Advances in Soil Moisture Retrieval from Multispectral Remote Sensing Using Unmanned Aircraft Systems and Machine Learning Techniques

Samuel N. Araya[1], Anna Fryjoff-Hung[2], Andreas Anderson[2], Joshua H. Viers[2,3], and Teamrat A. Ghezzehei[2,4]

[1]Earth System Science, Stanford University, Stanford, CA, USA
[2]Center for Information Technology in the Interest of Society and the Banatao Institute, University of California, Merced, CA, USA
[3]Department of Civil and Environmental Engineering, University of California, Merced, CA, USA
[4]Life and Environmental Science, University of California, Merced, CA, USA

Correspondence to: Samuel N. Araya (araya@stanford.edu)

Abstract. This study investigates the ability of machine learning models to retrieve surface soil moisture of a grassland area from multispectral remote sensing carried out using an unmanned aircraft system (UAS). In addition to multispectral images, we use terrain attributes derived from a digital elevation model and hydrological variables of precipitation and potential evapotranspiration as covariates to predict surface soil moisture. We tested four different machine learning algorithms and interrogated the models to rank the importance of different variables and to understand their relationship with surface soil moisture~~We developed machine learning models to retrieve surface soil moisture (0 – 4 cm) from high-resolution multispectral imagery using terrain attributes and local climate covariates. Using a small unmanned aircraft system (UAS) equipped with a multispectral sensor we captured high-resolution imagery in part to create a high-resolution digital elevation model (DEM) as well as quantify relative vegetation photosynthetic status. We tested four different machine learning algorithms.~~ . All the machine learning algorithms we tested were able to predict soil moisture with good accuracy. ~~The~~ The boosted regression tree algorithm was marginally the best ~~gave the best prediction~~ with mean absolute error of 3.8 % volumetric ~~water~~ moisture content. Variable importance analysis revealed that the four most important variables were ~~The most important variables for the prediction of soil moisture were~~ precipitation, reflectance in the red wavelengths, potential evapotranspiration, and topographic position indices (TPI). Our results demonstrate that the dynamics of soil water status across heterogeneous terrain may be adequately described and predicted by UAS remote sensing data and machine learning. Our modeling approach and the variable importance and relationships we have assessed in this study should be useful for management and environmental modeling tasks where spatially explicit soil moisture information is important.

List of Acronyms

ALE	Accumulated local effects
ANN	Artificial neural network

BRT	Boosted regression trees
DEM	Digital elevation model
MAE	Mean absolute error
MBE	Mean bias error
NDVI	Normalized difference vegetation index
NIR	Near-infrared
PET	Potential evapotranspiration
RF	Random forest
RMSE	Root mean square error
RVR	Relevance vector regression
SVR	Support vector regression
TDR	Time-domain reflectometry
TPI	Topographic position index
TTVI	Thiam’s Transformed vegetation index
UAS	Unmanned aircraft systems

1 Introduction

The relatively small quantity of water stored in the upper layers of soil plays a key role in terrestrial biology, biogeochemistry, and atmospheric water and energy fluxes. More than half of the solar energy absorbed by the land surface is used to evaporate water (Trenberth et al., 2009) and about 60% of terrestrial precipitation is returned to the atmosphere by evapotranspiration (Seneviratne et al., 2010).

In most environments, soil water storage mainly depends on precipitation and evapotranspiration (Hillel, 1998; Rana and Katerji, 2000), but the distribution of water in the soil is also dependent on the soil hydraulic properties, topography, and other environmental and belowground conditions (Korres et al., 2015; Vereecken et al., 2014; Western et al., 1999). Because of the complex interplay of these variables, it is a challenge to accurately estimate soil water.

It is ~~very difficult, or~~typically ~~even~~ impractical~~,~~ to acquire data on soil water dynamics by direct measurement ~~over~~ on ~~large~~ scales larger than a small experimental plot ~~,~~ and there is no robust approach to predict it. The scarcity of soil moisture ~~observations~~ data is a major impediment for the investigation of soil moisture-climate interaction. ~~New~~ Modern techniques ~~to~~ for large-scale measurement of soil moisture include the cosmic-ray soil moisture observing system~~,~~ (COSMOS~~,~~) and the GPS interferometric reflectometry (GPS-IR) based methods. The COSMOS employs a network of probes across the U.S. that estimate soil moisture by measuring cosmic-ray neutron radiation intensity above the land surface (Zreda et al., 2012). GPS based methods are also able to estimate soil moisture of a few square meters using GPS signal reflected from the soil. ~~For t~~These ~~new~~ techniques, while very promising still need to be refined for routine uses~~, the empirical confirmation of theoretical predictions of variable measurement depth and standardization of procedures still need to be refined~~ (Ochsner et al., 2013). ~~Remote sensing techniques can fill the need for spatial coverage. The availability and accuracy of satellite remote sensing~~

~~technologies have been steadily increasing over the years. However, satellite remote sensing methods are often constrained by their spatial and temporal resolution and shallow depth of measurement (Nichols et al., 2011).~~

Remote sensing methods of retrieving soil moisture provide an alternative to conventional methods of soil moisture measurement, which are impractical at large scales. Several remote sensing methods, particularly from spaceborne deployment, have been developed to retrieve soil moisture using optical, thermal infrared, and microwave sensors. Remote sensing methods enable spatially distributed and frequent observations over a large area, which is difficult to achieve using conventional field measurements (Barrett and Petropoulos, 2014; Petropoulos et al., 2015). A critical challenge to current remote sensing methods of retrieving soil moisture is the lack of imagery with optimum spatial resolutions appropriate for field-scale soil moisture studies and the low re-visit frequency of satellites (Barrett and Petropoulos, 2014; Das and Mohanty, 2006). Alternatives based on manned airborne platforms are limited due to their high operational costs. A significant challenge to remote sensing methods is estimating the root-zone soil moisture using surface observation obtained from remote sensing (Nichols et al., 2011; Ochsner et al., 2013).

Water is one of the most significant chromophores in soils and studies have shown that narrow band spectral information in the visible (0.4 – 0.7 µm), near-infrared (0.7 – 1.1 µm), and shortwave infrared (1.1 – 2.5 µm) regions can be used to estimate surface soil moisture (Ben-Dor et al., 2009; Malley et al., 2004). Soil reflectance in the visible to shortwave infrared spectral region generally decreases with an increase in soil moisture, with some parts of the spectrum showing a more pronounced decrease than others (Haubrock et al., 2008; Weidong et al., 2002). The hydroxide bond is the strongest absorber in the near-infrared region and free water in soil pores has strong absorption around 1.4 and 1.9 µm wavebands (Malley et al., 2004). Several hyperspectral techniques to estimate soil moisture content have been developed such as the Soil Moisture Gaussian Model (SMGM) (Whiting et al., 2004) and the Normalized Soil Moisture Index (NSMI) (Haubrock et al., 2008).

In the presence of vegetation cover, however, the ability to use soil reflectance to measure soil moisture is limited (Muller and Décamps, 2000). In addition, soil reflectance of solar radiation represents only the upper 50 µm of soil, and this makes it challenging to estimate moisture conditions in deeper layers (Malley et al., 2004). Most soil moisture remote sensing approaches operating in the optical range rely on developing an empirical spectral vegetation index (Barrett and Petropoulos, 2014). Several soil moisture measurement methods based on vegetation index proxies have been suggested as vegetation indexes are extremely sensitive to water stress, and they allow indirect estimates of soil moisture (Zhang and Zhou, 2016). Many studies have focused on deriving surface soil moisture content from synergistic use of remote sensing data acquired simultaneously in the optical and thermal infrared spectrum. The so-called ‘universal triangular relationship’ is a widely used method for estimating soil moisture (Nichols et al., 2011; Sobrino et al., 2014).

Retrieval of information from remote sensing measurements is based on the principle that changes in the chemical, physical,
Retrieval of information from remote sensing measurements is based on the principle that changes in the chemical, physical,

~~The r~~Retrieval of information from ~~measurements performed by~~ remote sensing measurements is based on the principle that changes in the chemical, physical, and/or structural characteristics of a target determine the variations of its electromagnetic response (Schanda, 1986). The task of retrieving information from remote sensing is complicated by several factors.~~:~~ Ali et al. (2015) ~~outline~~ summarize the ~~four general challenges of the~~ retrieval ~~problems~~challenges as~~:~~ (i) the often complex and non-linear relation between remote sensing measurement and target variables of interest; (ii) the ill-posed nature of the retrieval problem in that electromagnetic response of a target is typically the result of contributions from multiple target variables and similar electromagnetic responses may be associated with different physical variables; (iii) the mixed contribution of multiple objects represented within elementary resolution cell; and (iv) the influence of external disturbing factors such as noise, radiation components coming from surrounding of the investigated area and the atmosphere.

Soil moisture retrieval from remote sensing has traditionally been done by either empirical approaches or approaches based on an inversion of physical models. More recently, the use of machine learning techniques has gained increased attention because of their ability to tackle many of the limitations with the empirical and physical-model based approaches.

Physical-model-based approaches depend on understanding of mechanisms involving the interaction of electromagnetic radiation and the target variable. A wide variety of analytic electromagnetic models have been proposed in the literature. The thermal inertia approach (Price, 1977) is one such method that is most commonly used for soil moisture retrieval using thermal infrared (wavelengths between 3.5 and 14 µm) observation (Barrett and Petropoulos, 2014; Zhang and Zhou, 2016). Many new soil thermal inertia estimation methods continue to be developed (Price, 1985; Tian et al., 2015; Zhang and Zhou, 2016). The advantages of such physically based models are that they can operate in more general scenarios that are difficult to represent through the collection of in situ measurements. However, such models rely on simplifying the representation of a real phenomenon, which can reduce reliability. A major drawback of analytical models is their complexity and requirement for a large number of input parameters (Zhang and Zhou, 2016).

(Zhang and Zhou, 2016).

Empirical modeling approaches employ statistical regression techniques to develop a mapping function based on couples of in situ measurements of the target variable and corresponding remote sensing measurement ~~(Ali et al., 2015)~~. ~~Water is one of the most significant chromophores in soils and studies have shown that narrow band spectral information in the visible (0.4 – 0.7 µm), near-infrared (0.7 – 1.1 µm) and shortwave infrared (1.1 – 2.5 µm) regions can be used to estimate surface soil moisture~~

~~(Ben-Dor et al., 2009; Malley et al., 2004). Soil reflectance in the visible to shortwave infrared spectral region generally decreases with an increase in soil moisture, with some parts of the spectrum showing a more pronounced decrease than others (Haubrock et al., 2008; Weidong et al., 2002). The hydroxide bond is the strongest absorber in the near-infrared region and free water in soil pores has strong absorption around 1.4 and 1.9 µm wavebands (Malley et al., 2004). Several hyperspectral techniques to estimate soil moisture content have been developed such as the Soil Moisture Gaussian Model (SMGM) (Whiting et al., 2004) and the Normalized Soil Moisture Index (NSMI) (Haubrock et al., 2008).~~

~~In the presence of vegetation cover, however, the ability to use soil reflectance to measure soil moisture is limited (Muller and Décamps, 2000). In addition, soil reflectance of solar radiation represents only the upper 50 µm of soil, and this makes it challenging to estimate moisture conditions in deeper layers (Malley et al., 2004). Most soil moisture remote sensing approaches operating in the optical range relay on developing an empirical spectral vegetation index (Barrett and Petropoulos, 2014). Several soil moisture measurement methods based on vegetation index proxies have been suggested as vegetation indexes are extremely sensitive to water stress, and they allow indirect estimates of soil moisture (Zhang and Zhou, 2016). Many studies have focused on deriving surface soil moisture content from synergistic use of remote sensing data acquired simultaneously in the optical and thermal infrared spectrum. The so-called 'universal triangular relationship' is a widely used method for estimating soil moisture (Nichols et al., 2011; Sobrino et al., 2014).~~

The advantage of empirical relationships is that they are typically fast to derive and do not require too many inputs. However such models require ~~(Ali et al., 2015). The disadvantages of empirical models are the need for~~ higher quality ground measurement, which could be ~~time~~ time-consuming and expensive, and that the derived relationship is typically site and sensor dependent which limits the possibility to extend their use in a different area readily (Ali et al., 2015)~~.~~.

