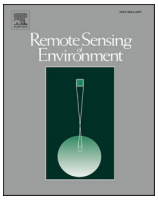
Remote Sensing of Environment 266 (2021) 112706

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse



Multi-layer high-resolution soil moisture estimation using machine learning over the United States

L. Karthikeyan [a,b], Ashok K. Mishra [b,*]

[a] Centre of Studies in Resources Engineering, IIT Bombay, Powai, Mumbai, India
[b] Department of Civil Engineering, Clemson University, Clemson, SC, USA

ARTICLE INFO

Editor: Jing M. Chen

Keywords:
Soil moisture
SMAP
Microwave remote sensing
Rootzone soil moisture
Multi-layer soil moisture
Machine learning

ABSTRACT

The lack of proper understanding of multi-layer soil moisture (SM) profile (signals) remains a persistent challenge in sustainable agricultural water management and food security, especially during drought conditions. We develop a machine-learning algorithm using the concept of *learning from patterns* to estimate the multi-layer SM information in ungauged locations firmly based on local knowledge of the climatic and landscape controls. The Contiguous United States (CONUS) is clustered into homogeneous regions based on the association between SM and climate and landscape controls. Extreme Gradient Boosting (XGBoost) algorithm is applied to homogenous regions to capture the complex relationship between appropriate predictor variables and in-situ SM at multiple layers over the CONUS. Soil Moisture Active Passive (SMAP) Level 4 (L4) surface (0–5 cm) and rootzone (0–100 cm) SM along with climate and landscape datasets are used as predictor variables. In-situ multi-layer SM recorded by Soil Climate Analysis Network (SCAN), Snow Telemetry (SNOTEL), and U.S. Climate Reference Network (USCRN) networks are utilized as predictands. XGBoost models are then trained region-wise and layer-wise to estimate multi-layer SM information at 5, 10, 20, 50, and 100 cm depths (five layers) at 1-km spatial resolution. Results indicate that the predictor variables have varying levels of influence on SM with changing soil depth, and meteorological variables have the least importance. Validation at 79 independent locations indicates the multi-layer SM estimates successfully capture temporal dynamics of SM, with most locations achieving ubRMSE less than 0.04 m^3/m^3. The high-resolution SM estimates offer spatial sub-grid heterogeneity compared to SMAP L4 SM.

1. Introduction

Soil Moisture (SM) quantifies the amount of water present in the pore spaces of the soil medium. SM plays an important role in studying land-atmosphere interactions (Seneviratne et al., 2010), numerical weather prediction models (Dirmeyer and Halder, 2016), land surface models (Koster et al., 2009), agriculture (Ma et al., 2013), irrigation assessment (Felfelani et al., 2018; Lawston et al., 2017), and monitoring of floods (Kim et al., 2019; Massari et al., 2015; Parinussa et al., 2016; Rahman et al., 2019) and droughts (Chawla et al., 2020; Mishra et al., 2017; Velpuri et al., 2016). Despite its wide-ranging applications, measuring SM locally at fine spatial scales over a large domain is challenging. This is due to the heterogeneous and higher variability of SM magnitudes across different landscapes (Karthikeyan et al., 2020). Besides, financial constraints are associated with establishing an in-situ SM network, and it is often challenging to capture adequate spatial coverage (Karthikeyan and Kumar, 2016).

Satellite remote sensing of SM is an efficient way of measuring soil moisture at large spatial scales. The microwave frequencies (typically less than 12 GHz) are sensitive to the dielectric property of soil, which is influenced by SM's variability. Therefore, microwave sensors equipped with L-, C-, and X- bands can measure SM. Microwave sensors are categorized into active (radar) and passive (radiometer) microwave sensors. Due to the difference in the acquisition, these sensors' resolution characteristics vary significantly (Karthikeyan et al., 2017a; Ulaby and Long, 2015). Currently, satellite sensors such as Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010a), Soil Moisture Ocean Salinity (SMOS) (Al Bitar et al., 2017; Kerr et al., 2001; Wigneron et al., 2021; Zhang et al., 2021b), Advanced Microwave Scanning Radiometer 2 (AMSR2) (Fujii et al., 2009; Koike et al., 2004; Parinussa et al., 2015), and Sentinel missions, among others, can retrieve SM at the global scale. These sensors measure the microwave radiations backscattered/emitted

* Corresponding author.
E-mail addresses: karthikl@iitb.ac.in (L. Karthikeyan), ashokm@g.clemson.edu (A.K. Mishra).

https://doi.org/10.1016/j.rse.2021.112706
Received 26 February 2021; Received in revised form 7 September 2021; Accepted 11 September 2021
Available online 23 September 2021

from the near-surface. Therefore, SM retrieved by these sensors would correspond to the top few centimetres of soil (typically ~5 cm) (Karthikeyan et al., 2017b). SM at rootzone at large spatial scales is generally estimated using a data assimilation scheme, wherein satellite surface SM retrievals are assimilated into a land surface model to update surface and rootzone SM states (Lievens et al., 2017; Sabater et al., 2007). The operational rootzone SM products are currently produced globally by assimilating SMAP SM into Catchment Land Surface Model (CLSM) (Reichle et al., 2019). Although low-frequency passive microwave sensors are useful to determine soil moisture, the spatial scales are limited to tens of kilometers and are often characterized by considerable uncertainty, which is unsuitable for agricultural water management that requires finer/local scale information.

Efforts are made to disaggregate the SM information to improve their applications by different stakeholders at localized scales (Brown et al., 2013). High resolution SM information is currently being utilized for drought monitoring (Gavahi et al., 2020; Fang et al., 2021), irrigation mapping (Dari et al., 2021) etc. The SMAP mission was launched to provide such high-resolution SM products by combining the retrievals from active and passive sensors. However, the failure of SMAP's radar instrument in 2015 resulted in the usage of several alternate sensors such as Copernicus Sentinel-1C-band radar along with SMAP radiometer in merging algorithms to obtain high-resolution SM (Das et al., 2019). Significant efforts are made to downscale passive microwave SM retrievals using radar backscatter measurements (Das et al., 2018; Das et al., 2010; Das et al., 2013; Piles et al., 2009; Wu et al., 2016). Optical and thermal sensors have the advantage of producing high-resolution maps. Given their physical relationship with SM, products such as Normalized Difference Index (NDVI) and Land Surface Temperature (LST) obtained from satellite sensors such as Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS), are widely used for SM disaggregation (Fang et al., 2018; Merlin et al., 2010; Peng et al., 2015; Piles et al., 2014). Attempts have been made to disaggregate SM using data assimilation (Hoeben and Troch, 2000; Lievens et al., 2017; Sahoo et al., 2013). Recently, machine learning techniques have gained popularity in the areas of gap filling and disaggregation of SM (Abbaszadeh et al., 2019; Abowarda et al., 2021; Fang and Shen, 2020; Kovačević et al., 2020; Liu et al., 2020a; Long et al., 2019; Mao et al., 2019). For SM downscaling, machine learning techniques typically use optical/thermal data and static geomorphological data (available at high resolution) as predictors (Abbaszadeh et al., 2019; Kovačević et al., 2020; Liu et al., 2020a; Long et al., 2019). A comprehensive review of downscaling techniques can be obtained from Peng et al. (2017) and Sabaghy et al. (2018).

It is important to note that the current efforts have focused on disaggregating surface SM, which corresponds to the top 5 cm of the soil layer. Few attempts are made to estimate rootzone SM at high resolution (Bablet et al., 2020; Dumedah et al., 2015; Merlin et al., 2006; Montaldo and Albertson, 2003). Attempts are made to estimate rootzone SM from surface measurements using an exponential filter (Albergel et al., 2008; Ford et al., 2014; Stefan et al., 2021; Tobin et al., 2017) and lagged soil moisture aggregation (Pal and Maity, 2019). Few studies applied machine learning models such as Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) model to estimate multi-layer SM up to 50 cm depth (Kornelsen and Coulibaly, 2014; O and Orth, 2020; Pan et al., 2017). Liu et al. (2016) estimated SM at multiple layers using a combination of Support Vector Machines (SVM) and Ensemble Kalman Filter (EnKF).

The above studies are either limited to laboratory/point scale or coarse spatial resolution or consider rootzone as a single layer, thus having a single SM value representing the entire layer. The root depth of plants varies according to the species and growth stage. Plants do not extract water uniformly throughout the root depth. For instance, the proportion of water extracted by corn plant's roots is divided into four quarters of 40%, 30%, 20%, and 10% with respect to four quarters of rootzone depth (Kranz et al., 2008). Given such an uneven distribution of water extraction, there is a need to determine SM's vertical distribution to accurately assess irrigation water requirements (Mishra et al., 2015). In heterogeneous agriculture landscapes, SM's vertical distribution is needed at a higher spatial resolution to cater to the spatial variability of SM at different layers under changing crop conditions. Besides, the vertical variability of SM is also influenced by the heterogeneity in soil texture.

In summary, the current efforts on SM estimation focus either on a) disaggregation of surface SM using data from active microwave/optical/thermal sensors, aided by data assimilation or machine learning techniques; or b) high-resolution rootzone SM estimation at multiple levels at point/laboratory scale; or c) high-resolution rootzone SM estimation at a large scale but coarse vertical resolution. We find that the large-scale estimation of high-resolution SM at multiple layers remains a grey area of research. Such SM profile estimates are essential for an accurate assessment of agricultural droughts and crop water management – a vital aspect in heterogeneous, fragmented agriculture systems (Mishra et al., 2015). We followed a four-step approach to estimate and validate multi-layer SM information over CONUS: (i) identify the homogeneous regions that explain the SM profile's statistical properties, (ii) apply the machine learning models to region-wise and layer-wise, (iii) assess the relative importance of predictors for SM estimation, and (iv) validate the spatio-temporal multi-layer SM estimates. Overall, we aim to address the following research questions:

1) How do the homogeneous regions vary over the CONUS based on the climate, landscape, soil, and vegetation characteristics that explain SM profile variability?

2) Can the machine learning models trained at in-situ locations within a region help us estimate multi-layer SM at ungauged locations?

3) Are the machine learning models helpful to infer dominant predictor variables that estimate SM across multiple soil layers? What is the impact of scaling on the relative importance of predictor variables?

4) Can the assimilated rootzone SM products and other predictors depict the temporal variations of multi-layer SM while maintaining high spatial resolution (sub-grid heterogeneity)?

The remainder of the manuscript is structured as follows. Section 2 presents a description of the data and methodology proposed in this paper. Section 3 presents the results and discussion. Section 4 presents important conclusions drawn from this work and the future scope.

2. Data and methods

2.1. Study area and datasets used

The proposed method is applied to the CONUS using the dense SM stations from three networks, SCAN, SNOTEL, and USCRN (Fig. S1 presents the station locations). Several of these stations collect multi-layer SM information. The analysis is carried out at a daily scale covering the period 31 March 2015 to 29 February 2019 (1431 days). Soil texture, elevation, vegetation, land surface temperature, precipitation, and SMAP Level 4 surface and rootzone SM products are used as predictors in a machine learning framework to estimate multi-layer high-resolution SM information. The additional information for these selected data sets is provided in Table 1. These data sets are regridded at 1 km resolution corresponding to the MODIS grid system to achieve uniform spatial resolution. The SM time series obtained from multiple in-situ stations present within a 1 km grid cell are averaged to obtained representative SM. As a result, 695 grid locations with 1 km resolution are generated for the CONUS. A detailed description of the usage of these datasets is presented in the methodology section.

2.2. Methodology

The SM temporal dynamics and spatial heterogeneity are controlled by geomorphological, topographical, meteorological, and vegetation characteristics. This study applies a machine-learning algorithm to

Table 1
List of datasets used in this study (Glossary: NSIDC – National Snow and Ice Data Center; USDA STATSGO – United States Department of Agriculture State Soil Geographic Database; DEM – Digital Elevation Model; USGS GTOPO30 – United States Geological Survey Global 30 Arc-Second Elevation; MODIS – Moderate Resolution Imaging Spectroradiometer; NDVI – Normalized Difference Vegetation Index; EVI – Enhanced Vegetation Index; GPP – Gross Primary Productivity; EOSDIS – Earth Observing System Data and Information System; LP DAAC – Land Processes Distributed Active Archive Center; SCAN – Soil Climate Analysis Network; SNOTEL – Snow Telemetry; USCRN – U.S. Climate Reference Network; ISMN – International Soil Moisture Network; CHIRPS – Climate Hazards Group InfraRed Precipitation with Station).

Dataset	Details	Source	Spatial Resolution	Temporal Resolution	Reference
SMAP Level 4 Soil Moisture	Version 4: Vv4030; Surface (0–5 cm) and Rootzone (0–100 cm) soil moisture products	NSIDC	9 km	3 h (rescaled to daily scale)	(Reichle et al., 2018)
Soil Texture	Variables: Sand, Silt, Clay fractions; Bulk Density; Depths: 5, 10, 20, 60, and 100 cm	Pennsylvania State University (http://www.soilinfo.psu.edu/) – CONUS-SOIL (Developed from USDA STATSGO)	1 km	Static	(Miller and White, 1998)
Elevation	DEM	USGS GTOPO30 https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30	30 arc sec (approximately 1 km)	Static	–
Vegetation	MODIS MOD13A2 v006 – NDVI, EVI MOD17A2H v006 – GPP	NASA EOSDIS LPDAAC	1 km (MOD13A2); 500 m (MOD17A2H) (resampled to 1 km)	16 days (MOD13A2); 8 days (MOD17A2H)	(Didan, 2015; Running et al., 2015)
LST	MODIS MOD11A1 v006 – LST Day and Night Times	NASA EOSDIS LPDAAC	1 km	Daily	(Wan et al., 2015)
Precipitation	CHIRPS v2.0 data	Climate Hazards Center, UC Santa Barbara	0.05°	Daily	(Funk et al., 2015)
In-situ soil moisture	SCAN, SNOTEL, USCRN Depths: 5, 10, 20, 50, and 100 cm	ISMN	Point scale (upscaled to 1 km resolution)	Daily	(Dorigo et al., 2011)

capture the complex relationship between these controlling variables and multi-layer in-situ SM and transfer it to the ungauged high-resolution (1 km) grid locations. The proposed model consists of three steps: 1) generation of homogenous regions, 2) setting up machine learning modeling framework between in-situ SM at multiple layers (predictand) and above-described features (predictors) region-wise and depth-wise, and 3) applying the calibrated model to generate multi-layer SM at ungauged 1 km spatial resolutions over CONUS. The rationale of setting up machine learning models in a decentralized manner for SM disaggregation is implemented earlier by Abbaszadeh et al. (2019). An overview of these steps is discussed in the following paragraphs.

