C. Index DataBase. 2012. Available online: https://www.indexdatabase.de/ (accessed on 3/5/2017).
28. Cannon, A.J.; Whitfield, P.H. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. *J. Hydrol.* **2002**, *259*, 136–151.
29. Onyari, E.K.; Ilunga, F.M. Application of MLP Neural Network and M5P Model Tree in Predicting Streamflow: A Case Study of Luvuvhu Catchment, South Africa. *Int. J. Innov. Manag. Technol.* **2013**, *4*, 11–15.
30. Deepa, C.; SathiyaKumari, K.; Pream Sudha, V. Prediction of the Compressive Strength of High-Performance Concrete Mix using Tree Based Modeling. *Int. J. Comput. Appl.* **2010**, *6*, 18–24.
31. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.
32. Ulaby, F.T.; Moore, R.K.; Fung, A.K. Microwave remote sensing: Active and passive. In *Microwave Remote Sensing: Active and Passive*; Volume III: From Theory to Applications; Remote Sensing Series; Artech House: Norwood, MA, USA, 1986; p. 4.
33. Lakens, D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc. Psychol. Pers. Sci.* **2017**, *8*, 355–362, doi:10.1177/1948550617697177.
34. Munoz-Carpena, R.; Shukla, S.; Morgan, K. Field Devices for Monitoring Soil Water Content. In *Southern Regional Water Program*; United States Department of Agriculture: Washington, DC, USA, 2004.
35. Ye, N.; Walker, J. P.; Wu, X., R.; de Jeu, Gao, Y.; Jackson, T. J.; Jonard, F.; Kim, E.; Merlin, O.; Pauwels, V.; Renzullo, L.; Rudiger, C.; Sabaghy, S.; von Hebel, C.; Yueh, S. H.; Zhu, L. The Soil Moisture Active Passive Experiments: Towards calibration and validation of the SMAP Mission. **2017**, Remote Sensing of Environment, In Review.