Some disadvantages specific to remote sensing methods in the optical and thermal infrared spectrum are the fact that these wavelengths have shallow soil penetration and require cloud-free conditions. Many of the optical and thermal infrared synergistic approaches require a wide range of both vegetation index and soil moisture conditions within a study region which cannot always be satisfied (Barrett and Petropoulos, 2014).

The advantages of machine learning techniques in remote sensing are their ability to learn and approximate complex non-linear mappings and the fact that no assumptions need to be made about data distribution. They can thus integrate data from different sources with poorly-defined or unknown probability density functions ~~(Ali et al., 2015). Machine learning techniques~~ and have often been shown to outperform other parametric approaches (Ali et al., 2015; Paloscia et al., 2008). ~~Furthermore, machine learning techniques improve with an increasing number of observed datasets.~~ Some of the limitations of machine learning methods are the need for a large number of training data, which require extensive ground truth datasets, and that machine learning methods are black boxes and only limited inference can be made about the relationships of different inputs.

Remote sensing from unmanned aircraft systems (UAS) has the potential to address several limitations of traditional remote sensing. The most attractive feature is their high spatial resolution, frequent or on-demand image acquisition, and low operating costs (Anderson and Gaston, 2013; Berni et al., 2009; Colomina and Molina, 2014; Elarab, 2016; Manfreda et al., 2018; Tmušić et al., 2020). UAS is an umbrella term that refers to the unmanned aircraft and the complementary ground control and communication systems necessary for air surveys (Singh and Frazier, 2018).

1.1 Objectives

The purpose of this research was to advance soil moisture change measurement, process understanding, and prediction using remote sensing products from UAS and machine learning methods. In this study, the spatial and temporal scale limitations were addressed by deploying multispectral remote sensing with small UAS and address the challenge of retrieving surface soil moisture changes using machine learning methods and fusing remote sensing data with ground data and meteorological data.

The specific goals of this study were to: (1) develop an adaptable method to retrieve information on surface soil moisture from small UAS remote sensing products and machine learning methods, (2) identify important reflectance and surface characteristics for the prediction of soil moisture changes (3) identify appropriate spatial resolutions of reflectance images and terrain variables for estimating soil moisture, and (4) explore the relation of soil moisture with surface properties.

2 ~~-~~Methods

~~Several machine learning algorithms exist for multivariate regression modeling. Artificial neural networks (ANN) are among the most commonly used algorithms for the retrieval of soil moisture from remote sensing (e.g., Hassan-Esfahani et al., 2015; Paloscia et al., 2008). In recent years, the support vector machine (SVM) and the similar support vector regression (SVR) algorithms have become popular in the retrieval of soil moisture (e.g., Ahmad, Kalra, & Stephen, 2010; Zaman & Mckee, 2014; Zaman, McKee, & Neale, 2012). Other popular machine learning algorithms include tree-based models such as the random forest (RF) and boosted regression trees (BRT).~~

~~2.1 Artificial Neural Network (ANN)~~

~~ANN models have been widely used in the development of PTFs (Matei et al., 2017; Pachepsky et al., 1996; Schaap et al., 2001; Zhang et al., 2018; Zhang and Schaap, 2017). ANNs are universal approximators that can approximate any nonlinear mapping. The feed-forward neural network is a popular variant of ANN. In this study, we implemented the feed-forward neural networks with a single hidden layer.~~

~~2.2 Support Vector Regression (SVR)~~

~~SVR is an adaptation of the support vector machine (SVM) for regression problems (Cortes and Vapnik, 1995; Drucker et al., 1997). The SVM learning is a generalization of 'maximal margin classifier.' The algorithm first maps the input variables into~~

~~a high-dimensional space using a fixed mapping function – a kernel function. The algorithm then constructs hyperplanes, which can be used for classification or, in the case of SVR, for regression. In this study, we use the Radial Basis Function kernel, which is one of the most commonly used kernels in SVR. Some advantages of SVR include the fact that they do not suffer from the problem of local minima, and that they have few parameters to tune when training the model.~~

~~2.3 Relevance Vector Regression (RVR)~~

~~Like SVM, the RVR originally introduced as a classification machine (Tipping, 2000). RVR is a Bayesian treatment of the SVM prediction function which avoids some of the limitations of SVM algorithms, such as reducing the use of basis functions and the need for optimizing the cost and the insensitivity parameters (Ben-Shimon and Shmilovici, 2006). Torres-Rua et al. (2016) successfully used the RVR algorithm to estimate surface soil moisture from satellite image and energy balance products.~~

~~2.4 Random Forest (RF)~~

~~RF are popular models that are relatively simple to train and tune (Hastie et al., 2009). They apply ensemble techniques by averaging a large number of individual decision tree-based models. Tree models are ‘grown’ by searching for a predictor that ensures the best split that results in the smallest model error. The individual trees in RF ensemble are built on bootstrapped training sample, and only a small group of predictor variables are considered at each split, this ensures that trees are de-correlated with each other (Breiman, 2001; James et al., 2013).~~

~~2.5 Boosted Regression Trees (BRT)~~

~~BRT is another form decision tree model ensemble enhanced by the gradient boosting approach. The gradient boosting algorithm constructs additive regression models by sequentially fitting ‘simple base learner’ functions (i.e., decision trees) to current pseudo-residuals at each iteration (Friedman, 2002). These pseudo-residuals are the gradient of the loss function being minimized. BRT models have shown considerable success and often outperform other machine learning algorithms in many situations (Elith et al., 2008; Natekin and Knoll, 2013). BRT models are also particularly adept for less-than-clean data (Friedman, 2001), which makes them particularly attractive in our work where the training data is compiled from various sources and different measurement methods which makes it prone to some inconsistencies.~~

~~Tree-based models, both the RF and BRT, have the advantage of being able to rank predictor variable’s relative importance. In these models, the approximate relative influence of a single predictor variable is calculated as the empirical improvement of predictions by splitting on that predictor at each node and then averaging the relative influence of the variable across all trees of the model (Ridgeway, 2012).~~

Methods

Multispectral images of the study area were collected on six different days throughout the 2018 water year using a UAS equipped with a multispectral camera. A High-resolution, cm-scale, digital elevation model (DEM) was generated from the stereo images using photogrammetric software, and multiple sets of terrain variables were calculated. Concurrently with the image acquisition flights, moisture content of top 3.8 cm of soil was measured at predefined sampling locations. The ground soil moisture measurements, multispectral reflectance, terrain variables, and rainfall and potential evapotranspiration (PET) data were then aggregated into a data table and used to train a machine learning model to predict the soil moisture. Figure 1 shows the model building process.

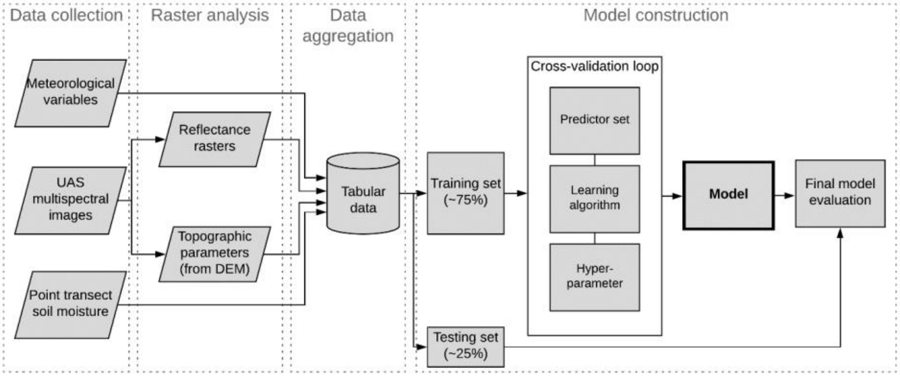

Figure 1 Process flowchart of model development.

3.12.1 Study Site

The study was conducted in a small grassland catchment at the Merced Vernal Pools and Grassland Reserve located about five kilometers northeast of the City of Merced, California. The grassland is used for livestock grazing, it has a Mediterranean climate with hot, dry summers and cool, wet winters with an average annual precipitation of 330 mm

~~The Merced Vernal Pools and Grassland reserve covers an area of about 26.6 km^2 and protects hundreds of ephemeral pools and wetlands~~ (Wong, 2014). ~~The reserve was historically and still is used for livestock grazing.~~

Our study site is a 0.6 km^2 area of land located within a sub-catchment that contributes to the Avocet Pond, a large stock pond located in the northeast corner of the Reserve (Figure 2). The catchment was selected because of an extensive hydrologic modeling study that was being conducted on the site at the time (Fryjoff-Hung, 2018).

Formatted: Justified

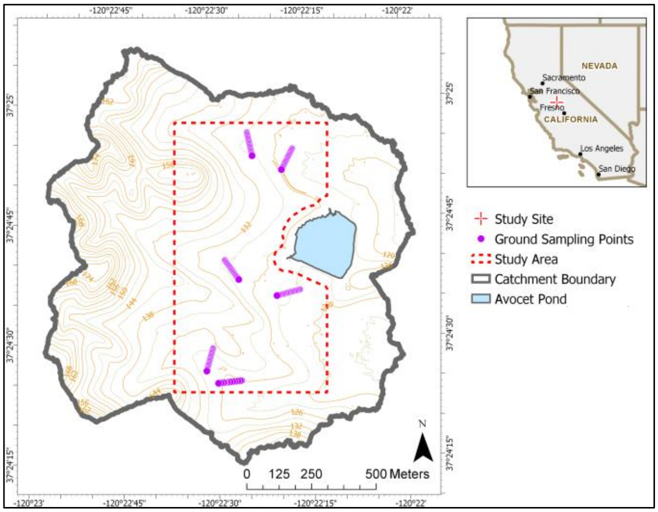

Figure 2 Map of Avocet Pond catchment showing the footprint of the study area, ground sampling points, and elevation contours in meters. Inset shows the location of the study site (red crosshair) in California.

The study area soils are dominated by Redding gravelly loam (Fine, mixed, active, thermic Abruptic Durixeralfs) soils. The elevation of the study area ranges from 118 to 162 m above sea level, and slope ranges from 0 to 31°. The distributions of elevation and slope are shown in supplemental Figure S1.

The ~~fact that~~ vernal pools ecology is predominantly controlled by large seasonal shifts and high spatial variability in hydrology, this makes the study site ~~particularly~~ attractive for our study~~to the proposed research~~. UAS remote sensing ~~have~~ has the potential to provide information at appropriate spatial and temporal scales for vernal pool studies (Stark et al., 2015). ~~Knowing soil moisture dynamics at a higher frequency is important, especially during the seasonal transition times.~~ The annual seasonal cycle of the study site is shown in Figure 3.

Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Transition to wet		**Wet**				**Transition to dry**		**Dry**			

Figure 3 Vernal pool annual moisture cycle.

~~3.2~~2.2 Data Collection and Preparation

The imagery was acquired on six days during the 2018 water year green-up and brown-down using a fixed-wing UAS with a multispectral camera onboard (See Table 3). Figure 4 shows a typical scene of the study site during wet and dry seasons. Point soil moisture measurements (top 4 cm) were collected with a time-domain reflectometry (TDR) probe across precise sampling transects identified with real-time kinematic (RTK) positioning survey. Daily rainfall and PET values were acquired from nearby weather stations.

Figure 4 Typical scene of the study area in April (left) and June(right) of 2018.