2.2.1. *Identification of homogeneous regions*

It is important to classify CONUS into homogeneous regions to capture the association between SM and controlling variables due to landscape and climate heterogeneity. Besides, constructing the machine learning algorithm for homogenous regions may address the problem of extrapolation (Reichstein et al., 2019). Such regional analysis improved the model performance in the past efforts of surface SM disaggregation (Abbaszadeh et al., 2019). In this work, besides soil texture (Abbaszadeh et al., 2019), we considered additional indicator variables that influence SM variability to holistically capture the processes that influence the multi-layer SM.

Although it will be possible to implement the downscaling method over legacy regions such as climate regions (Karl and Koss, 1984) or ecoregions (Omernik and Griffith, 2014), such regions have two limitations in the purview of this work. First, they may not account for the spatial homogeneity due to soil texture, vegetation, and topography, which are important inputs to predicting SM at fine spatial scales. Second, the number of regions may be too high (as in Level III or Level IV ecoregions of the CONUS) to ensure each region to have sufficient in-situ soil moisture data to set up the machine learning model. Since we are constrained by the location of in-situ stations (which are spread unevenly across the CONUS), a new set of homogeneous regions should be prepared to address the data inadequacy in each region. This step also adds more flexibility to implement the proposed method in other regions where in-situ soil moisture data is available.

The list of indicator variables includes (1) Daily precipitation characteristics (mean, standard deviation, and the number of rainy days); 2) Vegetation health based on NDVI, EVI, and GPP (mean, standard deviation, and range); 3) Day and night time LST (mean, standard deviation and diurnal LST range); 4) Soil Texture (% Clay, % Sand, % Silt, Bulk density); 5) Elevation; and 6) SMAP L4 Rootzone Soil Moisture (mean, standard deviation, range). Combining these multiple indicator variables was useful for generating homogeneous clusters at the catchment scale (Konapala and Mishra, 2020). Initially, we study the relationship between the above indicator variables and in-situ profile SM statistical parameters, such as mean, median, standard deviation, coefficient of variation, interquartile range (IQR), and range. These statistics are useful to capture the in-situ temporal dynamics of the SM profile. The in-situ profile SM at a station is estimated using a weighted average of SM observed at multiple depths. Similarly, soil texture data available at multiple layers is averaged to obtain profile representative soil texture. The weighting scheme of Ford and Quiring (2019) is utilized for this purpose. In this scheme, a sensor's top layer of control is assumed to be halfway between the current and the sensor above it, and the bottom layer of control is assumed to be halfway between the current and the sensor below it (Ford and Quiring, 2019). The layer thickness obtained through this procedure is considered as the weight of the corresponding layer. Pearson cross-correlation is used to assess the relationship between the indicator variables and in-situ SM profile parameters.

We apply a k-mean clustering algorithm (MacQueen, 1967) to 695 grid cells (where in-situ SM is available) using the indicator variables described above to generate homogenous regions. The in-situ stations are well spread out and are representative of different climate, vegetation, geomorphology, and topography conditions of the CONUS. It may be noted that location attributes are not considered in the list of indicator variables to negate the effect of unevenly located in-situ stations. In k-mean algorithm, each in-situ grid location, containing indicator variables as attributes, is initially assigned to k clusters randomly. Means of each of the indicator variables in a cluster are assigned as centroid of that cluster. Distances between each grid location and each of the k centroids are computed using Euclidean distance metric. Each grid location is assigned to the closest cluster centroid (in terms of Euclidean distance). The process of computing cluster centroids and subsequent assignment of clusters to each grid location continues iteratively until there is no change in clusters. During this process, each indicator variable is standardized across the in-situ grid locations (based on maximum

and minimum values) to generate a dimensionless time series. The optimum number of clusters are derived based on Xie-Beni (Xie and Beni, 1991) and Dunn (Dunn, 1974) indices. The ungauged grid cells are then assigned to one of the homogenous regions using the closest distance between the ungauged grid cell's indicator variables and the cluster centre as the criteria.

2.2.2. Setting up the machine learning algorithm

A machine learning algorithm is constructed for soil layers within each homogenous region based on the predictors (Table 1) and in-situ multi-layer SM as predictand. We use the state-of-the-art Extreme Gradient Boosting (XGBoost) machine learning algorithm for this purpose. XGBoost (Chen and Guestrin, 2016) gained popularity in the machine learning community, given its speed, efficiency, and accuracy. XGBoost found applications in recent years for classification and regression purposes in remote sensing and water resources fields (Chemura et al., 2020; Ni et al., 2020; Zhang et al., 2019a). XGBoost is also used for downscaling of groundwater (Zhang et al., 2021a), crop yield (Folberth et al., 2019), and surface soil moisture (Liu et al., 2020b). However, this algorithm has not been explored for multi-layer SM estimation.

XGBoost uses decision tree ensembles consisting of a set of classification and regression trees (CART). Consider a dataset with n samples and m features input $\mathbf{x}_i$ ($i = 1, 2, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^m$) and output y_i. The predicted values are obtained using the following equation.

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \Theta \tag{1}$$

where, f_k is k^{th} regression tree; Θ is the space of all possible regression trees (CARTs); $\widehat{y}_i$ is the predicted value corresponding to y_i, the observed output; K is the number of regression trees. Each regression tree (f_k) is an independent tree structure with leaf weights (denoted by w_t – the weight of t^{th} leaf). The following regularized objective function is minimized to learn the set of functions used in the model.

$$Obj = \sum_{i=1}^{n} e\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where, $e(\cdot)$ is the differentiable convex loss function computed between observed (y_i) and predicted values ($\widehat{y}_i$); n is the number of values; and $\Omega(f)$ is the regularization function, which controls the complexity of the function by penalizing the complex functions to avoid overfitting. $\Omega(f)$ is estimated from the following equation.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{t=1}^{T} w_t^2 \tag{3}$$

where, γ and λ are regularization parameters that control the tree's complexity; T is the number of leaves in a decision tree; w_t is the weight (score) associated with t^{th} leaf. XGBoost implements additive training, an iterative process, to obtain an optimum ensemble model. For this purpose, Eq. (2) is altered as follows.

$$Obj^{(p)} = \sum_{i=1}^{n} e\left(y_i, \widehat{y}_i^{(p-1)} + f^{(p)}(x_i)\right) + \Omega\left(f^{(p)}\right) \tag{4}$$

where, $\widehat{y}_i^{(p)}$ is the prediction at i^{th} instance p^{th} iteration. Inclusion of $f^{(p)}(x_i)$ in the loss function indicates a greedy addition of $f^{(p)}(x_i)$ that improves the model from Eq. (2). The loss function in Eq. (4) is approximated using a second-order Taylor approximation to increase the efficacy of optimization. The resultant objective function is minimized to obtain the optimal weights of leaves and subsequently determine tree structure quality.

For effective implementation of the XGBoost algorithm, it is essential to tune the model parameters. Six parameters are selected for this purpose. They include, 1) eta – controls the learning rate, 2) max_depth – maximum depth of a tree, 3) nrounds – number of iterations, 4) subsample – a fraction of observations to be randomly sampled for each tree, 5) colsample_bylevel – a fraction of columns to be randomly sampled for each new depth level, in each tree, 6) min_child_weight – a minimum sum of instance weight required in a leaf node to proceed with further partitioning. Further details on the selection of model parameters can be obtained from https://xgboost.readthedocs.io/en/latest/index.html.

It may be noted that gradient boosted trees of XGBoost have the capability of handling multicollinearity among predictors. The tree node splitting in decision trees is based on a reduction in node impurity measures. In XGBoost, it is based on a factor called gain, which describes the relative contribution of a feature to the model. When correlated variables are present in the input feature space, the split in decision trees will be based on only one of them (Kotsiantis, 2013). The subsequent split will happen based on uncorrelated variables. The correlated variables are naturally left out since there is minimal information gain obtained by splitting them. Since boosted trees in XGBoost use an ensemble of decision trees, the algorithm inherently avoids overfitting, which can be an issue when a single decision tree is used on multicollinear variables. XGBoost has another advantage of inherently handling the missing data. It achieves this task using sparsity-aware split finding algorithm (Chen and Guestrin, 2016). Through this algorithm, XGBoost defines an optimal default direction at tree nodes where missing data in a feature are encountered. The optimal direction is obtained by trying both the directions in a split and selecting the one which results in a maximum gain. The optimal direction is learned by visiting only non-missing observations. More details on the algorithm can be obtained from Chen and Guestrin (2016).

The model parameters are specified with candidate values (that lie in respective ranges), and parameter tuning is carried out using the grid search method. Overall, 80% of the data (predictor-predictand sets) is used to train the model, and the remaining 20% of the data unseen by the model is used to test the model. The training process is carried out for individual clusters and multi-layer SM information. This results in h*5 number of independent models, where h is the number of clusters generated over the CONUS, and 5 is the number of soil layers.

2.2.3. Application and validation and of the machine learning algorithm

Soil moisture values generated by the XGBoost model are initially evaluated for their accuracy in the model testing phase. After training the models, the relative importance of predictors in each cluster and soil layer is analyzed by computing the relative contribution of each predictor in each tree of the model. The calibrated XGBoost models at each soil layer are then applied at ungauged grid cells (based on the cluster the grid cell belongs to) to obtain 1 km resolution multi-layer SM. Fig. 1 presents the workflow of the proposed method to obtain 1 km resolution multi-layer SM.

For validation, the model-generated SM values are verified for accuracy at 79 grid locations containing in-situ data, which are not used for building the models. For this purpose, we use data from Little Washita, Fort Cobb, Texas soil observation network (TxSON) (Caldwell et al., 2019), and SoilSCAPE (Moghaddam et al., 2016) networks. TxSON and SoilSCAPE networks. These networks provide SM measurements at multiple layers up to 50 cm. Multiple stations located within a 1 km grid cell are upscaled to match the grid resolution using the arithmetic average. Tables S7-S10 present the location information of stations in these networks. There is no data available for 100 cm depth for validation. Therefore, we present the accuracy of the model for this layer in the testing phase. While building the model at 50 cm depth, we could not obtain soil texture information at this level since CONUS-SOIL data is only available at 60 cm depth. Therefore, we used soil data corresponding to 60 cm depth for the analysis. This approximation at 50 cm depth may cause some uncertainty in the model outcome. The accuracy assessment in the testing and validation phases is carried out

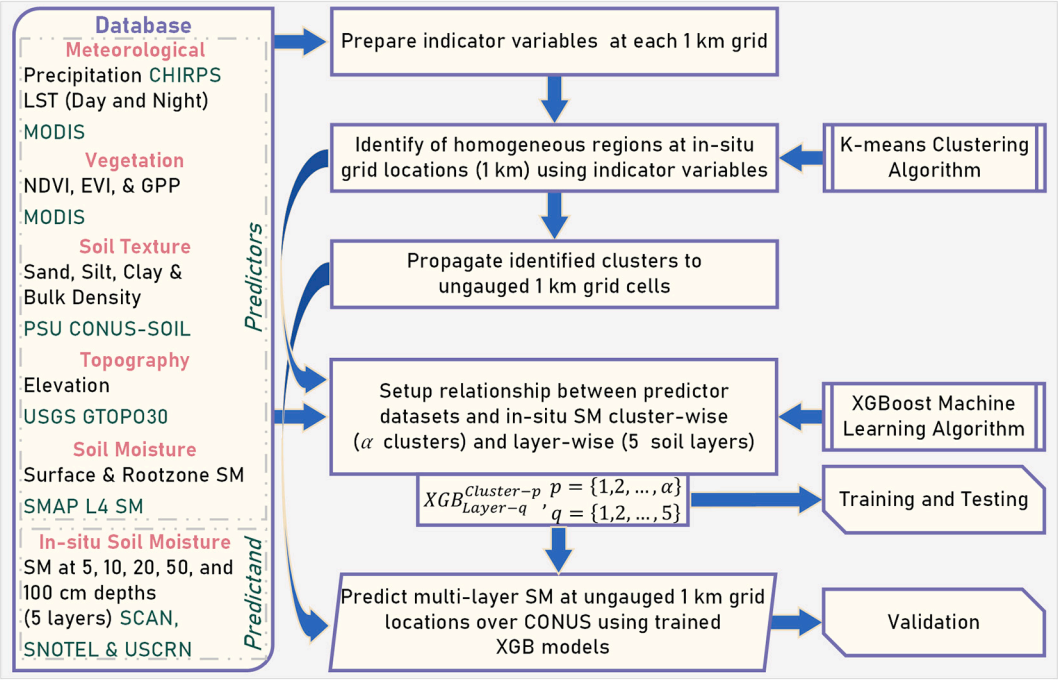


Fig. 1. Workflow of the proposed method to obtain 1 km resolution multi-layer SM over the CONUS.

using unbiased root mean square error (ubRMSE), Pearson correlation, and bias metrics (Gruber et al., 2020).