~~3.2.1~~2.2.1 Image acquisition and processing

~~The UAS remote sensing flights were conducted in the late mornings to mid-day during clear weather conditions. A single remote sensing campaign takes approximately three to four hours, meaning images were acquired between approximately 10:00 AM and 2:00 PM.~~

Multispectral images were acquired using *Parrot Sequoia* Sensor (Parrot SA, Paris, France) equipped with a sunshine sensor that measured irradiance at the sensor spectral wavebands for radiometric normalization. The camera is deployed on a fixed-wing unmanned aircraft (Finwing Sabre, Finwing Technology) with an average flight height of 120 m above ground level. Images of a calibrated reflectance panel (MicaSense, Inc, Seattle, ~~WA) were~~WA) were taken before each flight and used in the radiometric calibration of images. The UAS remote sensing flights were conducted between late mornings and early afternoons during clear weather conditions. A single remote sensing—soil moisture collection campaign takes between three to four hours. Images were acquired between 10:00 AM and 2:00 PM.

The Parrot Sequoia sensor captures four separable bands in the green, red, red edge, and near-infrared bands with a focal length

The Parrot Sequoia sensor captures four separable bands in the green, red, red edge, and near-infrared bands with a focal length of 3.98 mm and resolution of 1280x960 pixels. A fifth channel captures a high-resolution image in the visible spectrum with a focal length of 4.88 mm and a resolution of 4608x3456 (Pix4D, n.d.). ~~About 12,000 images~~ Images for the study area ~~are~~ were captured with a minimum of 85 percent overlap ~~per flight with~~ and ~~a~~ ground pixel resolution of ~~approximately~~ 10 to 15 cm. Images were scaled to uniform 15 cm pixel resolution post-processing. Images processing was done using Pix4D photogrammetry software (Pix4D, Lausanne, Switzerland). A digital elevation model (DEM) was photogrammetrically generated from the overlapping stereo-images~~Images are~~ and images~~-~~ were ~~mosaicked,~~ orthorectified, and radiometrically calibrated. ~~using Pix4D photogrammetry software (Pix4D, Lausanne, Switzerland). The Pix4D software also generates DEM photogrammetrically from stereo-images.~~

2.2.1.1 Geometric and radiometric corrections

Between 7 to 9 ground control point targets (GCPs) with precise locations identified by RTK survey were used for photo alignment. The mean georeferencing root-mean-square-errors (RMSE) of the GCPs ranged from 0.6 – 2 cm, and mean reprojection errors ranged from 0.1 – 0.2 pixel based on the bundle block adjustment error assessment report. DEM was generated using the structure-from-motion technique; noise filtering and mild surface smoothing (sharp smoothing) were applied to correct for noisy and erroneous points of the point cloud. The inverse distance weighting algorithm was used to interpolate between points to create the raster DEM.

Radiometric calibration by the Pix4D software considers the positional data, solar irradiance measurements, and gain and exposure data from the camera to convert raw digital numbers into sensor reflectance values. Sensor reflectance represents the ratio of the reflected light to the incoming solar radiation and provides a standardized measure that is directly comparable between images. Finally, surface reflectance is calculated in post-processing, taking into account the camera's orientation, the angle of the sun, and the known reflectance values of the calibration panel.

3.2.22.2.2 In situ soil moisture measurement

The moisture content of the top 4 cm soil was measured simultaneously with UAS remote sensing flights using FieldScout TDR-300 soil moisture meter equipped with a 3.8 cm probe (Spectrum Technologies Inc., IL, USA). The FieldScout TDR-300 measures volumetric water content using time-domain reflectometry with a resolution of 0.1% and an accuracy of ±3% (Spectrum Technologies Inc., 2009).

Accurate geolocation of the in-situ soil moisture measurement points is critical for over lay analysis of the ground measurements with the remote sensing products. To ensure accurate geolocations of the ground measurements, we identified six, 90 m long sampling transects and recorded survey-grade geolocation of the transect ends using RTK positioning survey. The transect ends were marked with a metal peg hammered into the ground. During each soil moisture measurement campaign, a 90 m tape measure was temporarily affixed between the two ends of the transect, and soil moisture measurements taken about every 10 m, noting the exact distance of the sampling point from the transect ends.

The sampling transects were laid out in a way that ensured they run over a variety of topographic variables in terms of flow

The sampling transects were laid out in a way that ensured they run over a variety of topographic variables in terms of flow

The sampling transects were laid out in a way that ensured they run over a variety of topographic variables in terms of flow accumulation, topographic wetness index, and stream networks. Furthermore, each sampling transect fell in a separate sub-basin within the Avocet basin.

each sampling transect fell in a separate sub-basin within the Avocet basin

.

~~Once we decided on the location of the sampling transects the location of the two ends was recorded accurately using RTK positioning survey and marked with a metal peg hammered into the ground to allow for repeated measurement at the same location. During the soil moisture measurement campaign, we temporarily affixed a 90 m tape measure at the two ends of the transect and took soil moisture measurement about every 10 m noting the exact distance of the sampling point from the transect ends.~~

Hydrological variables

Daily precipitation data was retrieved from the UC Merced weather station located approximately six km southwest of the study site (California Department of Water Resources, 2018). Daily reference evapotranspiration data was retrieved from the California Irrigation Management Information System's Merced station located approximately 10 km south of the study site (California Irrigation Management Information System, 2018). The Reference evapotranspiration is evapotranspiration from standardized grass calculated using the modified Penman (CIMIS Penman) and the Penman-Monteith equations (California Irrigation Management Information System, n.d.). In this study, we will refer to the reference evapotranspiration as the potential evapotranspiration (PET).

3.32.3 Data Processing

To prepare the data for machine learning. We compiled all the information into a table with the measured soil moisture from each sampling point and date organized into one column. Each row contained the accompanying information for that sampling point and time.

3.3.1 Feature engineering

~~*3.3.1.1 Geometric Correction*~~

~~Pix4D software was used to process the images. Between 7 to 9 ground control point targets (GCPs) with precise locations identified by RTK survey were used for photo alignment. The mean georeferencing root-mean-square-errors (RMSE) of the GCPs ranged from 0.6 – 2 cm, and mean reprojection errors ranged from 0.1 – 0.2 pixel based on the bundle block adjustment error assessment report. DEM was generated using the structure-from-motion technique; noise filtering and mild surface smoothing (sharp smoothing) were applied to correct for noisy and erroneous points of the point cloud. The inverse distance weighting algorithm was used to interpolate between points to create the raster DEM.~~

~~*3.3.1.2 Radiometric correction*~~

~~Along with capturing images in the four spectral bands, the multispectral camera records the location, orientation, and solar irradiation using its GPS, inertial measurement unit (IMU), and sunshine sensor, respectively. Radiometric calibration by the Pix4D software considers the positional data, solar irradiance measurements, and gain and exposure data from the camera to convert raw digital numbers into sensor reflectance values. Sensor reflectance represents the ratio of the reflected light to the incoming solar radiation and provides a standardized measure that is directly comparable between images. Finally, surface reflectance is calculated in post-processing, taking into account the camera's orientation, the angle of the sun, and the known reflectance values of the calibration panel.~~

Feature engineering

We calculated several variables based on the multispectral reflectance, terrain, and hydrological data to be used to train a machine learning model as predictor variables. A list of all the measured and calculated variables used in modeling soil moisture are given in Table 1.

~~*3.3.2.1 Reflectance based vegetation index*~~

~~We calculated the Thiam's Transformed vegetation index (TTVI) based on the red and near-infrared bands (Equation 1).~~

~~$$TTVI = \sqrt{\left|\frac{NIR - R}{NIR + R} + 0.5\right|} \qquad (1)$$~~

~~The TTVI is a transformation of the commonly used Normalized Difference vegetation index (NDVI). The reason for choosing of TTVI over NDVI is that it eliminates negative values and transforms NDVI histograms into a normal distribution.~~

~~*3.3.2.2 Terrain variables*~~

~~A list of the calculated topographic variables and their description is given in Table 1.~~ Topographic variables derived from DEM are scale-dependent, to account for this, we calculated all topographic variables on six different resolution DEM. ~~For this,~~Prior to calculating topographic variables, we ~~first~~ upscaled the DEM ~~from the original resolution of 6.85 cm~~ to 15, 30, 60, 100, 300, and 500 cm cell resolutions~~. We~~ and then calculated topographic variables ~~on~~ for all the resolutions.

The calculation of topographic position index (TPI) is a special case since it does not only depend on DEM resolution but also on the definition of inner and out radii of the annulus (see Equation (~~2~~1)).

$$TPI = Elevation - focal\ mean\big(annulus(Inner\ Radius, Outer\ Radius)\big) \qquad (\cancel{2}1)$$

We calculated TPI for different neighborhood sizes using the ArcGIS 10.5 Land Facet Corridor Tool (Jenness et al., 2013). We calculated TPI on three DEM resolutions (100, 300, and 500 cm) with two inner radii (1 and 3 cells) and three outer radii (3, 5, and 7 cells). A map of selected topographic variables is provided in supplemental materials, Figure S2.

Precipitation and PET are two important drivers of surface soil moisture. To account for antecedent conditions, we used the

~~*3.3.2.3 Meteorological variables*~~

Precipitation and PET are two important drivers of surface soil moisture. To account for antecedent conditions, ~~We~~ we used the cumulative water year precipitation and PET~~. We~~ as well as~~ also calculated~~ rolling sums of ~~those~~ both variables with different time windows before the measurement dates.~~, we~~ We calculated 1-, 2-, 3-, 7-, 15, and 30-day cumulative precipitation and PET before sampling dates and used those rolling sums as input.

Table 1 Measured and calculated data used for machine learning. All topographic variables are computed from the digital elevation

Table 1 Measured and calculated data used for machine learning. All topographic variables are computed from the digital elevation model. Descriptions and significance of topographic variables adapted from Wilson & Gallant (2000).

	Variable [unit]	**Description**	**Significance/ relation to soil moisture**
Measured	Soil moisture content [%]	Volumetric soil moisture content	Variable of interest.
	Daily rainfall [mm]	Daily rainfall from Precipitation ~~Gage (~~Gage (OTT Pluvio) with a windshield	Source of soil moisture
	Green [-]	Surface reflectance in the green wavelength band (530 – 570 nm)	Soil and vegetation reflectance change
	Red [-]	Surface reflectance in the red wavelength band (640 – 680 nm)	Soil and vegetation reflectance change
	Red-edge [-]	Surface reflectance in the red-edge wavelength band (730 – 740 nm)	Soil and vegetation reflectance change
	Near-infrared [-]	Surface reflectance in the near-infrared wavelength band (770 – 810 nm)	Soil and vegetation reflectance change
	Altitude [m]	Elevation (m)	Vegetation, potential energy
Calculated	Daily potential evapotranspiration [mm]	Reference evapotranspiration from standardized grass calculated using CIMIS Penman equation.	Major soil moisture loss pathway.
	Thiam’s Transformed Vegetation Index (TTVI) [-]*	$TTVI = \sqrt{\left\lvert \left(\frac{NIR - R}{NIR + R}\right) + 0.5 \right\rvert}$	Vegetation moisture stress.
	Slope [degrees]	Slope gradient (degrees)	Surface and subsurface flow velocity, runoff rate, vegetation, geomorphology
	Aspect [$\cos(degees)$]	Cosine transformed direction of maximum downward gradient (northerness).	North and south-facing slopes differ in solar insolation, PET, flora and fauna distribution, and abundance.
	Profile curvature [-]	Downslope curvature	Flow acceleration, erosion/deposition rate, geomorphology
	Plan ~~curvature[~~curvature [-]	Along-side curvature	Converging/diverging flow, soil characteristics
	Tangential ~~curvature[~~curvature [-]	Curvature in an inclined plane	Represents areas of convergent (concave) and divergent (convex) flow.
	Flow accumulation (MFD methods) (A) [cm^2]	Catchment area draining to pixel	Runoff volume, geomorphology

	Variable [unit]	Description	Significance/ relation to soil moisture
	Length-Slope factor, LS [-]	Length-slope factor from the Revised Universal Soil Loss Equation (RUSLE). For slope lengths <100 m and slopes <14 °: $LS = 1.4\left(\frac{A}{22.12}\right)^{0.4}\left(\sin\frac{S}{0.0896}\right)^{1.3}$	Calculates a spatially distributed sediment transport capacity
	Topographic position index, based SFD and MFD methods [-]	$TPI = Z_0 - \frac{1}{n_R}\sum_{i \in R} Z_i.$	
	Topographic wetness index, based on SFD and MFD methods [-]	$TWI = \ln\left(\frac{A}{\tan S}\right)$	Commonly used index to quantify topographic control on hydrological process.

* The TTVI is a transformation of the commonly used Normalized Difference vegetation index (NDVI). The reason for choosing the TTVI transformation is that it eliminates negative values and often transforms NDVI histograms into a more normal distribution. Normalizing machine learning inputs is considered good practice and aids models to converge faster.

~~3.3.3~~2.3.2 ~~Data aggregation~~Variable selection

~~We had a~~The total number of soil moisture measurements was 406. ~~soil moisture measurements across the six different measurement times~~This number determines the total length of the data ta, that is, the total number of rows. For each soil moisture point, we added columns with the corresponding date, reflectance, topographic and hydrological variables. The reflectance and topographic variables were extracted from the raster images using the raster-to-point data extraction tool in ArcGIS software taking the average value of a 1-meter diameter buffer around the points. The hydrological variables were taken to be the same for the entire study area and only changed based on the measurement days. ~~we extracted the raster values of reflectance and terrain variable by using raster to point data extraction tool in ArcGIS Pro using a 100 cm radius means. For this, ground coordinates of sampling locations were overlaid onto the geo-referenced image, and the pixel values across the bands representing the center of each sampled area were extracted to result in tabular data. We also included the meteorological variables for each soil moisture reading.~~

The data preparation resulted in 138 variables. We employed

~~*3.3.3.1* Variable transformation~~

Formatted: Normal

~~Prior to model training, all the predictor variables were standardized by centering the training variable's mean to zero and scaling by the variable's standard deviation, as shown in Equation (3):~~

~~$x' = \frac{x - \bar{x}}{\sigma_{\bar{x}}}$ (3)~~

variable selection (or feature selections) methods to identify a subset of relevant variables (features) from the larger set of potential predictors. The benefits of variable selection include improvement of model performance, reducing training and utilization times, and facilitating data understanding (Guyon and Elisseeff, 2003; Weston et al., 2003). We employed three methods of variable selection: tests of linear correlation and linear dependencies among variables, and recursive feature elimination. Recursive feature elimination involves removing the least important features whose omission has the least effect on training errors (Chen and Jeong, 2007; Guyon et al., 2002). We implemented a recursive feature elimination procedure during the coarse tuning of BRT, RF, ANN, and SVR algorithm models.

~~*3.3.3.2* Variable selection~~

Formatted: Normal

~~V~~variable selection (or feature selections) ~~involves~~ methods to~~the~~ identify a~~selection of~~ ~~a~~ subset of relevant variables (features) from ~~a~~ the larger set of potential predictors. The benefits of variable selection include improvement of model performance, reducing training and utilization times, and facilitating data understanding (Guyon and Elisseeff, 2003; Weston et al., 2003). We employed three methods of variable selection: tests of linear correlation and linear dependencies among variables, and recursive feature elimination. Recursive feature elimination involves removing the least important features whose omission has the least effect on training errors (Chen and Jeong, 2007; Guyon et al., 2002). We implemented a recursive feature elimination procedure during the coarse tuning of BRT, RF, ANN, and SVR algorithm models.

Following the variable selection procedure, of the 138 variables, 76 variables were removed based on linear correlation and linear dependencies among variables. An additional 16 were removed following the recursive feature elimination procedure. The final data used for building the models had 46 variables (Table 2), of which five are hydrological, nine are reflectance, and 32 are topographic variables. Variable categories that had no importance included the topographic wetness index (TWI), the reflectance in the red-edge band, and NDVI.

Table 2 Predictor variables used for machine learning models.

Table 2 ~~Important p~~Predictor variables used ~~in final~~for machine learning models.

Domain	**Variable**	**Scale***
~~Meteoric~~Hydrological	Potential evapotranspiration	1, 30
	Precipitation	1, 15, 30
Reflectance	Green	0.6, 1, 3
	Red	0.6, 1, 3
	Near Infrared	0.6, 1, 3
Topographic	Northerness	0.6, 1, 3, 5
	Slope	0.6, 1, 3, 5

	Flow Direction	0.6, 1, 3, 5
	Flow Accumulation	0.6, 1, 3, 5
	Curvature (Profile)	1, 3, 5, 50
	Curvature (Planform)	0.6, 1, 3, 5, 50
	Topographic Position Index	(1,3), (3,7), (3,9), (5,15), (9,21), (15,35), (15,100)

* Scale for raster products is pixel resolution in meters, and cumulative days for the ~~meteoric~~ hydrological variables. Topographic Position Index scale is a combination of the inner-outer diameters in meters (see equation (1)).

~~3.4~~2.4 Data Description

The six data collection days in the 2018 water year and summary site statistics are given in Table 3. ~~Measurements were done on six different days spread throughout the 2018 water year. The c~~Cumulative and 30-day rolling sums of precipitation and PET for the 2018 water year are shown in the supplemental materials Figures S3 and ~~and 1 to 30-day rolling sums of precipitation and PET for the sampling dates are shown in Figure~~ S4.

Table 3 Data collection days and site summary statistics.

	Date	Day of the water year	Cumulative water-year precipitation [mm]	Cumulative water-year PET [mm]	Mean soil moisture (and standard deviation) [%]	Sample count
1	2017-10-30	30	5.8	96.58	2.47 (1.09)	60
2	2017-12-04	65	34.3	149.16	12.26 (5.23)	60
3	2018-01-23	115	85.6	199.15	24.85 (7.79)	64
4	2018-04-04	186	133.3	371.81	20.27 (10.72)	92
5	2018-05-01	213	177.9	489.18	8.98 (3.93)	74
6	2018-05-24	236	177.9	619.26	4.42 (1.68)	56

Figure 5 shows the distribution of soil moisture and TTVI during the six measurement days. The soil moisture measurement followed the general precipitation patterns but was also influenced by immediate rainfall events; the highest soil moisture occurred on the only measurement day where it had rained the day before (January 23, 2018). The vegetation greenness, as measured with TTVI followed the 15-day cumulative rainfall well with maximum greenness occurring on April 4, 2018 (Day of water year 186), and sharply decreasing the following two months (Figure 5).

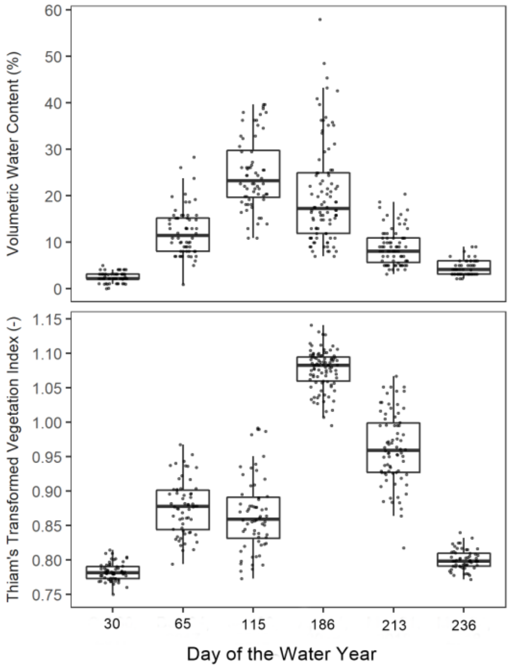

Figure 5 Measured soil moisture and vegetation index of the ground sampling locations from 1 m resolution raster.

Figure S5 in the supplemental materials shows the distribution of some terrain variables associated with the soil moisture sampling points and correlations between variables. The terrain variables for the ground sampling points show a reasonable distribution of values, while the distribution of elevation shows a bimodal distribution with ranges from 120 to 130 m, most of the other variables show a close to normal distribution. The only variables with Person's correlation above 0.5 are between TPI and curvature (Pearson's correlation = 0.67). The distribution of some variables in the data is shown in supplemental materials Figure S6.

3.52.5 Machine Learning Procedure

The overall machine learning procedure is illustrated in Figure 1. The computationally demanding steps of model training and testing were run at the Multi-Environment Research Computer for Exploration and Discovery (MERCED) high-performance computing cluster, at the University of California, Merced. The *caret* R package (Kuhn, 2017) was used to handle training and tuning procedures. The SVR and RVR algorithms were implemented using the *kernlab* package (Karatzoglou et al., 2004), RF algorithm was implemented using the *randomForest* package (Liaw and Wiener, 2015), and the BRT algorithm using the *xgboost* package (Chen and Guestrin, 2016).

Prior to model training, all the predictor variables were standardized by centering to mean zero and scaling by the variable's standard deviation (Equation (2)). Standardizing variables prior to model training is a good practice that minimizes issues of scale among input variables and often leads to better and faster training (Brownlee, 2020)

$$x' = \frac{x - \bar{x}}{\sigma_x} \tag{2}$$

where x' is the centered and scaled value of variable x and $\bar{x}$ and σ_x are the arithmetic mean and standard deviation of the variable.

2.5.1 Machine Learning Algorithms Used

Formatted: Heading 3

Several machine learning algorithms exist for multivariate regression modeling. Artificial neural networks (ANN) are among the most commonly used algorithms for the retrieval of soil moisture from remote sensing (e.g., Hassan-Esfahani et al., 2015; Paloscia et al., 2008). In recent years, the support vector machine (SVM) and the similar support vector regression (SVR) algorithms have become popular in the retrieval of soil moisture (e.g., Ahmad, Kalra, & Stephen, 2010; Zaman & Mckee, 2014; Zaman, McKee, & Neale, 2012). Other popular machine learning algorithms include tree-based models such as the random forest (RF) and boosted regression trees (BRT).

2.5.1.1 Artificial Neural Network (ANN)

ANN models have been widely used in the development of pedotransfer models (Matei et al., 2017; Pachepsky et al., 1996; Schaap et al., 2001; Zhang et al., 2018; Zhang and Schaap, 2017). ANNs are universal approximators that can approximate any nonlinear mapping. The feed-forward neural network is a popular variant of ANN. In this study, we implemented the feed-forward neural network with a single hidden layer which is considered sufficient for the majority of problems (Reed and Marks II, 1999).

2.5.1.2 Support Vector Regression (SVR)

SVR is an adaptation of the support vector machine (SVM) for regression problems (Cortes and Vapnik, 1995; Drucker et al., 1997). The SVM learning is a generalization of 'maximal margin classifier.' The algorithm first maps the input variables into a high-dimensional space using a fixed mapping function—a kernel function. The algorithm then constructs hyperplanes, which can be used for classification or, in the case of SVR, for regression. In this study, we use the Radial Basis Function kernel, which is one of the most used kernels in SVR. Some advantages of SVR include the fact that they do not suffer from the problem of local minima, and that they have few parameters to tune when training the model.

2.5.1.3 Relevance Vector Regression (RVR)

Like SVM, the RVR was originally introduced as a classification machine (Tipping, 2000). RVR is a Bayesian treatment of the SVM prediction function which avoids some of the limitations of SVM algorithms, such as reducing the use of basis functions and the need for optimizing the cost and the insensitivity parameters (Ben-Shimon and Shmilovici, 2006). Torres-Rua et al. (2016) successfully used the RVR algorithm to estimate surface soil moisture from satellite images and energy balance products.

2.5.1.4 Random Forest (RF)

RF are popular models that are relatively simple to train and tune (Hastie et al., 2009). They apply ensemble techniques by averaging a large number of individual decision tree-based models. Tree models are ‘grown’ by searching for a predictor that ensures the best split that results in the smallest model error. The individual trees in the RF ensemble are built on bootstrapped training sample, and only a small group of predictor variables are considered at each split, this ensures that trees are de-correlated with each other (Breiman, 2001; James et al., 2013).