3. Results and discussion

3.1. Relationship between indicator variables and SM profile parameters

Fig. 2 presents the correlation plot between the indicator variables and in-situ SM profile parameters. Mean and median SM have a similar dependence on the predictors. They are strongly correlated with soil texture properties with clay % and silt % content (due to higher moisture holding capacity) and negative correlation with sand % (due to lower porosity). A similar observation was made by Wang et al. (2017) while studying climate and soil effects on SM's spatial variability. Elevation exhibited a negative relationship with mean and median SM. Precipitation and vegetation predictors are positively correlated with mean and median SM. In their field-scale study, Jacobs et al. (2004) found soil properties, topography, and vegetation to influence mean SM variability.

Interestingly, the daytime mean LST is not correlated with mean and median SM, whereas nighttime LST is significantly correlated with these two variables. The relationship between SM and LST during the nighttime or early morning is highlighted by Zhao et al. (2018) in disaggregating surface SM. Conversely, the daytime LST standard deviation has a significant negative correlation, while the night time LST standard deviation is not correlated with mean and median SM. This could be attributed to the lower variability of nighttime LST (Rebetez, 2001). LST represents a snapshot of temperature conditions, whereas in-situ SM

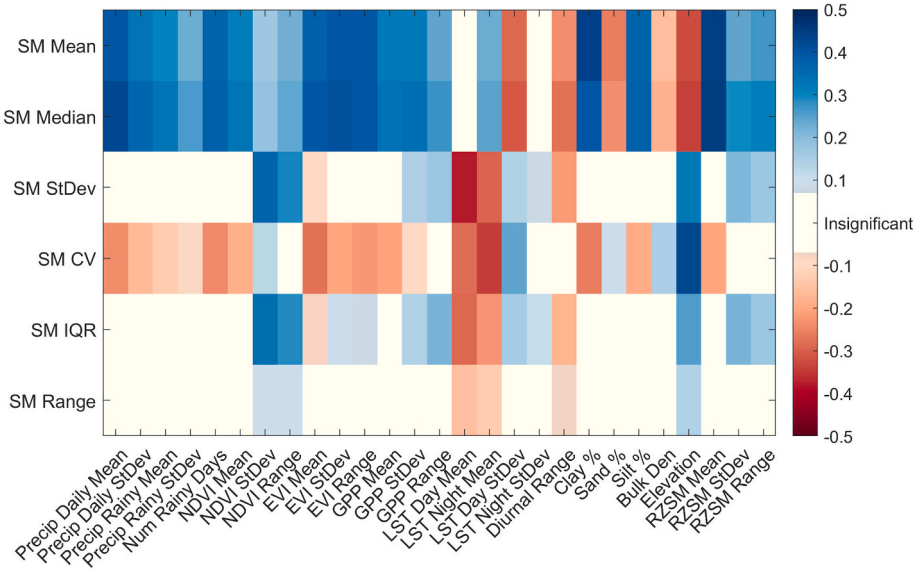


Fig. 2. Correlation plot between indicator variables (x-axis) and in-situ SM profile parameters (y-axis). Non coloured cells indicate a statistically insignificant correlation between two variables (at 95% confidence).

representing average SM at a daily scale could have also contributed to low correlations. Besides, the diurnal range of LST has a significant negative correlation with mean and median SM, i.e., the diurnal range is high in the dry regions where low SM is observed and vice versa. Both variables are positively correlated with rootzone SM predictors.

The SM profile parameters (e.g., standard deviation and IQR) have no significant correlation with precipitation. The standard deviation of SM is reported to vary with mean SM negatively, with the relationship governed mainly by soil hydraulic properties (Famiglietti et al., 2008; Vereecken et al., 2007). Given such a relationship, the correlation between SM's mean and standard deviation is expected to be close to zero. This could be the reason behind the insignificant correlation between the standard deviation of SM and the mean root zone SM. Both standard deviation and IQR of SM have a significant negative correlation with the day time and night time LST along with a diurnal range of LST. Although this indicates high variability of SM under wet conditions and vice versa, the inference may not be applicable at all locations since contrasting conclusions have been obtained in the past on the relationship between SM and LST (Famiglietti et al., 1998; Qiu et al., 2001). Although few studies reported correlations between SM and soil parameters (sand % and clay %) (Wang et al., 2017; Wu et al., 2020), such results are dependent on site characteristics. The elevation is found to be positively correlated with SM profile parameters.

The SM variability determined based on the coefficient of variation (CV) is significantly correlated with most of the predictors. It has shown a negative correlation with mean rootzone SM, which is in line with previous findings (Famiglietti et al., 2008; Jacobs et al., 2004; Korres et al., 2015). Most of the vegetation indicators (except NDVI) are negatively correlated with the CV of SM; this could be due to wet conditions (high SM) prevailing under greener vegetation (high magnitude of vegetation indices). A few indicators control the SM range (maximum-minimum), including standard deviation and range of NDVI, mean of LST (both day time and night time), diurnal range, and elevation.

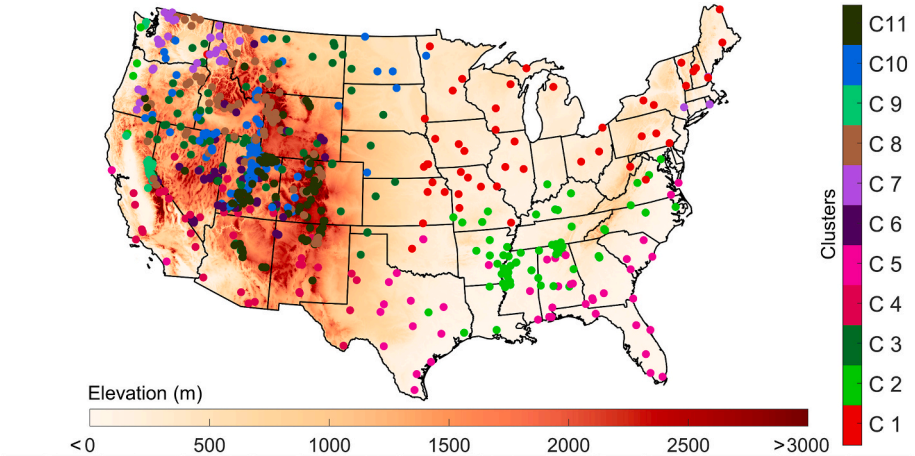
3.2. Identification of homogenous regions

A combination of predictor variables influences the SM characteristics (e.g., mean and Variability); however, their association varies in space and time. Therefore, we classify CONUS into several homogeneous regions defined by a more significant similarity among the predictors and predictand relationship. We applied the K-means clustering algorithm to 695 in-situ SM locations (where SCAN, SNOTEL, and USCRN stations are situated) using the predictors mentioned in Fig. 2. We identified 11 homogeneous regions based on the Xie-Beni and Dunn indices that are used commonly for determining the optimal number of clusters. The spatial locations of in-situ SM locations and their corresponding clusters are displayed in Fig. 3, and the number of SCAN, SNOTEL, and USCRN stations represents each cluster are also provided. Fig. 4 presents the variability of predictors across clusters.

The homogenous regions are very distinct (Fig. 3), highlighting the potential role of climate, vegetation, geomorphologic, and soil characteristics. The highest number of stations are located in clusters 10 (114 stations) and 11 (116 stations). Most of the stations in cluster 11 are located in high elevation forest regions (average elevation ~2800 m) of Colorado, Utah, and Arizona. Given the high altitude of stations in the cluster, the LSTs are relatively lower than other clusters. There is also high sand content with soil texture ranging from loam to sandy loam, resulting in a low range of the rootzone SM.

Cluster 10 is mostly spread across western CONUS and a portion of northern CONUS. Stations in this cluster are mostly located in loamy soil and situated in deciduous forested regions, which results in high variability (due to seasonality) of vegetation indicators. Similar high variability of vegetation indicators is also observed in the case of clusters 1 and 7. Stations of cluster 1 are mostly situated in the Midwest and Northeast CONUS, wherein croplands dominate the former, and the latter is covered with deciduous forests. Strong seasonality of vegetation in these land cover classes could have attributed to greater vegetation indicators variability than any other cluster. This cluster also has the highest average silt content of ~45%, with soils predominantly belonging to silt loam texture. This leads to soils having high available water content (promoting plant water uptake) and generally high moisture content. Notably, the average elevation of this cluster's stations (~ 342 m) is significantly lower than that of clusters 10 and 11.

Cluster 7 is mostly limited to the Northwest CONUS. Most of the stations are situated in evergreen needleleaf forests, with high mean and variability of vegetation indicators. Noticeably, this cluster has the lowest diurnal range due to the cold climate existing in this region. Stations in this cluster are located over silt loam soils with high sand and silt content. Therefore, the mean and range of root zone SM are higher than the conditions observed in cluster 11 (which also has high sand content). This observation is also supported by the higher amount of precipitation received by the stations than previously described clusters.



Clus	SCAN	SNOTEL	USCRN	Total	Clus	SCAN	SNOTEL	USCRN	Total	Clus	SCAN	SNOTEL	USCRN	Total
C1	28	0	21	49	C5	24	0	21	45	C9	0	29	2	31
C2	45	1	17	63	C6	9	32	2	43	C10	17	92	5	114
C3	30	19	23	72	C7	0	30	5	35	C11	0	115	1	116
C4	25	0	12	37	C8	1	86	3	90	**Total**	179	404	112	**695**

Fig. 3. Clusters of in-situ SM stations obtained from SCAN, SNOTEL, and USCRN networks overlayed on elevation map. [Note: the table provides a distribution of stations for eleven clusters across the three networks].

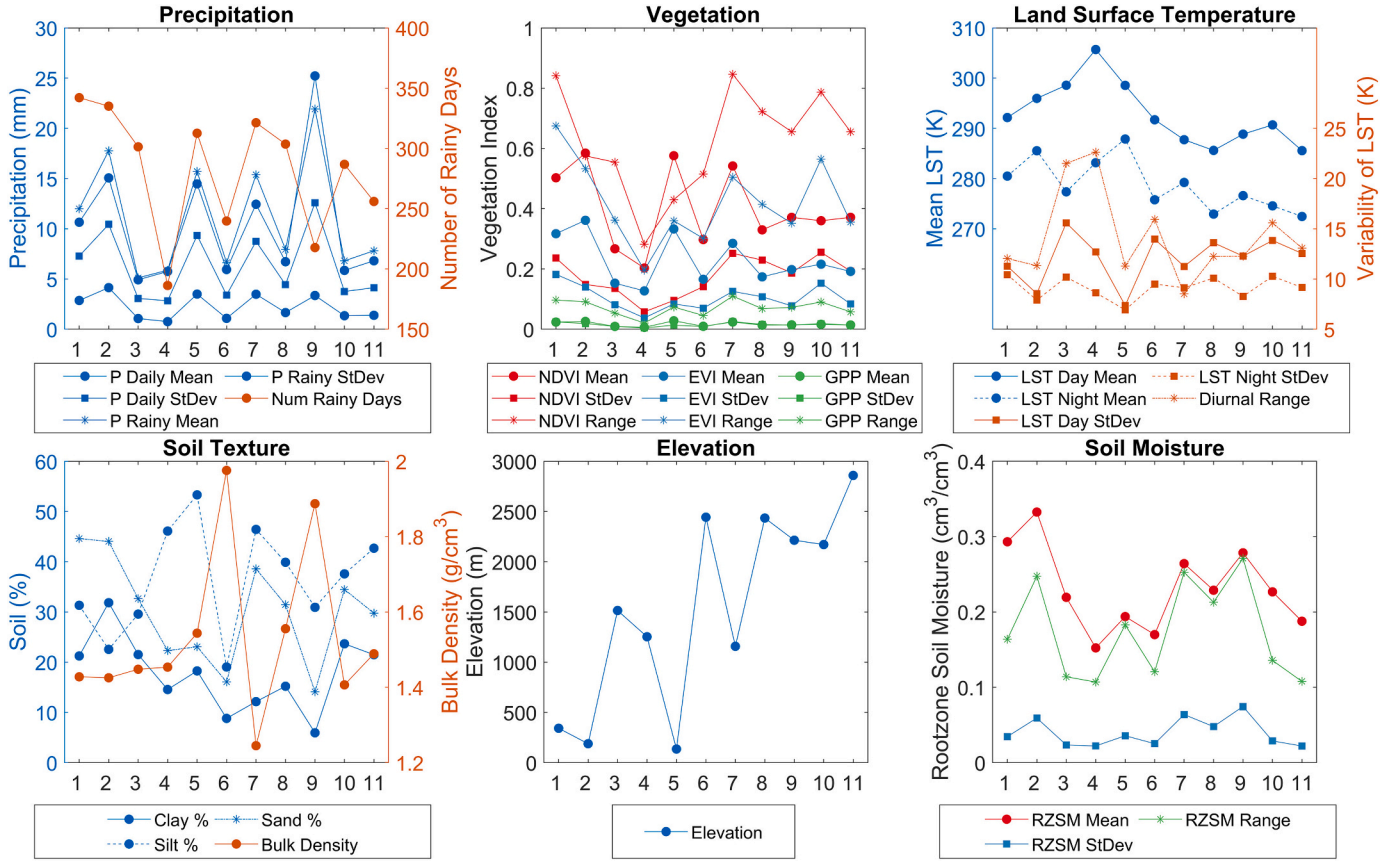


Fig. 4. Variability of predictor variables for eleven clusters. The horizontal axis in each figure indicates cluster number (consistent with that of Fig. 3). [Note: values provided in the plots represent the mean values of the selected variables].