2.5.1.5 Boosted Regression Trees (BRT)

BRT is another form decision tree model ensemble enhanced by the gradient boosting approach. The gradient boosting algorithm constructs additive regression models by sequentially fitting ‘simple base learner’ functions (i.e., decision trees) to current pseudo-residuals at each iteration (Friedman, 2002). These pseudo-residuals are the gradient of the loss function being minimized. BRT models have shown considerable success and often outperform other machine learning algorithms in many situations (Elith et al., 2008; Natekin and Knoll, 2013). BRT models are also particularly adept for less-than-clean data (Friedman, 2001), which makes them particularly attractive in our work where the training data is compiled from various sources and different measurement methods which makes it prone to some inconsistencies.

Tree-based models, both the RF and BRT, have the advantage of being able to rank predictor variable’s relative importance. In these models, the approximate relative influence of a single predictor variable is calculated as the empirical improvement of predictions by splitting on that predictor at each node and then averaging the relative influence of the variable across all trees of the model (Ridgeway, 2012).

3.5.12.5.2 Training-testing set splits

The data was split into training and testing sets of approximately 75-25 percent, respectively (i.e., approximately 300 and 100 records). The testing set was a hold-out set used only to evaluate final trained models.

The training-testing set splitting was done based on random selection of transect. For the testing set, two transects are randomly selected on four randomly selected sampling dates, and one transect is randomly selected on the remaining two sampling dates. To minimize bias that may result from the training-testing set split, we generated 30 unique training-testing set splits and trained 30 separate models based on each separate training set. The performance of each model was assessed on its respective testing set. Similar performance of the individual models would indicate that bias due to the training-testing set split is minimal. The justification for this sub setting procedure is: (1) the selection of entire transects as testing sets avoids the possible data leakage between the training and testing sets due to spatial autocorrelation since samples in a transect are located close to each other—a simple random splitting would not avoid this potential problem; (2) all six sampling dates are represented in the training set—models are trained on the entire range of time and soil moisture changes; and (3) the testing set is between 25 to 30 percent of the data (between 100 to 125 samples).

The distribution of samples across the sampling dates and transects for the training and testing sets are shown in supplemental materials Figures S7 and S8, respectively.

On average the training-testing split was 294 samples in the training sets and 113 samples in the testing sets. All the training sets have samples from all the six sampling dates and transects. While all sampling dates are represented in each testing set, on average, there are five transects in each testing set.

3.5.22.5.3 Cross-validation procedure

The selection of optimal model parameters in the model training process was done by the cross-validation method. Cross-validation is done to estimate the test error rate by holding out a subset of the training data (i.e., validation set) from the fitting process and then applying the fitted model to predict the validation subset. A 30-fold cross-validation set was generated by randomly splitting the training data into 80-20 percent training-validation split by randomly selecting a single transect every day. Optimum model parameters were selected using a comprehensive grid search method.

3.5.32.5.4 Model assessment

3.5.3.12.5.4.1 Performance

The final performance of models was assessed on the separate hold-out test dataset that was not used in the model training. The performance of models is measured in terms of mean absolute error (MAE), mean bias error (MBE), and the coefficient of determination (R^2) determined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$ (~~4~~3)

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$ (~~5~~4)

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$ (~~6~~5)

Where N is the number of observations; y is the measured value; $\hat{y}$ is the predicted value; and $\bar{y}$ is the mean of measured values.

The MAE indicates the average deviation of predictions from the measured value with smaller values indicating better performance. The MBE measures the average systematic bias, positive or negative values indicate the average tendency of the predicted values to be larger or smaller than the measured values, respectively. The R^2 measures the correspondence between predicted and measured data with higher values indicating stronger correspondence. The MAE was chosen over RMSE as it is a more appropriate measure when averaging (Willmott and Matsuura, 2005).

3.5.3.22.5.4.2 Variable importance

The predictor variable importance is the statistical significance of each predictor variable with respect to its effect on the generated model. For the tree-based models, RF and BRT, variable importance is calculated internally within the model algorithm. For the rest of the machine learning models, we calculated the predictor variable importance by recursive feature elimination method, which is done by recursively removing predictors before training a model and evaluating the change in model performance. In this method, to account for possible bias in variable subset selection (Ambroise and McLachlan, 2002; Hastie et al., 2009), we included a separate layer of 10-fold cross-validation to the entire sequence modeling steps.

3.5.3.32.5.4.3 *Effect of predictor variables*

The relationship between the predictor variables and outputs for a black-box model can be analyzed using model-independent methods such as partial dependence plots or accumulated local effects (ALE) plots (Apley, 2016; Greenwell, 2017). These plots help explain the relationship between the outcome of black-box supervised machine learning models and predictors of interest. We use the ALE plots to analyze the effect of selected predictor variables. Although similar, the ALE plots are preferred over partial dependence plots for their speed and their ability to produce unbiased plots when variables are correlated (Apley, 2016). The value of the ALE is centered so that the mean effect is zero; it can be interpreted as the effect of the variable on the outcome at a certain value compared to the average prediction of the data. For example, an ALE estimate of -2 when a variable of interest has a value of 3, then the prediction is lower by 2 compared to the average prediction (Molnar, 2019).

3.6 Results and Discussion

~~Preliminary UAS image processing—radiometric calibrations and, orthomosaic and DEM generation—is done using Pix4D photogrammetry software (Pix4D, Lausanne, Switzerland). Raster rescaling, terrain analysis, and spatial data visualization are done in ArcGIS Pro software (ESRI, Redlands, CA, USA). The machine learning process—data preparation, model tuning, and prediction—were done in R (R Core Team, 2019).~~

~~4~~ 3 Results and Discussion

4.13.1 Model Performance

All the tested machine learning algorithms were able to predict soil moisture with good accuracy. Both BRT and RF algorithms however had a slightly better performance with MAE of less than 4 % soil moisture content. Figure 6 shows the ~~Performances~~ performances of the five machine learning algorithms [illegible] The relatively high accuracy performance of BRT and RF models is consistent with other studies that find ensemble decision-tree-based regression models perform better than many other machine learning algorithms (Caruana and Niculescu-Mizil, 2006); particularly in terrain and soil spatial predictions (Hengl et al., 2017, 2018; Keskin et al., 2019; Nussbaum et al., 2018; Szabó et al., 2019). Of the two best algorithms, BRT performs marginally better and for this reason, we use only present variable importance and predictor effect analysis done with the BRT model in the coming sections. Despite being marginally inferior to BRT, ~~The~~ the RF model has several advantages over BRT and the other algorithms. RF is much easier and faster to train compared to the other machine learning algorithms used. Since the ensemble trees are independent in RF model, the ‘forest’ can be grown simultaneously, which dramatically increases processing efficiency in parallel computing. In addition, the RF model has few hyperparameters to tune. In contrast, the ensemble trees in the BRT algorithm must be grown sequentially since each new tree is dependent on the previous ensemble (which makes parallel processing challenging). Training a BRT model requires tuning multiple hyperparameters—seven in our implementation of the BRT model compared to two for the RF model. The performance of individual models across the 30 different training-testing splits was comparable and seems consistent with bias based on individual testing set selection to be minimal.

materials ~~Figure 7~~ Figure S9. ~~On average the ensemble of the BRT models had MAE of 3.77 % across all 30 testing sets. Given that the locations of ground sampling points were clustered in the six transects, selection of testing and training datasets can lead to spatial bias. To minimize such biases, the transects for the testing dataset were randomly for each sampling day. Comparison of the performance of models across the different training-testing splits suggests that the potential bias based on testing set selection was minimal.~~

The ~~individual~~ performances of the individual 30 tuned BRT models on their respective testing sets are shown in supplemental materials ~~Figure 7~~ Figure S9. ~~On average the ensemble of the BRT models had MAE of 3.77 % across all 30 testing sets. Given that the locations of ground sampling points were clustered in the six transects, selection of testing and training datasets can lead to spatial bias. To minimize such biases, the transects for the testing dataset were randomly for each sampling day. Comparison of the performance of models across the different training-testing splits suggests that the potential bias based on testing set selection was minimal.~~

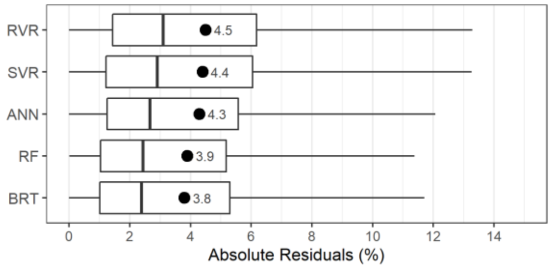

Figure 6 Distribution of residuals and MAE on the testing set by the type of machine learning algorithm. Filled circles and values to their right indicate the average MAE.

Measured versus model predicted soil moisture contents for the testing data sets are plotted in Figure 7. The 1 to 1 comparison in Figure 7 shows the individual predictions by all 30 models. The plot shows a general increase in error at higher soil moisture levels. In addition, the model prediction appears capped around 40 % soil moisture content, this is likely the result of a lack of training data points with values above that soil moisture (see supplemental Figure 5). The marginal boxplot on the y-axis of Figure 7 illustrates this point, values above 40% soil moisture content are over 1.5 times the interquartile range above the upper quartile and are plotted as outliers.

~~**Figure 7 Distribution of residuals and MAE on the testing set for the 30 BRT models. Filled circles indicate MAE and solid vertical line indicates the average MAE overall individual models.**~~

Measured versus model predicted soil moisture contents for the testing data sets are plotted in Figure 7. The 1 to 1 comparison in Figure 7 shows the individual predictions by all 30 models. The plot shows a general increase in error at higher soil moisture levels. In addition, the model prediction appears capped around 40 % soil moisture content, this is likely the result of a lack of training data points with values above that soil moisture (see supplemental Figure 5). The marginal boxplot on the y-axis of Figure 7 illustrates this point, values above 40% soil moisture content are over 1.5 times the interquartile range above the upper quartile and are plotted as outliers.

~~The m~~Measured versus model predicted ~~water~~ soil moisture contents ~~from the~~for the testing data sets ~~are compared~~are plotted in ~~with the prediction by all the trained models in~~ Figure 7~~Figure 8~~. The 1 to 1 comparison in Figure 7 shows the individual predictions by all 30 models. The plot shows a general increase in error at higher soil moisture levels. In addition, the model prediction appears capped around 40 % soil moisture content, this is likely the result of a lack of training data points with values above that soil moisture (see supplemental Figure 5). The marginal boxplot on the y-axis of Figure 7 illustrates this point, values above 40% soil moisture content are over 1.5 times the interquartile range above the upper quartile and are plotted as outliers.

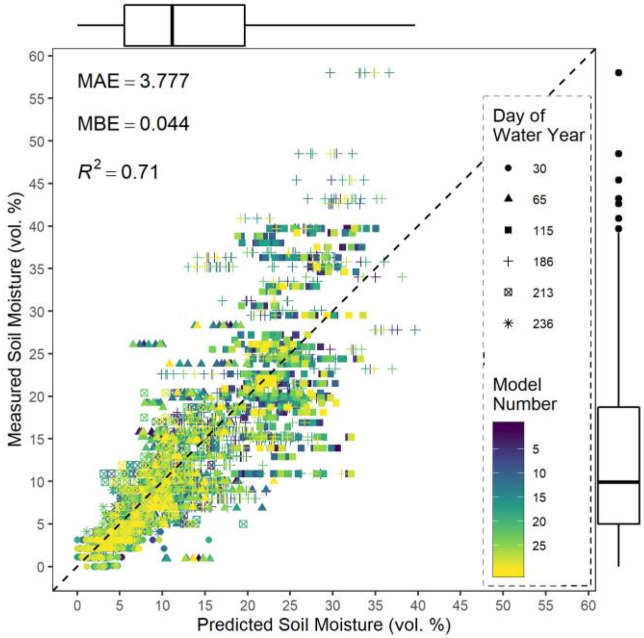

Figure 7 Scatter plot of the measured versus predicted soil moisture content for the testing sets. Marginal boxplots show the distributions of measured and predicted values. MAE, MBE, and R2 are averaged across the 30 models.