Clusters 2 and 5 exhibited similar precipitation indicators as that of cluster 7. However, each of them is distinct in terms of other indicators. Stations representing cluster 2 are located in the Southcentral and Northwest coastal CONUS. Given a high number of rainy days coupled with high clay content, the average rootzone SM in this cluster is highest than that of other clusters. Besides, several stations are situated along the Mississippi river banks, which could have also contributed to high rootzone SM. A similar observation is made with stations located near the water bodies situated in northern Alabama.

Furthermore, the clustering algorithm could distinguish stations in southern Missouri, predominantly covered with mixed forests, cluster 2, from stations located in northern Missouri's croplands (categorized into cluster 1). Cluster 5 borders cluster 2 and is situated in Southeast CONUS, Texas, and some California stations. Despite the contrast in the stations' longitudinal location in these regions, they are categorized into a single cluster given their high sand content. This cluster is found to have the highest average mean NDVI compared to that of other clusters. Interestingly, stations in coastal plains and portions of Texas's central plains (covered with vegetation) are categorized into cluster 2. In contrast, stations in great plains and mountains and basins regions of Texas (where arid conditions persist) are categorized into clusters 3 and 4.

Cluster 3 is spread over a larger geographical area compared to other clusters. Stations of this cluster are located in the shrublands/grasslands of West CONUS and great plains. Despite a large spatial extent, this cluster has stations that experience low precipitation mean and variability. On the other hand, these stations have the highest LST variability compared to other clusters. Cluster 4 is mostly spread across the shrublands of Southwest and West CONUS. This cluster has the lowest mean and variability of vegetation indicators, the highest average daytime LST, and the lowest mean precipitation and number of rainy days. Given such dry conditions alongside moderately high sand content, this cluster has the lowest mean and variability of rootzone SM. Noticeably, the densely vegetated high altitude SNOTEL stations located in central Arizona are categorized mostly into cluster 11.

Cluster 9 has densely located stations in the high altitudes (average elevation ~2435 m) of California's Sierra Nevada region. This cluster is the wettest of all clusters, with mean rainy-day precipitation of ~22 mm. The mean vegetation indicators and LST indicators of this cluster are similar to clusters 9 and 10. This could be due to the similar high-altitude topography of the three clusters. The mean rootzone SM of stations in this cluster is moderately high despite sandy loam soils (low moisture-holding capacity). This could be attributed to the high amount of precipitation received by the cluster. Stations of cluster 6 are situated in the high-altitude areas of Southwest CONUS. This cluster is noticed to have the highest soil bulk density compared to other clusters. Lastly, cluster 8 is mostly comprised of SNOTEL stations situated in the densely vegetated high-altitude areas of West CONUS. It is unique from other high-altitude clusters in terms of low clay content.

The CONUS is classified into 11 homogenous regions using the classification scheme developed based on the in-situ SM locations. The ungauged 1 km grid cells are assigned to one of the eleven clusters using the lowest distance between indicator variables as the criteria (see methodology Section 2.2). Fig. S2 presents the resulting clusters of homogeneous regions for the CONUS. Clusters are found to be mostly contiguous despite not providing location attributes as indicator variables.

3.3. *Application of machine learning algorithm for multi-layer SM estimation*

The XGBoost machine Learning algorithm is set up individually at

each cluster and each soil layer, resulting in 55 models (11 models × 5 soil layers). Fig. 5 present the in-situ testing performance in terms of ubRMSE, correlation, and bias, respectively. Table 2 presents the cluster- and layer-wise median values of performance metrics during the model testing phase.

Overall, XGBoost models performed with reasonable accuracy across most of the clusters and soil layers. We found XGBoost to outperform Multiple Linear Regression (MLR) model during the testing process (Table S1). The MLR-based model set up involved replacing XGBoost models with MLR models. The stations located in cluster 4 consistently produced low ubRMSE for predictions at 5 cm. This could be due to SM's low variability associated with the dry conditions, which results in lower magnitude of ubRMSE (Entekhabi et al., 2010b). Besides, SM retrievals' accuracy is generally higher in the arid regions (Zhang et al., 2019b). Despite the large spatial extent of clusters 1 and 3, their ubRMSE and correlation values are less than 0.04 m^3/m^3 (within the acceptable limit) and greater than 0.85, respectively. This highlights the robustness of the model performance in homogenous regions with similar geomorphology and climatology. The model performance in cluster 11 appears low due to the high ubRMSE, low R, and high bias, especially in Utah and Colorado. This could be due to mountainous topography, which induces bias in the brightness temperature observations (Li et al., 2013). However, given the limited geographical extent of this cluster (Figs. 3 and S2), its performance shall not affect other CONUS regions. Generally, the model performance in the central plains is distinctly lower compared to other regions. This could be attributed partly to land surface model structure and parameters uncertainties, given that SMAP brightness temperatures have not significantly improved the skill of CLSM simulations over North America (Dong et al., 2019). Apart from high elevation clusters, the performance of cluster 5 is found to be low. However, this behavior is limited to a few isolated stations located in Texas, Alabama, and Florida within the cluster. Similar performance (few stations located in Texas, Alabama, and Florida) is noticed in the results of Abbaszadeh et al. (2019) concerning disaggregation of SMAP surface SM.

The number of stations with 10 cm depth retrievals is comparatively less than the number of stations included in 5 cm depth retrievals. This is noticed especially in Nevada, Utah, Idaho, Wyoming, and Colorado, which are dominated by SNOTEL stations that lack probes at 10 cm and 100 cm depths. Despite training with a different set of models, the error patterns depict similar spatial characteristics as the 5 cm case. Recently, Akbar et al. (2018) found a high correlation between USCRN SM measurements made at 5 cm and 10 cm depths. This explains the similarity in spatial patterns. Most of the stations produced SM with ubRMSE less

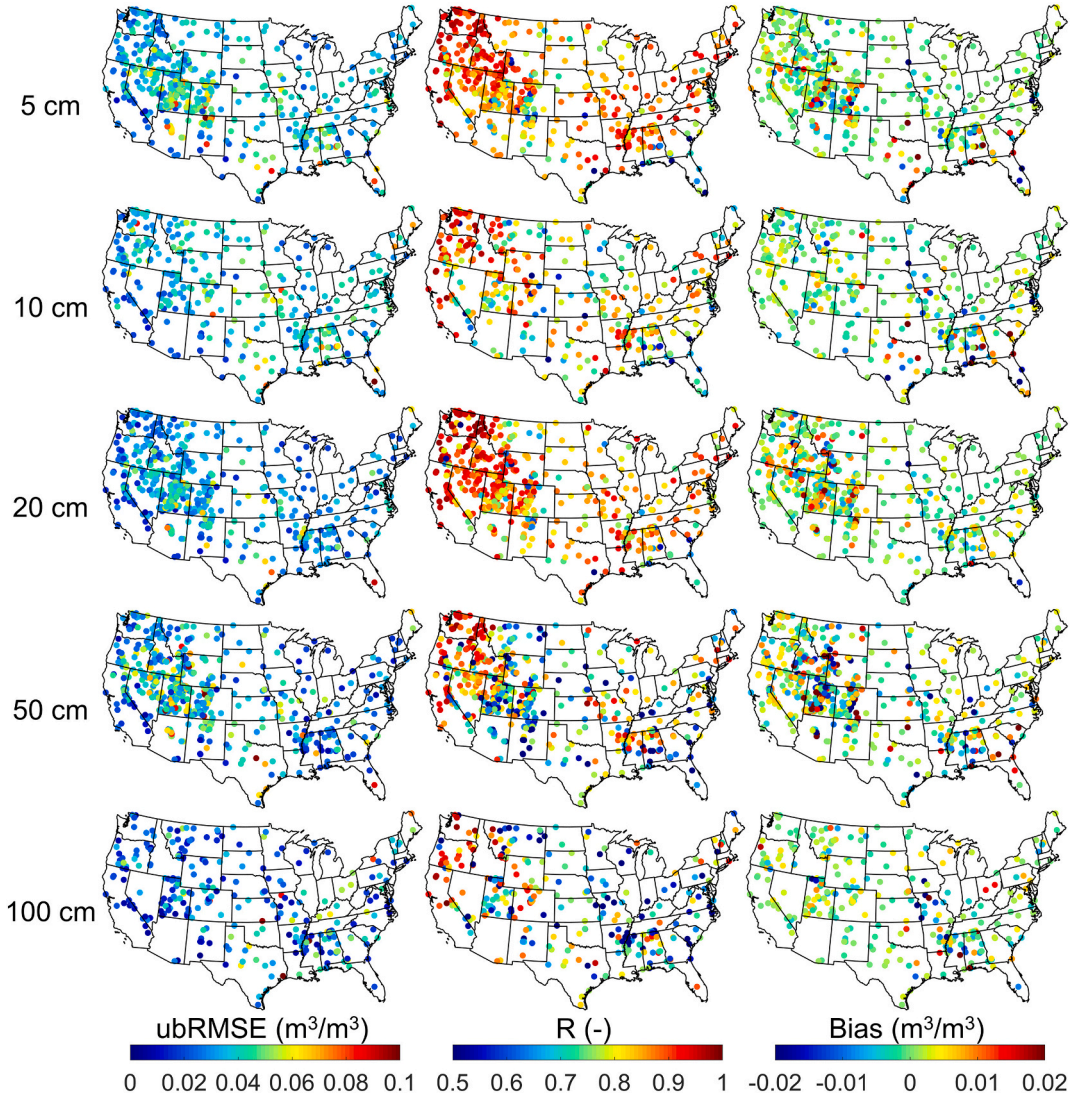


Fig. 5. Goodness of fit statistics (ubRMSE, R, and bias) across SCAN, SNOTEL, USCRN locations estimated during the model testing phase. Performance metrics are calculated between model prediction and in-situ SM data sets.

Table 2
Cluster-wise median values of performance metrics of the model during testing phase across five soil layers. ubRMSE is in m^3/m^3, R is unitless, and Bias is in m^3/m^3.

Cluster	5 cm			10 cm			20 cm			50 cm			100 cm		
	ubRMSE	R	Bias	ubRMSE	R	Bias	ubRMSE	R	Bias	ubRMSE	R	Bias	ubRMSE	R	Bias
1	0.0373	0.8560	0.0003	0.0403	0.7869	0.0000	0.0269	0.8537	−0.0003	0.0270	0.7777	0.0008	0.0267	0.6813	0.0009
2	0.0350	0.8955	0.0004	0.0395	0.8151	−0.0004	0.0289	0.8580	0.0002	0.0243	0.8197	0.0005	0.0279	0.6910	0.0001
3	0.0360	0.8667	0.0001	0.0366	0.7939	−0.0004	0.0312	0.8694	0.0014	0.0331	0.7626	−0.0009	0.0229	0.7180	−0.0007
4	0.0236	0.8288	0.0000	0.0224	0.7458	0.0003	0.0177	0.8461	−0.0001	0.0191	0.7568	0.0017	0.0104	0.7095	0.0000
5	0.0449	0.7724	0.0045	0.0411	0.7505	0.0032	0.0314	0.7807	−0.0002	0.0285	0.6518	−0.0005	0.0325	0.7346	0.0003
6	0.0404	0.8614	−0.0001	0.0319	0.8578	0.0009	0.0317	0.8914	0.0000	0.0345	0.8162	0.0013	0.0252	0.7842	0.0009
7	0.0267	0.9497	0.0005	0.0309	0.9283	−0.0002	0.0240	0.9520	0.0003	0.0258	0.9410	0.0022	0.0181	0.9364	−0.0001
8	0.0372	0.9028	0.0001	0.0364	0.9036	0.0001	0.0342	0.9125	0.0019	0.0368	0.8192	−0.0002	0.0183	0.8236	0.0004
9	0.0322	0.9118	0.0002	0.0166	0.9830	0.0005	0.0276	0.9401	0.0003	0.0396	0.9097	0.0039	0.0135	0.9828	−0.0002
10	0.0396	0.8876	0.0004	0.0398	0.8151	0.0008	0.0356	0.8863	0.0011	0.0430	0.7861	0.0005	0.0290	0.7364	0.0007
11	0.0535	0.7826	0.0022	0.0412	0.8577	0.0000	0.0395	0.8365	0.0018	0.0471	0.6615	0.0003	0.0303	0.8765	0.0012
Average	**0.0369**	**0.8650**	**0.0008**	**0.0342**	**0.8398**	**0.0004**	**0.0299**	**0.8752**	**0.0006**	**0.0326**	**0.7911**	**0.0009**	**0.0232**	**0.7886**	**0.0003**

than 0.04 m^3/m^3. Cluster 9 is found to perform with the highest accuracy. This is due to the presence of only three stations that contained SM observations at 10 cm depth in this cluster. Although cluster 4 trained with low ubRMSE, its correlation is lower compared to the 5 cm case. A similar observation is made with cluster 5, wherein the correlation of few stations in Texas, Alabama, and Florida are found to be lower (~0.5).

Soil moisture estimates at 20 cm depth are found to have higher accuracy than other depths, especially in terms of correlation metric (although 100 cm depth estimates have better ubRMSE, the number of stations available for testing is comparatively lesser). Nearly 70% of the stations located in cluster 4 have lower ubRMSE (ubRMSE $\leq$0.04 m^3/m^3). This is expected due to the background aridity, which results in low variability of SM in the deeper layers. The high-altitude clusters (6, 8, 9, 10, and 11) have performed well in both ubRMSE and correlation. However, clusters 3, 8, 10, and 11 depicted a slight wet bias in the SM retrievals (Fig. 5). In general, the models' training accuracy is high in clusters 2 and 7, mostly located in the regions of high average precipitation and SM.