4.23.2 Predictor Variable Importance

Variable importance analysis of the models revealed that the hydrologic variables of precipitation and PET were among the top variables in terms of predictive importance. This is to be expected given that these two variables represent the major source and loss pathway for surface soil moisture. Reflectance in the red band was by far the most important of all the reflectance bands. Following these three variables, topographic variables—TPI and curvature in particular—were most important in most models. Topography has a strong control on soil moisture distribution at landscape scales (Sørensen et al., 2006). While the TPI was the most important topographic variable in determining soil moisture. A surprising finding was that TWI was found to be not an important predictor despite numerous studies finding it important in explaining surface soil moisture (e.g. Moore et al., 1988; Western et al., 1999) Despite calculating TWI in multiple ways and at multiple scales, it consistently failed the variable selection procedure for all algorithms we tested. The reasons for this could be that TWI is of less significance at explaining soil moisture distribution at a small scale. Upon observing a similar lack of correlation between TWI and soil moisture, Famigliettiet et al. (1998) suggest that TWI is more appropriate for predicting soil moisture of an entire unsaturated zone profile, and not just the surface layer. Yet another, the reason could be that since slope and flow accumulation, the constituent parts of TWI, are already included in the models it was deemed redundant (not providing unique information) to the models. This would be consistent with the models finding NDVI not important when the constituent bands (red and NIR bands) were found to be important.

It is also worth noting when considering variable importance's that different variable may have different importance depend on how wet or dry the soil condition is. Western et al. (1999), for example, found that flow accumulation was the best predictor for soil moisture distribution during wet conditions while the potential solar radiation index was a better predictor during dry conditions.

The relative importance of predictors for the BRT model is shown in Figure 8. The predictors have been grouped by variable type (lumping the same variables regardless of variable specifications such as summation window for precipitation or pixel resolution for the raster). The only temporally dynamic variables in our model are the hydrological and reflectance variables, the topographic variables are not time-dependent and their variables need only be generated once for a study area. A more detailed variable importance plot is provided in the supplemental materials Figure S10).

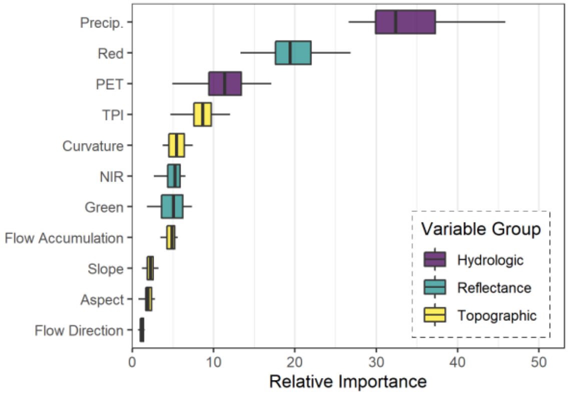

Figure 8 Relative variable importance distribution of the 30 BRT models aggregated by variable type.

4.33.3 Effect of Predictor Variables

We used the ALE plots to investigate the nature of the relationship between the predictor variables and soil moisture. Figure 9 shows the partial effect of red reflectance and three of the most important topographic variables: TPI, profile curvature, and flow accumulation. Given the high importance of these topographic variables, it is useful to understand how these variables relate to soil moisture and identify possible thresholds of significant changes. Predicted soil moisture generally increased with flow accumulation across all scales.

The relationship of curvature to soil moisture is a little more complex, soil moisture tends to decrease as the surface became less convex. However, the trend reverses and soil moisture tends to increase as surface curvature transitioned from convex to a concave surface (approximately between profile curvature values -5 to +5) before the effect reverses again at higher concavity surfaces. A possible explanation for this behavior might be that more flat surfaces (with near 0 curvature value) are associated with higher slope areas, as in immediately following a ridge top which is of higher convexity. The decreasing trend of soil moisture at increasing concavity (profile curvature value above +5) is harder to explain. However, at lower scales. i.e., 3 and 5 and resolution DEM, soil moisture did peak at convex to concave transition (near 0) but there was almost no noticeable decreasing pattern at higher curvature values (concavity) values (not presented in this manuscript). With the lowest resolution (50 m DEM) soil moisture continued to increase with an increase in concavity of surface.

Of all the topographic variables we calculated, perhaps TPI is the most scale dependent variable. Surprisingly, TPI across all scales had a similar relation with soil moisture. Negative TPI values indicate surface that trend towards valleys, zero values indicate flat areas if the slope of the surface is shallow or mid-slope areas. For areas with significant slopes, positive TPI values indicate surfaces that trend towards ridgetops (Jenness et al., 2013). Across all scales, there was a u-shaped relation between TPI and soil moisture with soil moisture decreasing as negative TPI values moved towards zero and soil moisture increasing as TPI moved from zero to positive values. This pattern is consistent with valleys and ridge tops being wetter than mid-slope areas. We computed TPI at several scales, across all machine learning algorithms TPI with 15 and 35 m inner and outer diameters, TPI(15,35), had the highest variable importance among all topographic variables .

~~Topography has a strong control on soil moisture distribution at landscape scales (Sørensen et al., 2006). While the TPI was the most important topographic variable in determining soil moisture, the TWI was found to be not an important predictor.~~ Although the machine learning models are considered non-spatial models, that is, they do not consider sampling location information and spatial autocorrelations (Georganos et al., 2019; Hengl et al., 2018). The inclusion of spatially dependent variables (specifically: curvature, flow accumulation, and TPI) as predictors, however, means that the models do account for a significant amount of spatial information. The inclusion of such variables ~~to an extent and this~~ should make the predictions more spatially relevant.

The red band was the most predictive of the three bands and the red-edge band was found to be not an important predictor. The two spectral vegetation indices we tested, i.e., NDVI and TTVI, were found to be not important but their constituent bands, the red and NIR were important. The lower importance of NIR compared to the red band in the prediction of surface soil moisture was particularly surprising given the higher sensitivity of the NIR band to plant moisture stress and the fact that our study area was almost entirely covered with vegetation.

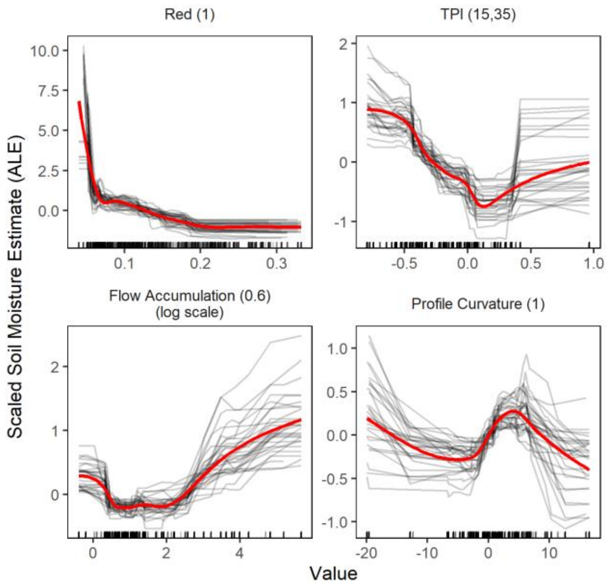

Figure 9 ALE plots for four selected high importance predictor variables. The black curves represent the individual effects of the 30 models, and red curves are smoothed trendlines overall individual models. Marks along the x-axis show the distribution of data in the model training set.

4.43.4 Spatial Prediction of Soil Moisture

The final utility of training a machine learning model was to be able to produce a spatially resolved soil moisture map. In Figure 10, we have predicted surface soil moisture for the six sampling days of which we had multispectral images. Ideally, in the future, the only new inputs required to produce a soil moisture prediction map for our study site is UAV based multispectral images and hydrologic variables of precipitation and PET which are available from nearby weather stations. As can be seen in Figure 10, while the mean moisture content was largely driven by the day (which in turn is controlled by antecedent precipitation and PET), the distribution appears to closely follow topographic attributes. This is particularly visible in a zoomed-in map (Figure 11). Ridges appear drier while valleys appear wetter, furthermore, northern facing ridges appear slightly drier than south-facing slopes. Probably due to slightly higher vegetation cover. Density plots showing the distribution of soil moisture predictions over the test area for each of the six days is provided in supplemental material Figure S11.

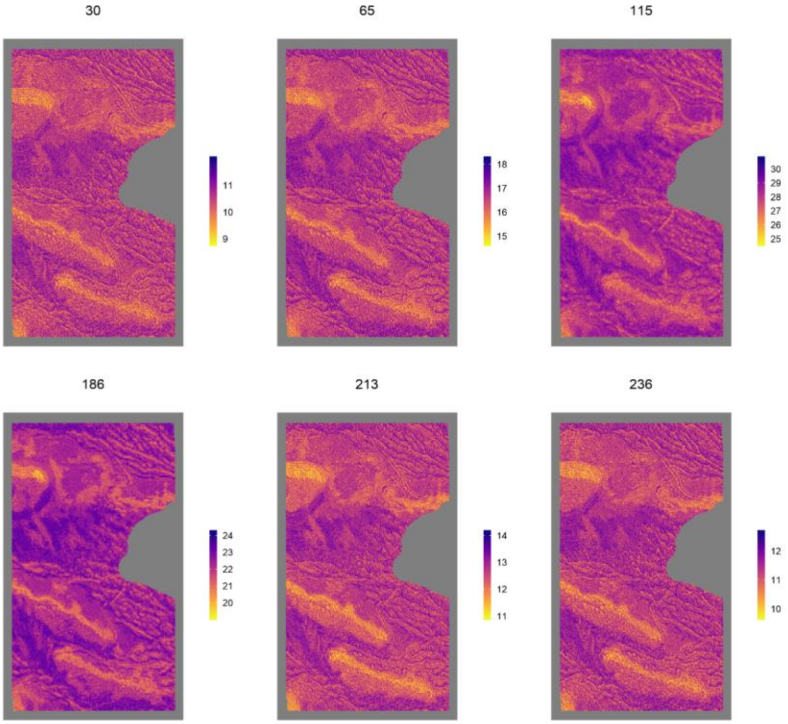

Figure 10 Predicted soil moisture content (%) over the study area for the six days sampled. Days of the water year 30, 65, 115, 186, 213, and 236 are 10/30/2017, 12/4/2017, 1/23/2018, 4/4/2018, 5/1/2018, and 5/24 /2018, respectively).

A close-up map of soil moisture prediction for January 23, 2018, is shown in Figure 11 and shows that soil moisture varies

A close-up map of soil moisture prediction for January 23, 2018, is shown in Figure 11~~Figure 12~~ and shows that soil moisture varies considerably with topography. Tracks made by repeated passage of vehicles are for example clearly identifiable as areas of low soil moisture in the close-up map.~~.~~

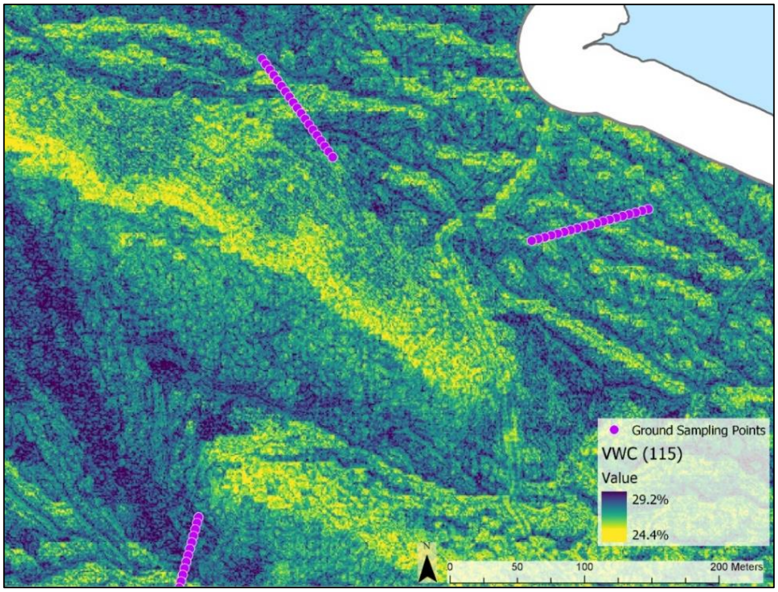

Figure 11 Zoomed-in map of predicted soil moisture content (%) map for water year day 115 (January 23, 2018).