Approximately 93% and 43% of 695 stations have SM at 50 cm and 100 cm depths, respectively. The models performed reasonably well based on the ubRMSE metric. However, lower correlation values are noticed across clusters in the 50 cm layer compared to predecessor layers. Typically, the stations located in high elevation clusters (8–11) and few stations in clusters 2, 4, and 5 witnessed a correlation less than 0.5. As we progress to deeper layers, SM's sensitivity to surface and climate processes reduces compared to the influence of other variables, including soil texture and surface SM (Pan et al., 2017). Therefore, the involvement of climate variables could have resulted in a decline in SM estimation model performance of SM estimation at100 cm depth. To gain further insights, the relative importance of predictors towards estimating SM in each soil layer is assessed.

3.4. Relative importance of predictors

The relative importance of predictors for the five soil layers is presented in Fig. 6 (details in Section 2.2). Bars indicate the variability of relative importance across the eleven clusters. Surface SM estimates from L4 SMAP have the highest importance while predicting SM at 5 cm depth. This is followed by elevation, vegetation attributes, and soil texture features. LST attributes, along with precipitation, have the lowest importance among the predictors. It is necessary to note that the importance of predictors is determined at 1 km spatial resolution. At finer spatial scales, it is reported that SM variability is controlled predominantly by the topography and soil textural parameters, and the role of meteorological variables is significant only at regional/watershed scales (Crow et al., 2012; Gaur and Mohanty, 2016). Interestingly, NDVI, GPP, and closely followed by EVI, have high importance. The influence of vegetation is significant in some studies, whereas few studies reported limited influence on SM variability. For example, vegetation is a key controlling variable for the variability of in-situ SM data (Teuling and Troch, 2005) and (Baroni et al., 2013); whereas, vegetation has a limited role at scales closer to 1 km (Gaur and Mohanty, 2016) and (Joshi and Mohanty, 2010). These studies are based on local observations, and the inferences can be site-specific. The present results depict a holistic view of the geomorphological, topographical, meteorological, and vegetation controls of SM profile at 1 km resolution.

SMAP surface and rootzone SM are used as a predictor to estimate SM at 10, 20, 50, and 100 cm depths. At 10 cm depth, SMAP surface SM has the highest importance, followed by the rootzone SM. This can be attributed to the higher correlation of SM between 5 cm and 10 cm depths (Akbar et al., 2018). Besides, the high variability of the importance of surface SM can be attributed to the variability in cluster characteristics, which are defined by the indicator variables. Some of these variables (related to precipitation, vegetation, and soil texture) are found to influence the characteristic length scale of SM (Akbar et al.,

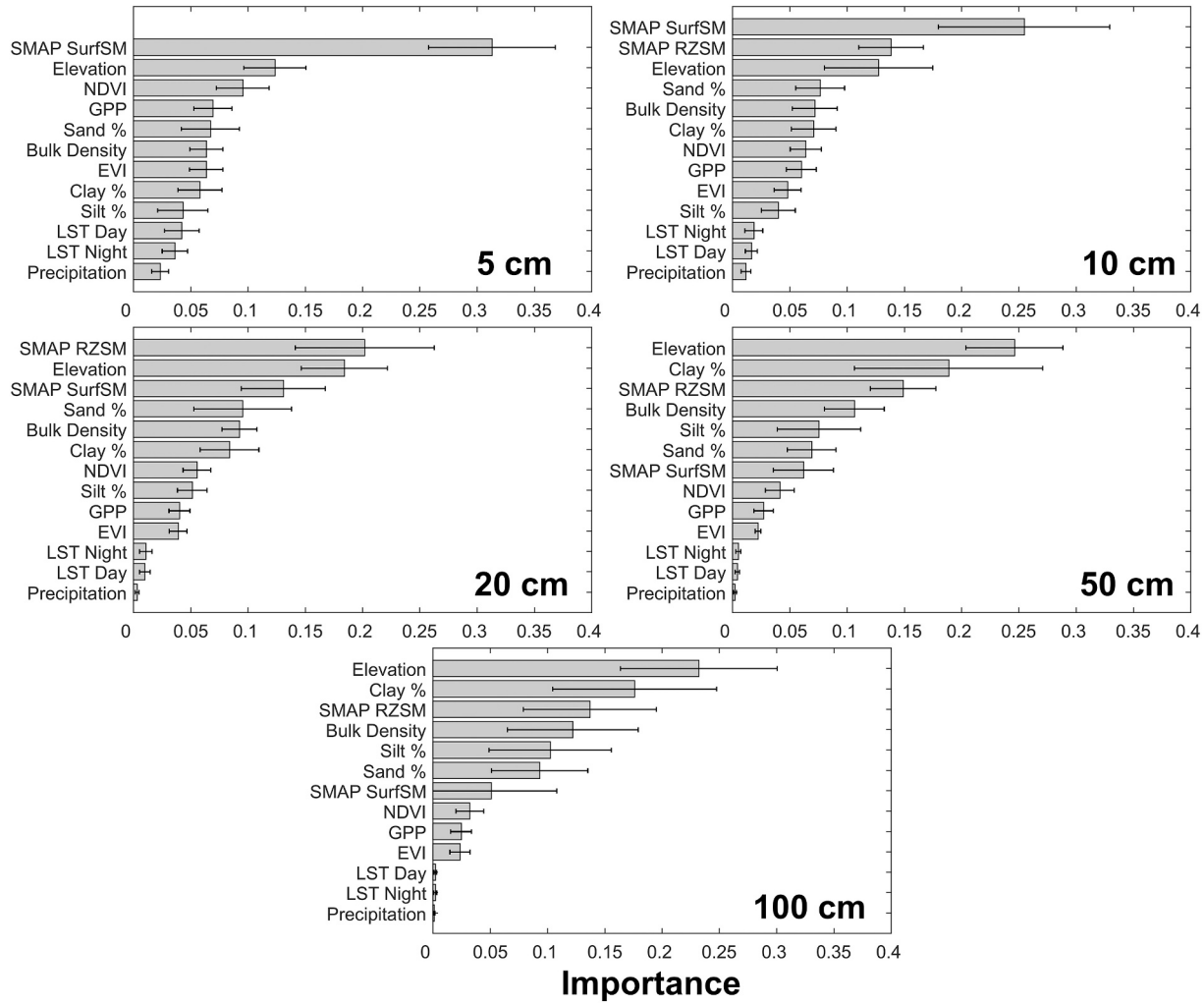


Fig. 6. The relative importance of predictors across five soil layers. Error bars associated with each variable represent their spread (i.e., variability) calculated based on standard deviation across the eleven clusters.

2018). Soil textural properties have greater importance over vegetation indicators in this layer, an opposite of which is noticed at 5 cm depth.

Rootzone SM is the dominant predictor at 20 cm depth. Interestingly, elevation has greater importance compared to surface SM. Besides, the soil textural parameters have greater importance, and the importance of vegetation further diminished at 20 cm depth compared to 10 cm depth. In the case of 50 cm and 100 cm depths, the order of predictors' importance remains the same. Noticeably, rootzone SM is the third significant predictor, while elevation and clay (%) have greater significance in these two layers. Few studies found topography and soil textural properties to influence rootzone SM's Variability (Pan et al., 2017; Shi et al., 2014). The importance of surface SM dropped noticeably from 20 to 50 cm depth, signifying the surface and deeper SM values' decorrelation. Vegetation has higher importance compared to meteorological predictors LST and precipitation. Some studies reported NDVI and EVI's potential to predict rootzone SM at SCAN sites (Schnur et al., 2010; Wang et al., 2007). Lastly, LST and precipitation predictors have a lower influence as we progress to deeper layers, which indicates a reduction of their control on deeper SM dynamics.

3.5. Evaluation of high-resolution multi-layer SM estimates

3.5.1. Validation of SM temporal dynamics

Soil moisture estimates obtained from the multi-layer downscaling algorithm are validated using the actual SM data obtained from the TxSON, Little Washita, Fort Cobb, and SoilSCAPE networks. Several sensors in these networks have in-situ SM at 5, 10, 20, 50, and 100 cm depths. These observed in-situ data sets are not included in the model-building stage. We considered all the stations with multi-layer SM observations for the validation since SMAP L4 SM (a key input for XGBoost model) is found to perform reasonably well even with sparse networks (Reichle et al., 2018). Fig. 7 presents the boxplots of performance metrics of predicted SM at 5 cm computed at in-situ stations across four networks. Table S2 presents the performance metrics at these stations. The model is validated at 31, 21, 16, and 7 grid locations that fall under Little TxSON, Washita, Fort Cobb, and SoilSCAPE networks, respectively. Table S3 presents the station-wise number of observations used for the computation of performance metrics. Results generally report low ubRMSE and high correlations in most of the grid locations. The median ubRMSE values are less than 0.04 m^3/m^3 (the mission accuracy target for SMAP mission) for TxSON and Fort Cobb networks. Correlation values are in the range 0.64–0.91, 0.50–0.84, 0.55–0.86, and 0.68–0.95, for TxSON, Little Washita, Fort Cobb, and SoilSCAPE networks respectively. The median correlation values are on a higher side compared to SMAP L4 (version 4) surface SM product's correlation metrics (Fig. 2 in (Reichle et al., 2019)). Although XGBoost models have been trained with negligible bias concerning SCAN, SNOTEL and USCRN networks (Fig. 5), some bias still persists in the model validation stage. This could be partly due to the wet bias in the SMAP L4 surface SM product (version 4) (Reichle et al., 2019; Reichle et al., 2018). Wet bias

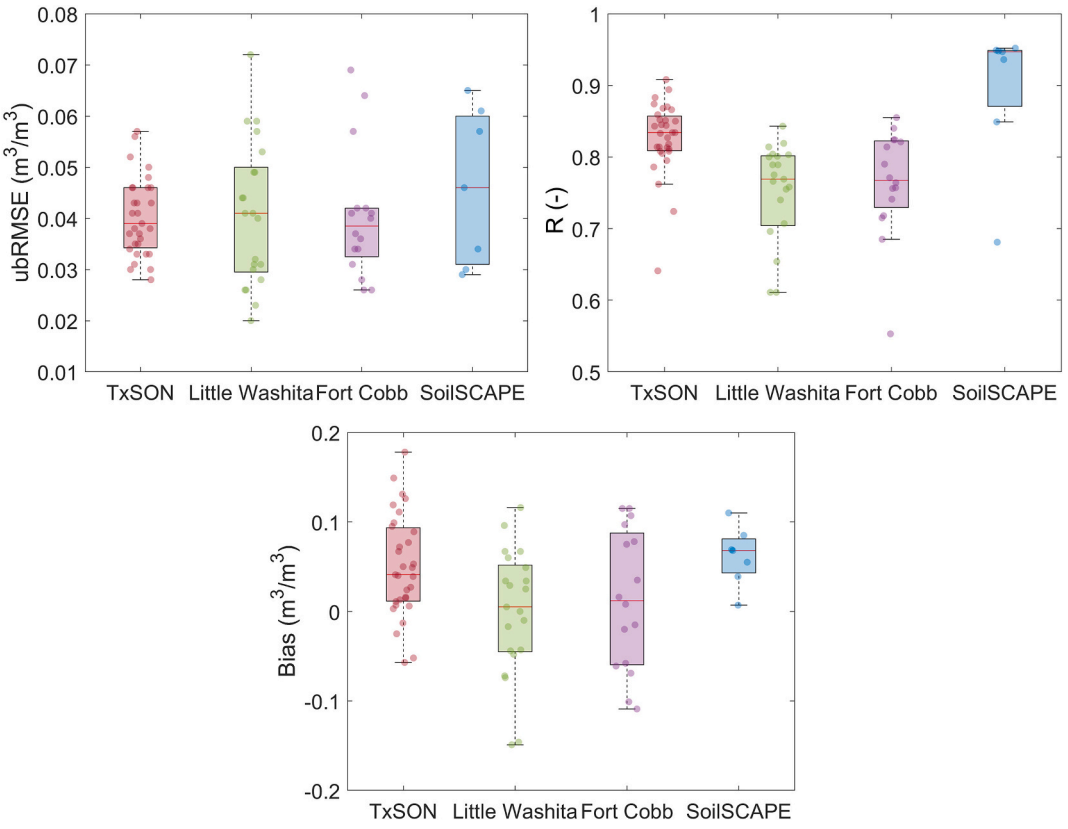


Fig. 7. Boxplots of ubRMSE, R, and Bias computed at in-situ stations of four networks, TxSON, Little Washita, Fort Cobb, and SoilSCAPE at 5 cm depth.

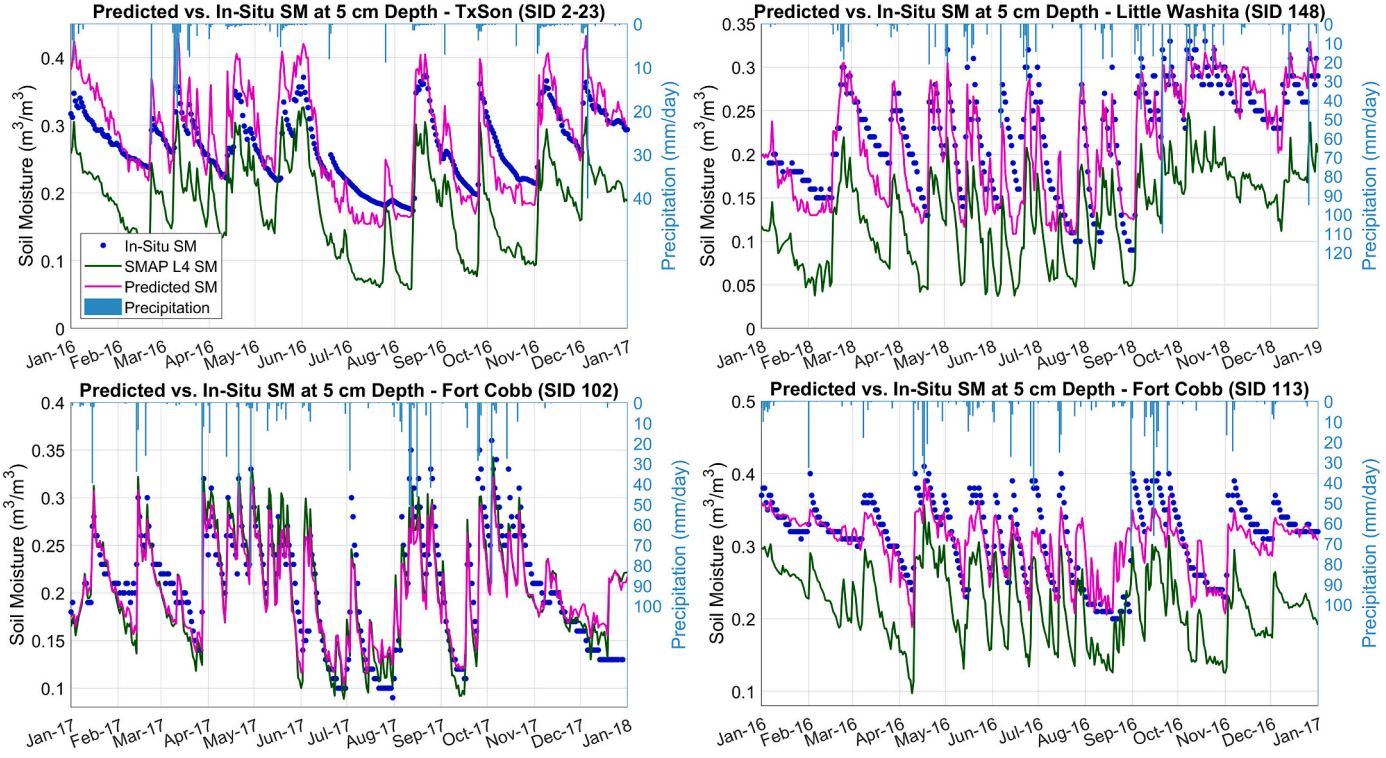


Fig. 8. Time series of 5 cm depth at different locations. Each plot contains in-situ, predicted, and SMAP L4 SM along with daily precipitation.

is noticed, especially during the dry-down periods due to the lower dynamic range of the SM estimates. This is noticed particularly in the SoilSCAPE sites for downscaled SM, which resulted in a relatively higher wet bias than other networks.

We compared the in-situ data, downscaled SM, and SMAP L4 SM (at 9 km) along with local precipitation to study the temporal dynamics of SM (Fig. 8). There is generally a good agreement between the observed and downscaled SM time series. The high SM variability (despite low rainfall conditions compared to other sites) associated with the station (TxSON site 2–23) is located in loam surface textured soil is adequately captured by the proposed method. The downscaled SM is noticed to follow the dry-down and wet-up patterns in agreement with the in-situ SM and precipitation intensities. During the August 2016 rainfall event at TxSON site 2–23, the in-situ data did not capture the increase in SM, whereas the proposed method depicted an increase in the moisture content. This indicates a possibility that the uncertainties in the predictor information (mainly soil texture and precipitation) can propagate into downscaled SM retrievals. We find that much of the dry bias existing in SMAP L4 surface SM is resolved by the downscaled SM product, except for some residual dry bias in dry-down periods from June to October 2016.

A similar observation is made in Little Washita site 148, wherein much of the dry bias of SMAP L4 SM is resolved by the downscaled SM. The proposed method could capture an increase in the moisture content due to frequent rainfall during the September and October months of 2018. Besides, faster dry-downs are observed in both SMAP L4 and downscaled timeseries, which is not the case with the in-situ data. This could be due to the SMAP SM's characteristic of rapid dry-down after the rainfall (Shellito et al., 2016). There is a strong agreement of downscaled and SMAP L4 SM timeseries with the in-situ data, in the case of Fort Cobb site 102. The downscaled SM could adjust some dry bias of SMAP L4 SM during the dry period of September 2017. The downscaled SM matched well with the in-situ data at Fort Cobb site 113. Compared to SMAP L4 SM, the proposed method improved the ubRMSE (~0.042 m^3/m^3 for SMAP L4) and bias (~0.10 m^3/m^3 for SMAP L4) metrics at this site.

In the case of 10 cm depth, validation could be carried out only using TxSON network. Fig. 9 presents the boxplots of performance metrics of predicted SM at 10 cm depth. Table S4 presents station-wise performance metrics along with the number of observations used per station for validation. There is an improvement in the ubRMSE at 10 cm depth compared to that of the surface (5 cm). The time series plots of downscaled and in-situ SM for four sample locations are presented in Fig. 10. There is a general agreement between the temporal variability of the two datasets. We noticed the downscaled SM be more reactive to the surface processes than in-situ data, for example, an increase in SM during rainfall events in August 2016 at 10–5 and 2–23 sites and in December 2017 at L-4, 2–17 sites. This could be attributed to SMAP surface SM variations, which are found to have the highest importance for estimating SM at 10 cm depth (Fig. 6). Besides, the correlated nature of SM at 5 cm and 10 cm depths, as discussed earlier, can be observed for TxSON 2–23 site downscaled SM (Figs. 8 & 10), wherein an agreement between several peak events is noticed. The greater agreement can also be due to the soil texture homogeneity in these two layers (presence of

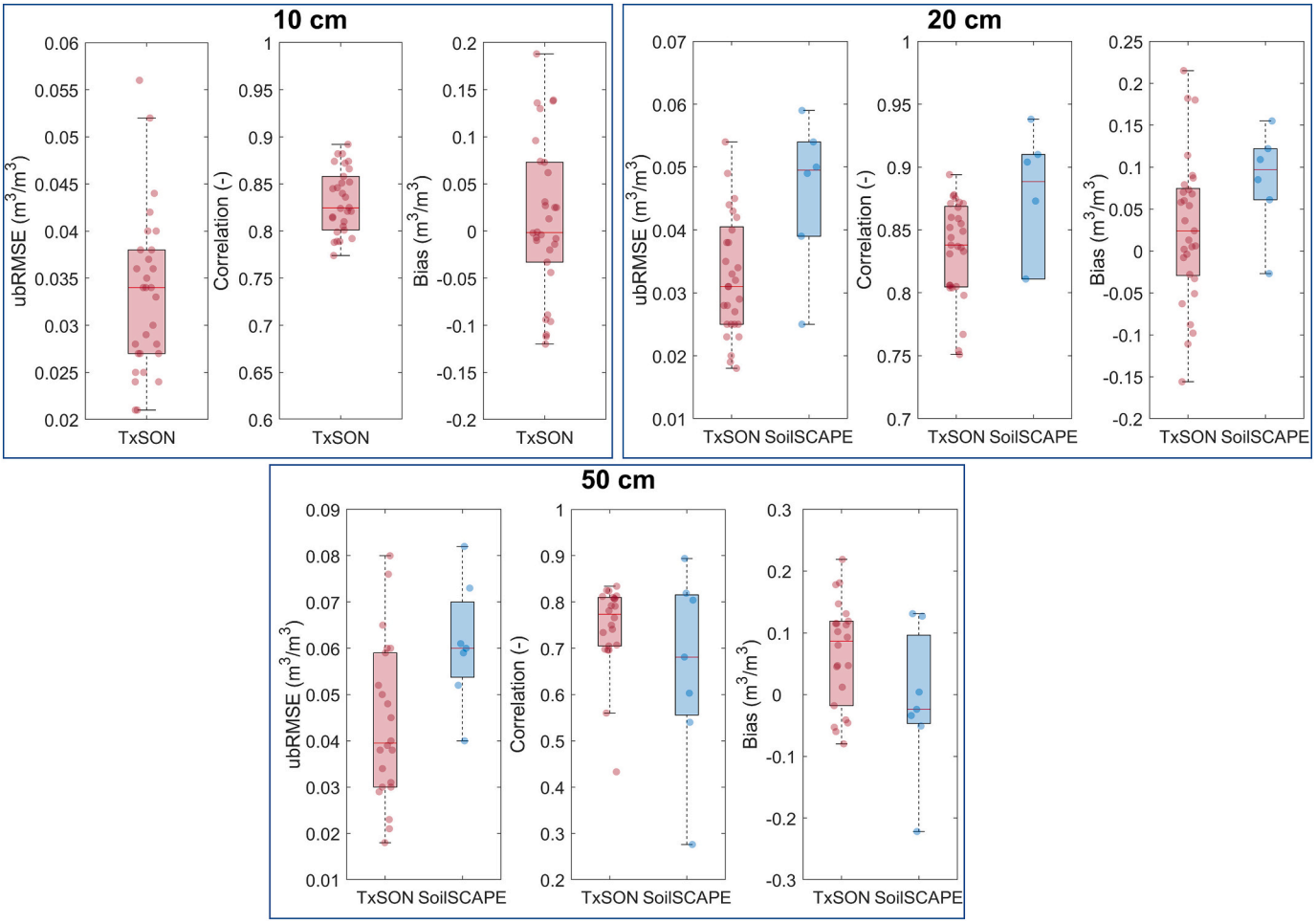


Fig. 9. Boxplots of ubRMSE, R, and bias computed for soil moisture predicted at 10, 20, and 50 cm depths at in-situ stations. Soil moisture at 10 cm depth is validated using TxSON network. Soil moisture at 20 and 50 cm depths are validated using TxSON and SoilSCAPE networks.

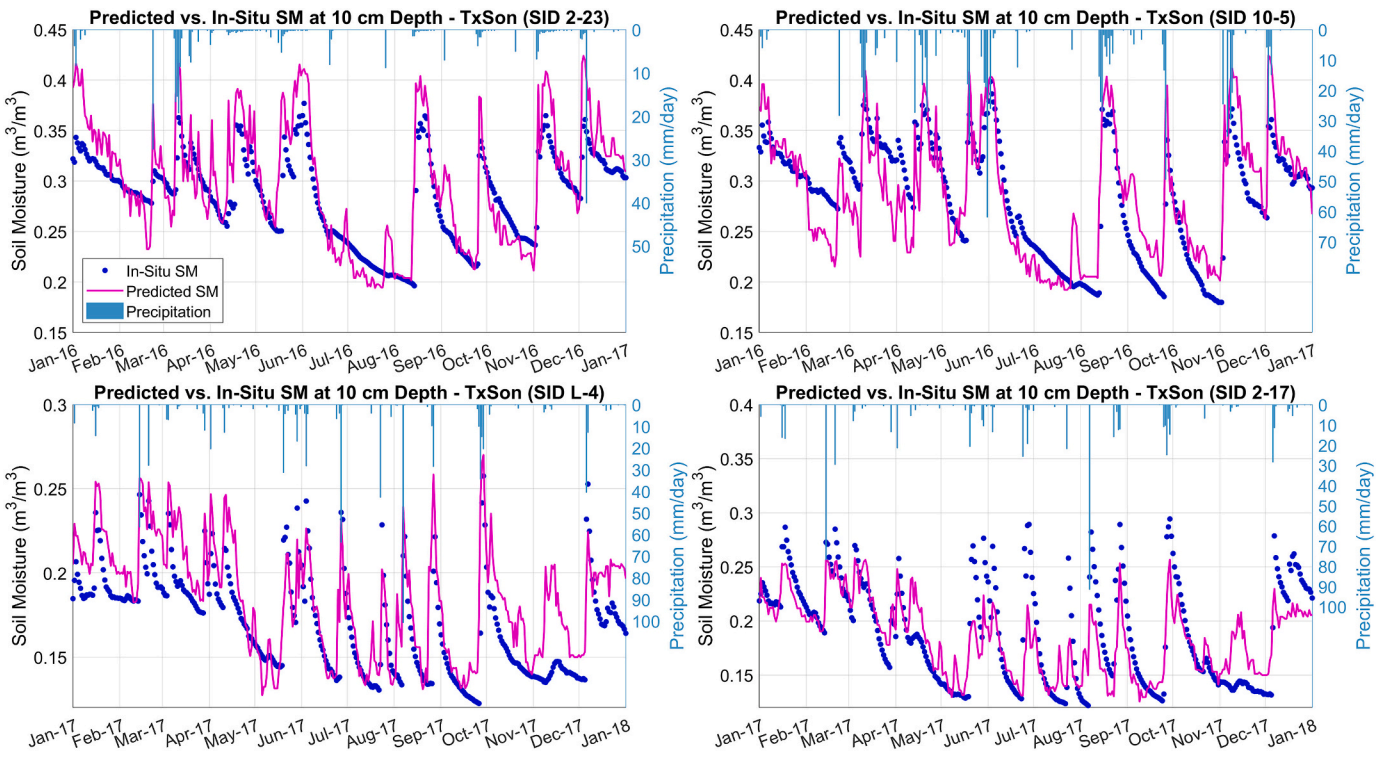


Fig. 10. Time series of 10 cm depth at different locations. Each plot contains in-situ and predicted SM along with daily precipitation.

clay soil). SM at 10 cm exhibited a lower dynamic range than that of 5 cm depth at this location. This could be due to SM's persistent behavior at 10 cm depth, although further investigations are needed on this aspect.