4 Conclusion

Our study addressed the following questions: How effectively can machine learning methods be employed to retrieve soil moisture from a combination of topographic and UAV remote sensing data? What are the most important predictors of surface soil moisture in our study area? And what is the nature of the relationship between predictor variables and soil moisture? Our approach can be summarized as follows: We took multispectral images of grassland in the visible-near-infrared range using a UAS. Using photogrammetric analysis of the images we produced a high-resolution digital elevation model of the study site which we used to calculate several topographic variables at different scales. Simultaneously with the UAS imaging flights we took about 400 in situ surface soil moisture measurements. Using those ground truth measurements, we trained machine learning models to predict soil moisture from the multispectral images, topographic variables as well as precipitation and evapotranspiration data. We finally interrogated the machine learning models to understand the importance of the different variables and elucidate the nature of the relationship between variables and soil moisture.

~~5~~

What makes our study stand out is that we used UAS based remote sensing to investigate soil water outside of the relatively homogenous farm plots. Our study site had uneven topography and an ecology dominated by large seasonal shifts in moisture. The fact that it is neither bare nor a monoculture makes the interpretation of multispectral images challenging and justifies the use of machine learning methods that can generalize complex relationships. Our findings showed that it is possible to use machine learning to interpret soil moisture from surface characteristics recorded from a UAS. This research serves as a proof of concept that surface soil moisture can be interpreted with reasonable accuracy from multispectral UAS remote sensing using machine learning methods.

Based on our findings we conclude that all the popular machine learning algorithms we tested (ANN, SVR, RVR, RF, and BRT) are acceptable for modeling surface soil moisture with the conditions of our study. We particularly found that the decision tree-based methods, and particularly RF to be a versatile algorithm. Even though only provides a marginal improvement in prediction compared to the other algorithms is particularly simple to train and requires the least amount of computational resource.

Analysis of the models revealed that hydrologic variables of precipitation and evapotranspiration are the most important predictor variables for surface soil moisture. However spatial distribution of soil moisture is highly dependent on topographic variables. Topographic variables such as flow accumulation and TPI are particularly important in ameliorating the non-spatial nature of such machine learning models by including spatial variables. Based on our study, a DEM of about 1 m resolution is sufficient. Although we had resolution as high as 15 cm, we found that excluding such high-resolution topographic variables did not substantially reduce the model performances. Performance.

Partial dependence analysis of how the input variables relate to soil moisture is important in understanding the mechanisms that control soil moisture over the landscape. Using a type of partial dependency plots we were able to investigate the nature of the relationships. Surface curvature showed a complex relationship with soil moisture. There was a negative relationship as surface curvature transition moved from more convex to less convex. In the middle ranges where convex surface transitions to upward concave surface soil moisture increase as expected. However, this relation reverses at higher concavities. The partial effect of the curvature variable is such that the mildly concave surfaces tend to be wetter than more concave surfaces. However, the magnitude of changes associated with curvature was small. TPI was the most important of the topographic variables and the partial dependency plots in the form of ALE (Figure 9) show that it affects soil moisture at a relatively higher magnitude. The general nature of the relation can be summarized as that valleys and ridgetops as defined by TPI tend to be wetter surfaces compared to areas in between. Valleys also tend to be moister compared to ridge tops.

5 Outlook

As a data mining technique, machine learning model performance and reliability are closely tied to the quantity of data. Although the number and spatial coverage of ground sampling points were reasonable for the purposes of our research, future studies would benefit from significantly increasing the number of ground measurements as well as flight frequency.

Multi-year studies are eventually needed to ensure that the model can be used reliably for future predictions. As data-based methods, machine learning methods are highly unreliable if used to forecast situations they have not been trained on.

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Another important consideration for future studies is to move beyond surface soil moisture to deeper layers. Although more challenging to implement, studies of deeper soil root-zone soil moisture are more ecologically relevant in Although more challenging to implement, studies of deeper soil root-zone soil moisture are more ecologically relevant in many. Expanding the reflectance information beyond multispectral bands could lead to important improvements in soil moisture prediction. Although lightweight thermal or hyperspectral sensors are currently very expensive and may not be financially feasible for routine applications at this time, those technologies might become more affordable in the future.

The possibility of using the high-resolution topographic variables from UAS with cheaper and more widely available ~~reflectance data from~~ satellite images~~remote sensing~~ is an interesting ~~topic~~ avenue worth investigating ~~to study. This would mean the UAS would only be flown once in an area and future and past predictions of soil moisture can be estimated using satellite images and meteoric data from stations. This would also be ideal to downscale satellite data by integrating high-resolution topographic information from UAS remote sensing.~~ Although the machine learning models are high performing and generalizable models, they are non-spatial models that do not consider sampling location information and spatial autocorrelations. This can potentially compromise the model's ability to appropriately address spatial heterogeneity (Georganos et al., 2019; Hengl et al., 2018). Hengl et al. (2018) introduce a novel method to incorporate spatial information into a non-spatial machine learning model by including distances between sampling points as predictor variables, and they show this method (although still in its formative stage) has comparable accuracy to kriging methods. The potential to improve soil moisture predictions by using such spatially integrated methods should be considered in future research.

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Field Code Changed

Field Code Changed

Acknowledgments

This work was made possible by the support of U.S. Fish and Wildlife Service Agreement #P1740401 as administered by the California Department of Fish and Wildlife; Monique Kolster and the UC Merced Vernal Pools and Grassland Reserve of the

Natural Reserve System; and Francesca Cannizzo of the UC Merced Physical and Environmental Planning team. We gratefully acknowledge computing time on the Multi-Environment Computer for Exploration and Discovery (MERCED) cluster at UC Merced, which was funded by National Science Foundation Grant No. ACI-1429783.

References

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

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M. and Notarnicola, C.: Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data, Remote Sens., 7(12), 16398–16421, doi:10.3390/rs71215841, 2015.

Ambroise, C. and McLachlan, G. J.: Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci., 99(10), 6562–6566, doi:10.1073/pnas.102102699, 2002.

Anderson, K. and Gaston, K. J.: Lightweight unmanned aerial vehicles will revolutionize spatial ecology, Front. Ecol. Environ., 11(3), 138–146, doi:10.1890/120150, 2013.

Apley, D. W.: Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. [online] Available from: https://arxiv.org/ftp/arxiv/papers/1612/1612.08468.pdf (Accessed 13 June 2019), 2016.

Barrett, B. W. and Petropoulos, G. P.: Satellite Remote Sensing of Surface Soil Moisture, in Remote Sensing of Energy Fluxes and Soil Moisture Content, edited by G. P. Petropoulos, pp. 85–120, CRC Press., 2014.

Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Taylor, G. R., Hill, J., Whiting, M. L. and Sommer, S.: Using Imaging Spectroscopy to study soil properties, Remote Sens. Environ., 113(SUPPL. 1), S38–S55, doi:10.1016/j.rse.2008.09.019, 2009.

Ben-Shimon, D. and Shmilovici, A.: Kernels for the Relevance Vector Machine - An Empirical Study, in Advances in Web Intelligence and Data Mining, edited by M. Last, P. S. Szczepaniak, Z. Volkovich, and A. Kandel, pp. 253–263, Springer-Verlag GmbH., 2006.

Berni, J. A. J., Zarco-Tejada, P. J., Suárez, L., González-Dugo, V. and Fereres, E.: Remote sensing of vegetation from UAV platforms using lightweight multispectral and thermal imaging sensors, Int. Arch. Photogramm. Remote Sens. Spat. Inform. Sci, 38, 6 pp, doi:10.1007/s11032-006-9022-5, 2009.

Breiman, L.: Random Forest, Mach. Learn., 45(1), 5–32, doi:10.1023/A:1010933404324, 2001.

Brownlee, J.: Better Deep Learning Train Faster, Reduce Overfitting, and Make Better Predictions, 1.8., Machine Learning Mastery Pty., 2020.

California Department of Water Resources: UC Merced Weather Station, [online] Available from: http://cdec.water.ca.gov/dynamicapp/staMeta?station_id=UCM (Accessed 1 February 2019), 2018.

California Irrigation Management Information System: CIMIS Station Report, [online] Available from: https://cimis.water.ca.gov/WSNReportCriteria.aspx (Accessed 1 February 2019), 2018.

California Irrigation Management Information System: Data Overview, [online] Available from: https://cimis.water.ca.gov/Resources.aspx (Accessed 19 April 2019), n.d.

Caruana, R. and Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms, in Proceedings of the 23rd international conference on Machine learning - ICML ’06, pp. 161–168, ACM Press, New York, New York, USA., 2006.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining - KDD ’16, pp. 785–794, ACM Press, New York, New York, USA., 2016.

Chen, X. and Jeong, J. C.: Enhanced Recursive Feature Elimination, Proc. - 6th Int. Conf. Mach. Learn. Appl. ICMLA 2007, 330–335, doi:10.1109/ICMLA.2007.35, 2007.

Colomina, I. and Molina, P.: Unmanned aerial systems for photogrammetry and remote sensing: A review, ISPRS J. Photogramm. Remote Sens., 92, 79–97, doi:10.1016/j.isprsjprs.2014.02.013, 2014.

Cortes, C. and Vapnik, V. N.: Support-Vector Networks, Mach. Learn., 20(3), 273–297, doi:10.1023/A:1022627411411, 1995.

Das, N. N. and Mohanty, B. P.: Root Zone Soil Moisture Assessment Using Remote Sensing and Vadose Zone Modeling, Vadose Zo. J., 5(1), 296, doi:10.2136/vzj2005.0033, 2006.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V. N.: Support vector regression machines, in Advances in Neural Information Processing Systems 9, vol. 1, edited by M. C. Mozer, M. I. Jordan, and T. Petsche, pp. 155–161, MIT Press. [online] Available from: http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf, 1997.

Elarab, M.: The Application of Unmanned Aerial Vehicle to Precision Agriculture: Chlorophyll, Nitrogen, and Evapotranspiration Estimation, Utah State University., 2016.

Elith, J., Leathwick, J. R. and Hastie, T.: A working guide to boosted regression trees, J. Anim. Ecol., 77(4), 802–813, doi:10.1111/j.1365-2656.2008.01390.x, 2008.

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

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat., 29(5), 1189–1232, doi:10.1017/CBO9781107415324.004, 2001.

Friedman, J. H.: Stochastic gradient boosting, Comput. Stat. Data Anal., 38(4), 367–378, doi:10.1016/S0167-9473(01)00065-2, 2002.

Fryjoff-Hung, A. F.: 2D Hydrodynamic Modeling for Evaluating Restoration Potential of a Vernal Pool Complex, University of California, Merced., 2018.

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N. O., Wolff, E. and Kalogirou, S.: Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling, Geocarto Int., 1–12, doi:10.1080/10106049.2019.1595177, 2019.

Greenwell, B. M.: pdp: An R Package for Constructing Partial Dependence Plots, R J., 9(1) [online] Available from: https://github.com/bgreenwell/pdp/issues. (Accessed 13 June 2019), 2017.

Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection, J. Mach. Learn. Res., 3(3), 1157–1182, 2003.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V.: Gene selection for cancer classification using support vector machines, Mach. Learn., 389–422, doi:10.1023/A:1012487302797, 2002.

Hassan-Esfahani, L., Torres-Rua, A., Jensen, A. and McKee, M.: Assessment of surface soil moisture using high-resolution multi-spectral imagery and artificial neural networks, Remote Sens., 7(3), 2627–2646, doi:10.3390/rs70302627, 2015.

Hastie, T., Tibshirani, R. and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer New York, New York, NY., 2009.