Fig. 9 presents the boxplots of performance metrics of validation of SM predicted at 20 cm depth. Table S5 presents these performance metrics along with number of observations used for validation at 29 and 6 stations of TxSON and SoilSCAPE networks, respectively. From 20 cm onward, the temporal dynamics of SM departed from surface SM dynamics. The downscaling algorithm extracts information of SM at 20 cm depth using rootzone SM as the most important predictor (Fig. 6). So, it would be important for the layer-wise estimates to capture the dynamic

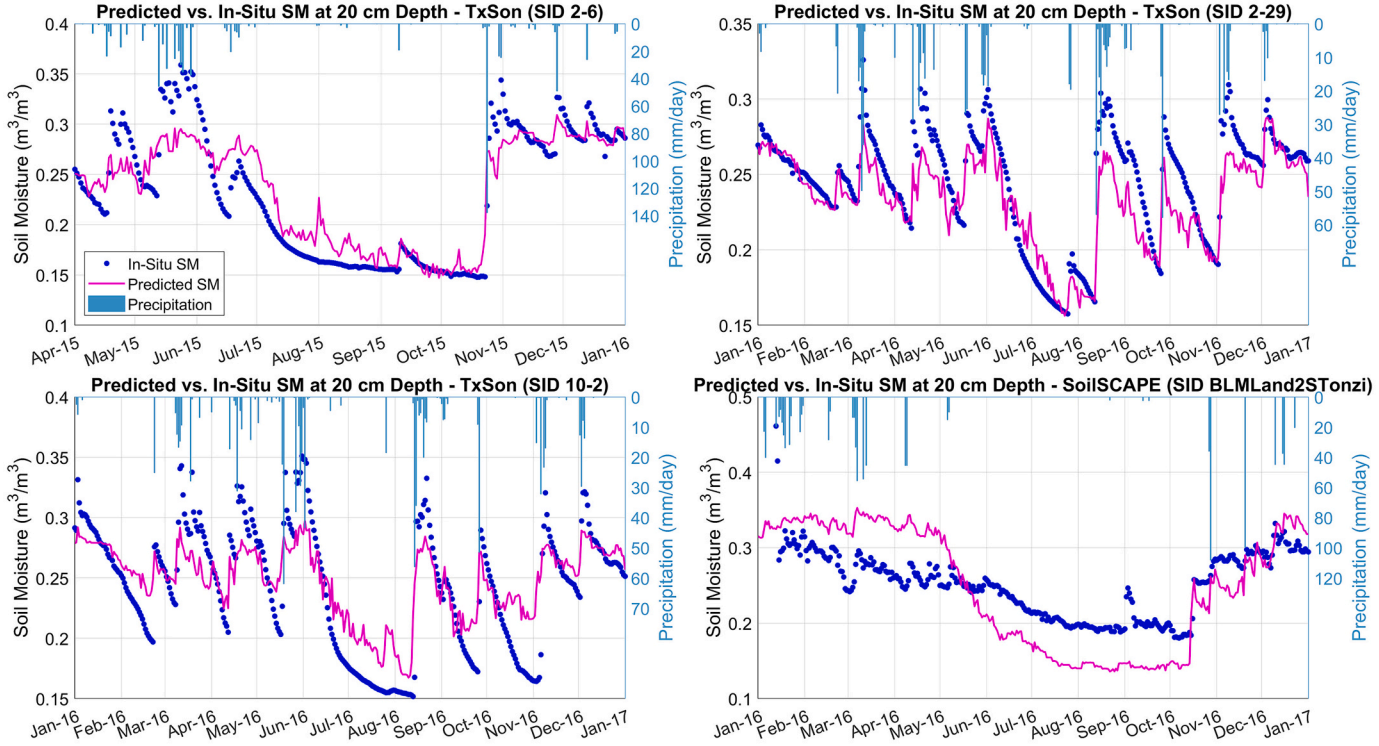


Fig. 11. Same as Fig. 10 but for soil moisture at 20 cm depth at different locations.

range of SM. Table 4 indicates that the downscaled SM has a noticeable bias with respect to in-situ data. Reichle et al. (2019) noted that the bias in rootzone SM does not imply a degradation of the result. If we set aside the bias component of the error, we find that the downscaled SM could predict the dynamic range of the SM reasonably well at 20 cm depth. To describe further, we plotted the downscaled bias-corrected (with respect to mean) SM against in-situ SM at six locations of TxSON in Fig. S3. Some of the sites indicated low ubRMSE values. However, it is partly contributed by the reduced dynamic range of rootzone SM of SMAP Level 4 data caused by the reduction of upward recharge in the CLSM (Reichle et al., 2019). In the case of SoilSCAPE stations, the performance is slightly degraded, particularly in terms of ubRMSE. This is because the downscaled SM could not predict the high variability during the peak SM periods. However, such a high variability is not noticed in TxSON stations. Apart from this issue, the proposed method depicted reasonable skill in modeling the dynamic range and temporal variations of the in-situ SM.

The timeseries of in-situ and downscaled SM (along with precipitation) at 20 cm depth at four locations are presented in Fig. 11. The timeseries plots indicated that the proposed method could estimate SM during the dry-down and wet-up periods with reasonable accuracy. The downscaled SM exhibited a greater high-frequency variability compared to that of in-situ data. This could be due to a) the internal variability of the machine learning model and b) model sensitivity to precipitation events. The downscaled SM exhibited a slightly lower dynamic range, with peaks being underestimated during the wet periods. However, it could successfully model the wet-up period during November 2015. The proposed method could reasonably well model the temporal dynamics of TxSON 2–29 site with accurate dry-downs. In the case of SoilSCAPE site, apart from biases in wet (January to May 2016) and dry (June to November 2016) periods, a faster dry-down is noticed during May 2016. This could be attributed to the sandy loam soil texture input at this location, which has a greater hydraulic conductivity, resulting in faster dry-down.

Fig. 9 presents the boxplots of validation performance metrics of SM predicted at 50 cm depth. The downscaled SM at 50 cm is validated at 22 and 7 stations of TxSON and SoilSCAPE networks, respectively. Table S6 presents station-wise performance metrics along with a number of observations used for validation. The performance of the downscaled product is lower compared to the results of 20 cm depth. This is mainly because of the differences in the dynamic range of downscaled and in-situ SM at this depth. It may be noted that soil texture properties have high importance while estimating SM at 50 cm depth (Fig. 6). Since PSU USDA STATSGO soil texture data was not available exactly at 50 cm depth, we selected a next layer available at 60 cm depth. Therefore, uncertainties in soil texture information can influence the estimates of SM. In their attempt to estimate SM at 20 cm and 50 cm depths, Pan et al. (2017) also found reduced skill of SM estimation at 50 cm compared to 20 cm. Besides, the low dynamic range of SMAP Level 4 SM could have also limited XGBoost's skill to estimate the complete range of SM at this depth. The timeseries plots at four TxSON sites at 50 cm depth are presented in Fig. 12. These plots indicate that the downscaled SM at 50 cm can broadly capture the temporal dynamics of in-situ SM. The diminished impact of meteorological inputs is evident through a reduced high-frequency variability in SM, which is observed estimates at shallower depths. Although the low dynamic range, as mentioned earlier, is evident in these plots dry-down and wet-up patterns of downscaled SM are consistent with that of the in-situ data. In the next section, we present the spatial maps of multi-layer downscaled SM.

3.5.2. *Spatial patterns of multi-layer downscaled SM*

For illustration, we present the multi-layer downscaled maps for 9th April 2015 (Fig. 13). The downscaled maps depict the general spatial characteristics of SM over the CONUS. Despite setting up a machine learning model cluster-wise and layer-wise, the spatial contiguity of SM patterns is attained, which indicates the efficacy of the proposed model. There is consistency with respect to the wetting caused by accumulated precipitation (Fig. S4). East CONUS (excluding Florida) and Northwest Coast CONUS experienced rainfall during 3–9 April 2015 (one week). Therefore, these regions are predominantly wet in the SM maps. Southwest and Central CONUS mostly did not receive any rainfall, resulting in dry conditions. The extent of drying reduced in these regions

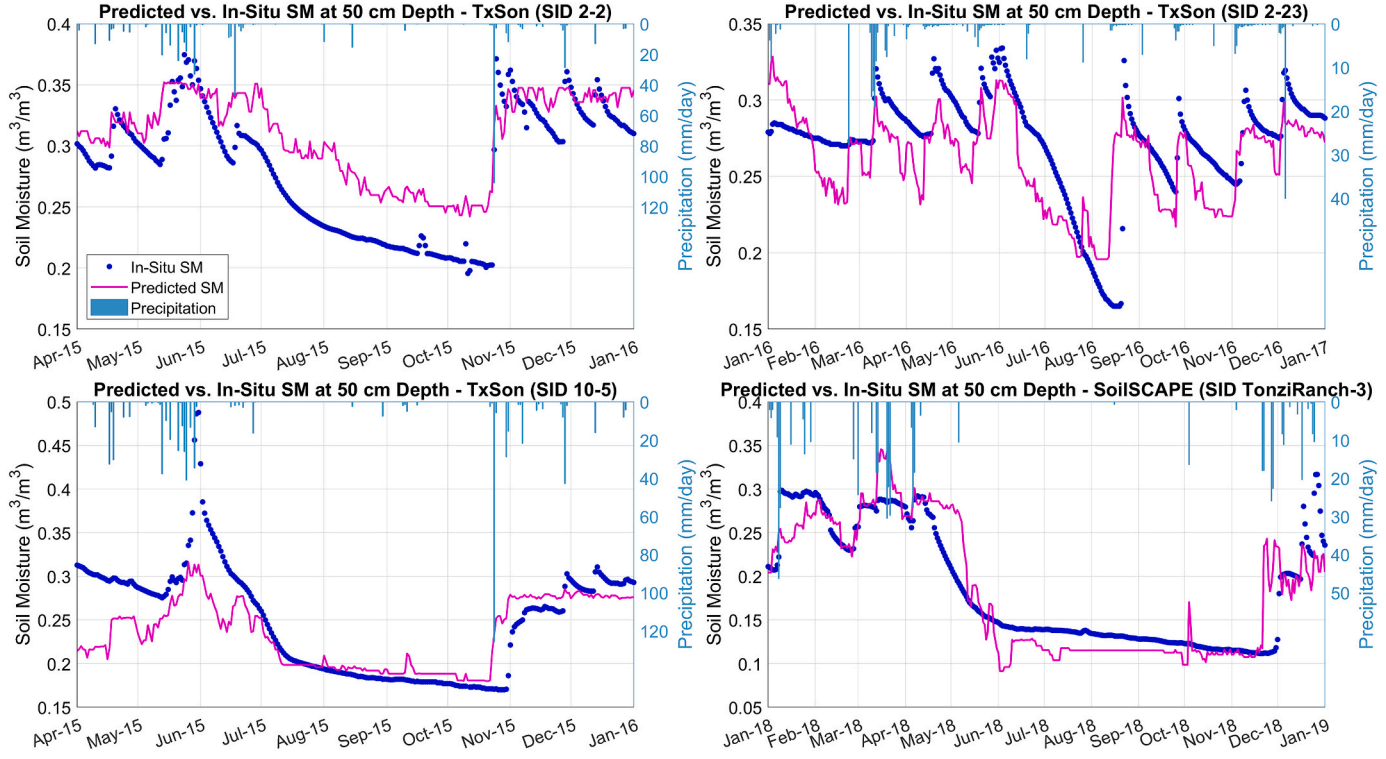


Fig. 12. Same as Fig. 10 but for soil moisture at 50 cm depth at different locations.

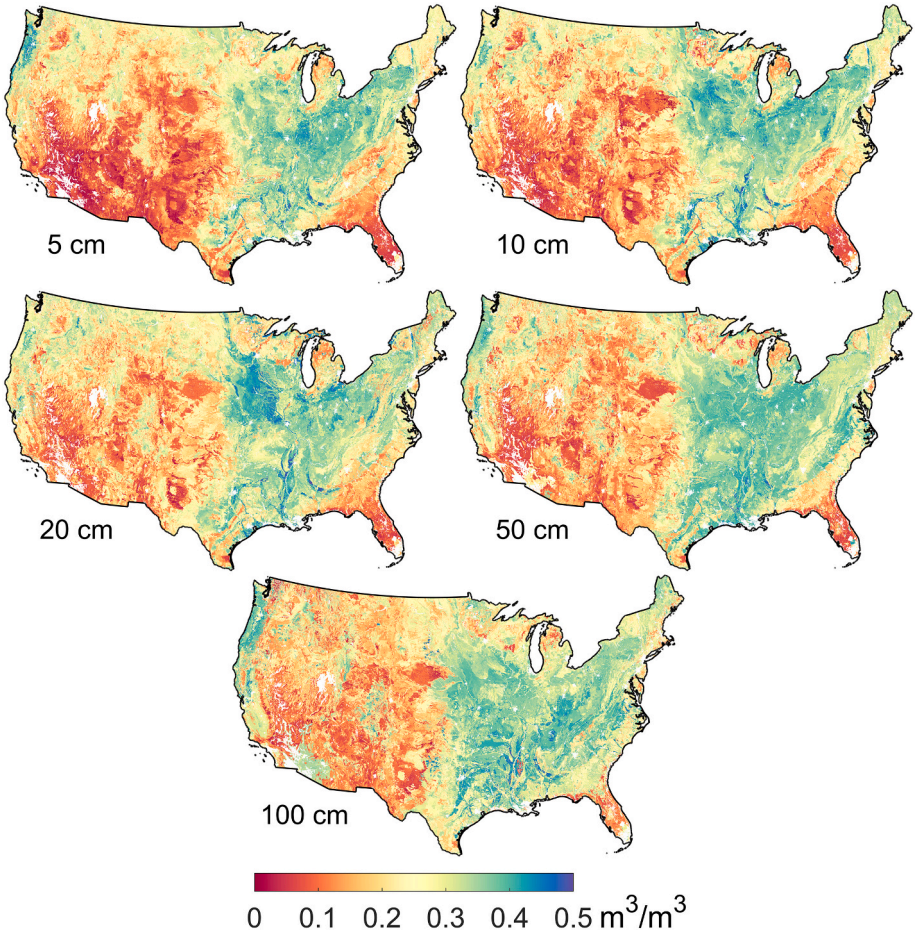


Fig. 13. High-resolution (1 km) multi-layer SM maps for five soil layers on 09th April 2015.

in the deeper layers, which could be due to the persistence behavior of rootzone SM. Besides, the spatial patterns of SM in the dry regions (mainly in West CONUS) varied significantly with changing depth. This could be due to the short hydrological length scales caused by decorrelation between the surface and rootzone SM (Akbar et al., 2018; Hirschi et al., 2014). The Midwest region experienced an increase in moisture content till 20 cm depth, followed by drying. A reverse mechanism is noticed in the Northwest coast, wherein the moisture content decreased till 20 cm depth, followed by an increase in the deeper layers. This could be possibly due to the vertical heterogeneity of soil texture and differences in the land cover conditions (croplands in the Midwest region and dense forests in the Northwest region), which result in variable rooting depths. These two factors affect the soil water movement (Fan et al., 2017).