Haubrock, S.-N., Chabrillat, S., Lemmnitz, C. and Kaufmann, H.: Surface soil moisture quantification models from reflectance data under field conditions, Int. J. Remote Sens., 29(1), 3–29, doi:10.1080/01431160701294695, 2008.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, edited by B. Bond-Lamberty, PLoS One, 12(2), e0169748, doi:10.1371/journal.pone.0169748, 2017.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M. and Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 6, e5518, doi:10.7717/peerj.5518, 2018.

Hillel, D.: Environmental Soil Physics, Academic Press, San Diego, CA., 1998.

James, G., Witten, D., Hastie, T. and Tibshirani, R.: An Introduction to Statistical Learning, Springer New York, New York, NY., 2013.

Jenness, J., Brost, B. and Beier, P.: Land Facet Corridor Designer, , (July) [online] Available from: www.corridordesign.org, 2013.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A.: kernlab: An S4 Package for Kernel Methods in R (version 0.9-25), J. Stat. Softw., 11(9), 1–20 [online] Available from: http://www.jstatsoft.org/v11/i09/, 2004.

Keskin, H., Grunwald, S. and Harris, W. G.: Digital mapping of soil carbon fractions with machine learning, Geoderma, 339, 40–58, doi:10.1016/j.geoderma.2018.12.037, 2019.

Korres, W., Reichenau, T. G., Fiener, P., Koyama, C. N., Bogena, H. R., Cornelissen, T., Baatz, R., Herbst, M., Diekkrüger, B., Vereecken, H. and Schneider, K.: Spatio-temporal soil moisture patterns – A meta-analysis using plot to catchment scale data, J. Hydrol., 520, 326–341, doi:10.1016/j.jhydrol.2014.11.042, 2015.

Kuhn, M.: The caret Package, [online] Available from: http://topepo.github.io/caret/index.html (Accessed 1 June 2017), 2017.

Liaw, A. and Wiener, M.: Package “randomForest” (v.4.6-12), R Packag., XXXIX(1), doi:10.5244/C.22.54, 2015.

Malley, D. F., Martin, P. D. and Ben-Dor, E.: Application in Analysis of Soils, in Near-Infrared Spectroscopy in Agriculture, edited by R. Craig, R. Windham, and J. Workman, pp. 729–784, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, USA., 2004.

Manfreda, S., McCabe, M., Miller, P., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., Ben Dor, E., Helman, D., Estes, L., Ciraolo, G., Müllerová, J., Tauro, F., de Lima, M., de Lima, J., Maltese, A., Frances, F., Caylor, K., Kohv, M., Perks, M., Ruiz-Pérez, G., Su, Z., Vico, G. and Toth, B.: On the Use of Unmanned Aerial Systems for Environmental Monitoring, Remote Sens., 10(4), 641, doi:10.3390/rs10040641, 2018.

Matei, O., Rusu, T., Petrovan, A. and Mihuț, G.: A Data Mining System for Real Time Soil Moisture Prediction, Procedia Eng., 181, 837–844, doi:10.1016/j.proeng.2017.02.475, 2017.

Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. [online] Available from: https://christophm.github.io/interpretable-ml-book/, 2019.

Moore, I. D., Burch, G. J. and Mackenzie, D. H.: Topographic Effects on the Distribution of Surface Soil Water and the Location of Ephemeral Gullies, Trans. ASAE, 31(4), 1098–1107, doi:10.13031/2013.30829, 1988.

Muller, E. and Décamps, H.: Modeling soil moisture–reflectance, Remote Sens. Environ., 76(2), 173–180, doi:10.1016/S0034-4257(00)00198-X, 2000.

Natekin, A. and Knoll, A.: Gradient boosting machines, a tutorial, Front. Neurorobot., 7(DEC), 1–11, doi:10.3389/fnbot.2013.00021, 2013.

Nichols, S., Zhang, Y. and Ahmad, A.: Review and evaluation of remote sensing methods for soil-moisture estimation, J. Photonics Energy, 2(1), 028001, doi:10.1117/1.3534910, 2011.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E. and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, 4(1), 1–22, doi:10.5194/soil-4-1-2018, 2018.

Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. a., Draper, C. S., Hagimoto, Y., Kerr, Y. H., Njoku, E. G., Small, E. E. and Zreda, M.: State of the Art in Large-Scale Soil Moisture Monitoring, Soil Sci. Soc. Am. J., 77(6), 1888, doi:10.2136/sssaj2013.03.0093, 2013.

Pachepsky, Y. A., Timlin, D. and Varallyay, G.: Artificial Neural Networks to Estimate Soil Water Retention from Easily Measurable Data, Soil Sci. Soc. Am. J., 60(3), 727–733, doi:10.2136/sssaj1996.03615995006000030007x, 1996.

Paloscia, S., Pampaloni, P., Pettinato, S. and Santi, E.: A Comparison of Algorithms for Retrieving Soil Moisture from ENVISAT/ASAR Images, IEEE Trans. Geosci. Remote Sens., 46(10), 3274–3284, doi:10.1109/TGRS.2008.920370, 2008.

Petropoulos, G. P., Ireland, G. and Barrett, B. W.: Surface soil moisture retrievals from remote sensing: Current status, products & future trends, Phys. Chem. Earth, 83–84, 36–56, doi:10.1016/j.pce.2015.02.009, 2015.

Pix4D: Learn more about Sequoia, [online] Available from: https://pix4d.com/sequoia-faq/ (Accessed 14 November 2016), n.d.

Price, J. C.: Thermal inertia mapping: A new view of the earth, J. Geophys. Res., 82, 2582–2590, 1977.

Price, J. C.: On the analysis of thermal infrared imagery: The limited utility of apparent thermal inertia, Remote Sens. Environ., 18(1), 59–73, doi:10.1016/0034-4257(85)90038-0, 1985.

Rana, G. and Katerji, N.: Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review, Eur. J. Agron., 13(2–3), 125–153, doi:10.1016/S1161-0301(00)00070-8, 2000.

Reed, R. and Marks II, R. J.: Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, MIT Press, Cambridge, Massachusetts., 1999.

Ridgeway, G.: Generalized Boosted Models: A guide to the gbm package., 2012.

Schaap, M. G., Leij, F. J. and van Genuchten, M. T.: Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, J. Hydrol., 251(3–4), 163–176, doi:10.1016/S0022-1694(01)00466-8, 2001.

Schanda, E.: Physical Fundamentals of Remote Sensing, Springer-Verlag, Berlin, Germany., 1986.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B. and Teuling, A. J.: Investigating soil moisture-climate interactions in a changing climate: A review, Earth-Science Rev., 99(3–4), 125–161, doi:10.1016/j.earscirev.2010.02.004, 2010.

Singh, K. K. and Frazier, A. E.: A meta-analysis and review of unmanned aircraft system (UAS) imagery for terrestrial applications, Int. J. Remote Sens., 39(15–16), 5078–5098, doi:10.1080/01431161.2017.1420941, 2018.

Sobrino, J., Mattar, C., Jiménez-Muñoz, J. C., Franch, B. and Corbari, C.: On the Synergy between Optical and TIR Observations for the Retrieval of Soil Moisture Content, in Remote Sensing of Energy Fluxes and Soil Moisture Content, edited by G. P. Petropoulos, pp. 363–390, CRC Press., 2014.

Sørensen, R., Zinko, U. and Seibert, J.: On the calculation of the topographic wetness index: Evaluation of different methods based on field observations, Hydrol. Earth Syst. Sci., 10(1), 101–112, doi:10.5194/hess-10-101-2006, 2006.

Spectrum Technologies Inc.: FieldScout TDR 300 Soil Moisture Meter, 2009.

Stark, B., McGee, M. and Chen, Y.: Short wave infrared (SWIR) imaging systems using small Unmanned Aerial Systems (sUAS), in 2015 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 495–501, IEEE., 2015.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K. and Pásztor, L.: Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics, Hydrol. Earth Syst. Sci., 23(6), 2615–2635, doi:10.5194/hess-23-2615-2019, 2019.

Tian, J., Su, H., He, H. and Sun, X.: An Empirical Method of Estimating Soil Thermal Inertia, Adv. Meteorol., 2015, 1–9, doi:10.1155/2015/428525, 2015.

Tipping, M. E.: The Relevance Vector Machine, in Advances in Neural Information Processing Systems 12, edited by S. A. Solla, T. K. Leen, and K. Muller, pp. 652–658, MIT Press. [online] Available from: http://papers.nips.cc/paper/1719-the-relevance-vector-machine.pdf, 2000.

Tmušić, G., Manfreda, S., Aasen, H., James, M. R., Gonçalves, G., Ben-Dor, E., Brook, A., Polinova, M., Arranz, J. J., Mészáros, J., Zhuang, R., Johansen, K., Malbeteau, Y., de Lima, I. P., Davids, C., Herban, S. and McCabe, M. F.: Current Practices in UAS-based Environmental Monitoring, Remote Sens., 12(6), 1001, doi:10.3390/rs12061001, 2020.

Torres-Rua, A., Ticlavilca, A., Bachour, R. and McKee, M.: Estimation of Surface Soil Moisture in Irrigated Lands by Assimilation of Landsat Vegetation Indices, Surface Energy Balance Products, and Relevance Vector Machines, Water, 8(4), 167, doi:10.3390/w8040167, 2016.

Trenberth, K. E., Fasullo, J. T. and Kiehl, J.: Earth’s Global Energy Budget, Bull. Am. Meteorol. Soc., 90(3), 311–323, doi:10.1175/2008BAMS2634.1, 2009.

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

Weidong, L., Baret, F., Xingfa, G., Qingxi, T., Lanfen, Z. and Bing, Z.: Relating soil surface moisture to reflectance, Remote Sens. Environ., 81(2–3), 238–246, doi:10.1016/S0034-4257(01)00347-9, 2002.

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

Weston, J., Elisseeff, A., Schölkopf, B. and Tipping, M.: Use of the Zero-Norm with Linear Models and Kernel Methods, J. Mach. Learn. Res., 3, 1439–1461, 2003.

Whiting, M. L., Li, L. and Ustin, S. L.: Predicting water content using Gaussian model on soil spectra, Remote Sens. Environ., 89(4), 535–552, doi:10.1016/j.rse.2003.11.009, 2004.

Willmott, C. and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res., 30(1), 79–82, doi:10.3354/cr030079, 2005.

Wilson, J. P. and Gallant, J. C.: Digital Terrain Analysis, in Terrain Analysis: principles and applications, edited by J. P. Wilson and J. C. Gallant, pp. 1–21., 2000.

Wong, K.: Merced Vernal Pools Joins Natural Reserve System, Univ. Calif. News, 22nd January [online] Available from: http://universityofcalifornia.edu/news/merced-vernal-pools-join-natural-reserve-system, 2014.

Zaman, B. and Mckee, M.: Spatio-Temporal Prediction of Root Zone Soil Moisture Using Multivariate Relevance Vector Machines, Open J. Mod. Hydrol., 4(3), 80–90, doi:dx.doi.org/10.4236/ojmh.2014.43007, 2014.

Zaman, B., McKee, M. and Neale, C. M. U.: Fusion of remotely sensed data for soil moisture estimation using relevance vector and support vector machines, Int. J. Remote Sens., 33(20), 6516–6552, doi:10.1080/01431161.2012.690540, 2012.

Zhang, D. and Zhou, G.: Estimation of Soil Moisture from Optical and Thermal Remote Sensing: A Review, Sensors, 16(8), 1308, doi:10.3390/s16081308, 2016.

Zhang, Y. and Schaap, M. G.: Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3), J. Hydrol., 547, 39–53, doi:10.1016/j.jhydrol.2017.01.004, 2017.

Zhang, Y., Schaap, M. G. and Zha, Y.: A High-Resolution Global Map of Soil Hydraulic Properties Produced by a Hierarchical Parameterization of a Physically Based Water Retention Model, Water Resour. Res., 54(12), 9774–9790, doi:10.1029/2018WR023539, 2018.

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T. and Rosolem, R.: COSMOS: the COsmic-ray Soil Moisture Observing System, Hydrol. Earth Syst. Sci., 16(11), 4079–4099, doi:10.5194/hess-16-4079-2012, 2012.