The proposed method downscales the coarse resolution SMAP L4 SM product, primarily using fine resolution soil texture, elevation, and vegetation products. To assess the sub-grid heterogeneity and the spatial consistency with SMAP L4 product, we produced two high-resolution SM maps, one corresponding to the surface (0–5 cm) and another to rootzone (0–100 cm). The rootzone map is generated by computing the weighted average of high-resolution multi-layer SM maps. Fig. 14 presents these two maps along with SMAP L4 surface, and rootzone SM maps for 9th April 2015 plotted over a portion of Kansas state. The figure also presents the 2016 Land Use Land Cover (LULC) map obtained from USGS National Land Cover Database (NLCD) (Wickham et al., 2021). The region is selected considering its land cover heterogeneity. In the case of surface SM, the downscaling method could resolve the spatial heterogeneity of SM in the region. The predicted SM offers heterogeneity in low SM conditions prevailing in the west, and southwest portions of the map covered predominantly by grasslands (Fig. 14(a)). There is also a difference in the SM values of croplands compared to that of pastures and forested pixels. Such differences are not evident in the 9 km SMAP Level 4 map (Fig. 14(b)). For instance, the croplands along the Missouri River (upstream of Kansas City) and the west of Kansas City have distinguishable SM compared to other regions. Although the spatial heterogeneity is driven by high-resolution elevation, vegetation, and soil texture information, there is a greater influence of variations of surface SM, which influenced the spatial heterogeneity of high-resolution SM. Rootzone SM is found to have a lower SM dynamic range compared to surface SM. Cropland regions have a wetter rootzone SM compared to surface SM. The fine resolution features are prominent in high-resolution rootzone SM. This could be due to the higher relative importance of elevation and soil texture in some layers (Fig. 6). Notably, the spatial heterogeneity in rootzone SM broadly matches with LULC variations despite LULC not being used in the input feature space of machine learning models. Although it would be interesting to study the irrigation processes at fine resolution as attempted recently (Abbaszadeh et al., 2021), the irrigation signals may not be prominent since we use SMAP Level 4 product. It may be noted that before assimilation, SMAP data is bias corrected to match the dynamic range of CLSM SM. This process may suppress irrigation signals captured by SMAP SM alone (Lawston et al., 2017). We may include Level 3 satellite SM alongside the Level 4 assimilated SM product in the input feature space to capture such information. This step shall be attempted in future revisions of the algorithm.

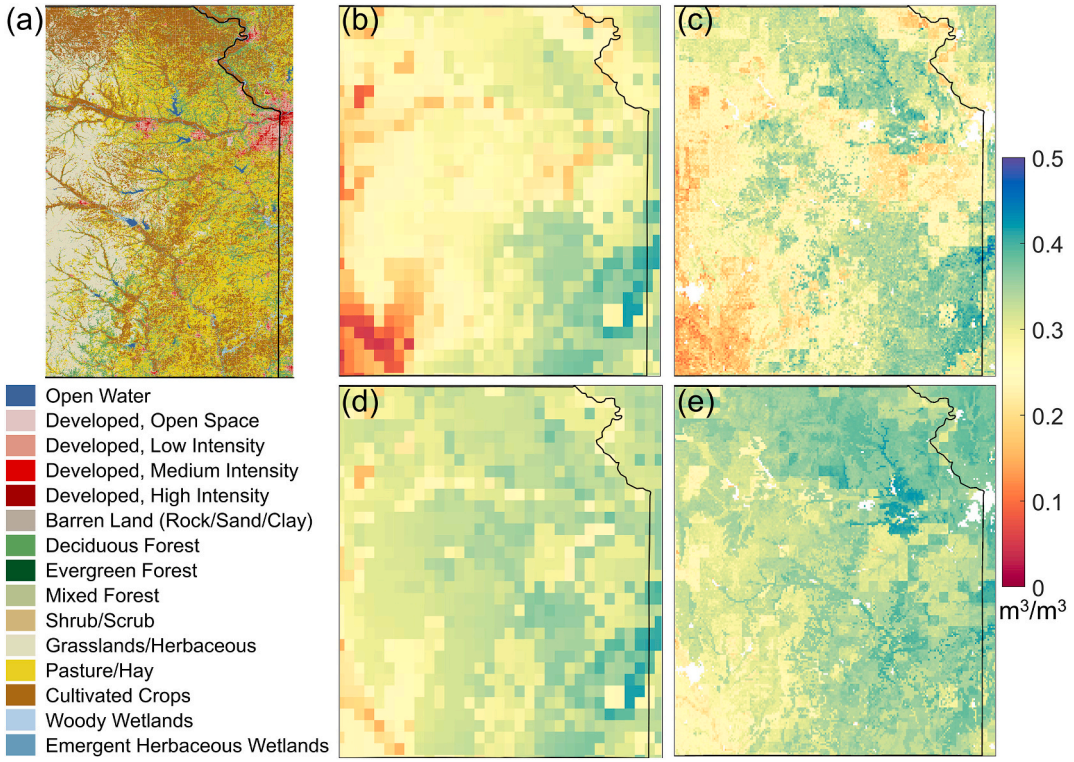


Fig. 14. Comparison of 1 km surface and root zone SM maps with SMAP L4 data on 9th April 2015 encompassing latitude range 37°N–40°N and longitude range 97.5°W–94.5°W (a portion of Kansas state). (a) 2016 LULC map from USGS NLCD (Wickham et al., 2021), (b) SMAP Level 4 9 km surface SM (0–5 cm), (c) Predicted 1 km surface SM, (d) SMAP Level 4 9 km rootzone SM (0–100 cm), (c) Predicted 1 km rootzone SM. Note: White portions in (c) and (e) represent either water bodies or built-up areas. The predominant white patch on the north east portion of the map is of Kansas City.

4. Summary and conclusions

Multi-layer soil moisture (SM) has several potential applications in the fields of water resources and agriculture. Knowledge of SM in the deeper layers is currently being estimated using a land surface model ingested with satellite surface SM in a data assimilation framework. However, such information limits to providing SM profile data (one representative value for the entire soil depth). Through this work, we attempt to discretize SM profile information at high-resolution (1 km) over the CONUS using multi-layer in-situ data and multiple variables that regulate SM dynamics using a machine learning modeling framework. The machine learning model is setup layer-wise in regions formed using various indicators that influence SM profile patterns.

The cluster analysis resulted in contiguous nature of homogenous regions with distinct geomorphological, topographical, meteorological, and vegetation characteristics. During the testing phase, the model performance is higher in arid regions. The error characteristics behavior mostly remains similar for 5 cm and 10 cm depths, which may be due to similarity in the temporal dynamics of SM at these two depths. The model performance is better at 20 cm depth compared to the deeper layers. Elevation, soil texture, and vegetation features are found to have greater importance in estimating SM. With the increase in depth, the importance of vegetation indicators reduced, and the importance of soil texture indicators increased. Meteorological indicators have the lowest importance across all soil layers. A reasonable performance was observed during the model validation stage for high-resolution multi-layer SM estimates. The models performed well for most of the locations across all the soil layers with ubRMSE less than 0.04 m^3/m^3. However, a slight decrease in correlation is noticed in the deeper layers compared to surface layers (5 cm and 10 cm depths). XGBoost algorithm could capture the temporal dynamics of SM in the deeper layers reasonably well. The new SM estimates could produce sub-grid heterogeneity owing to the high-resolution soil texture, elevation, and vegetation patterns. It could also accurately depict the spatial variability of SM.

4.1. Future scope

Achieving high-resolution rootzone SM is identified as a challenging task (Peng et al., 2020). This work proposes a method to address this challenge using the potential of machine learning tools. In terms of the model, improvements are necessary to achieve the dynamic range of SM more accurately in the deeper layers. Usage of dimensionality reduction techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) along with clustering algorithm and other widely used deep learning architectures such as Convolution Neural Networks (CNN) and LSTM models (Shen, 2018) can be explored.

The relative importance of predictors can be used for revising the predictors' set to improve the SM accuracy (e.g., omit precipitation and LST for SM prediction in deeper layers). To enable capturing irrigation water use at high resolution, the potentiality of including satellite SM in the input feature space shall be explored. This can improve the quality of predictions in data-scarce regions where SMAP can provide valuable SM information that cannot be accurately modelled by a land surface model (Dong et al., 2019). In addition, antecedent SM and precipitation conditions can be included in the input feature space to improve the accuracy of SM profile predictions (Pal and Maity, 2019; Pan et al., 2017). In the future, gap filled techniques that play a vital role in filling missing data in LST due to cloud cover (Long et al., 2020; Shiff et al., 2021) can be used to provide temporally consistent information in the input feature space. This step reduces control of XGBoost to handle missing data by itself and can improve the SM prediction accuracy. To gain a better understanding on the influence of above-described changes, a comprehensive uncertainty analysis concerning the choice of (a) input datasets, (b) clustering algorithms, (c) input feature space, (d) machine

learning algorithm, would be needed. The machine learning tools can capture the complex relationships between different hydrological processes to predict multi-layer SM estimates, however a large volume of data is required to train the models. In the future, we shall explore ways to implement this technique to other regions using limited in-situ information of rootzone soil moisture.

Data

SMAP Level 4 (Version 4: Vv4030) surface and rootzone soil moisture product is obtained from https://nsidc.org/data/spl4smau/versions/4/; Soil Texture data is obtained from http://www.soilinfo.psu.edu/; Elevation data is obtained from USGS GTOPO30 – https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30; MODIS NDVI/EVI MOD13A2 v006 product is obtained from https://lpdaac.usgs.gov/products/mod13a2v006/; MODIS GPP MOD17A2H v006 product is obtained from https://lpdaac.usgs.gov/products/mod17a2hv006/; MODIS LST Day and Night times MOD11A1 v006 product is obtained from https://lpdaac.usgs.gov/products/mod11a1v006/; CHIRPS v2.0 precipitation data is obtained from https://www.chc.ucsb.edu/data/chirps; In-situ soil moisture data is obtained from https://ismn.geo.tuwien.ac.at/en/; TxSON in-situ soil moisture data is obtained from https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/JJ16CF; Little Washita and Fort Cobb in-situ soil moisture data are obtained from http://ars.mesonet.org/; SoilSCAPE in-situ soil moisture data is obtained from https://daac.ornl.gov/LAND_VAL/guides/SoilSCAPE.html;

Declaration of Competing Interest

We have no conflict of interest to report.

Acknowledgments

This study was supported by the National Science Foundation (NSF, USA) award # 1653841 and 1841629. We thank the four anonymous reviewers and Dr. J.-P. Wigneron for their helpful comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2021.112706.

References

Abbaszadeh, P., Moradkhani, H., Zhan, X., 2019. Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. Water Resour. Res. 55, 324–344.
Abbaszadeh, P., Moradkhani, H., Gavahi, K., Kumar, S., Hain, C., Zhan, X., Duan, Q., Peters-Lidard, C., Karimiziarani, S., 2021. High-resolution SMAP satellite soil moisture product: exploring the opportunities. Bull. Am. Meteorol. Soc. 102, 309–315.
Abowarda, A.S., Bai, L., Zhang, C., Long, D., Li, X., Huang, Q., Sun, Z., 2021. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. Remote Sens. Environ. 255, 112301.
Akbar, R., Short Gianotti, D., McColl, K.A., Haghighi, E., Salvucci, G.D., Entekhabi, D., 2018. Hydrological storage length scales represented by remote sensing estimates of soil moisture and precipitation. Water Resour. Res. 54, 1476–1492.
Al Bitar, A., Mialon, A., Kerr, Y.H., Cabot, F., Richaume, P., Jacquette, E., Quesney, A., Mahmoodi, A., Tarot, S., Parrens, M., 2017. The global SMOS level 3 daily soil moisture and brightness temperature maps. Earth Syst. Sci. Data 9, 293–315.
Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J.-C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Piguet, B., Martin, E., 2008. From near-Surface to Root-Zone Soil Moisture Using an Exponential Filter: An Assessment of the Method Based on In-Situ Observations and Model Simulations.
Bablet, A., Viallefont-Robinet, F., Jacquemoud, S., Fabre, S., Briottet, X., 2020. High-resolution mapping of in-depth soil moisture content through a laboratory experiment coupling a spectroradiometer and two hyperspectral cameras. Remote Sens. Environ. 236, 111533.
Baroni, G., Ortuani, B., Facchi, A., Gandolfi, C., 2013. The role of vegetation and soil properties on the spatio-temporal variability of the surface soil moisture in a maize-cropped field. J. Hydrol. 489, 148–159.
Brown, M.E., Escobar, V., Moran, S., Entekhabi, D., O'Neill, P.E., Njoku, E.G., Doorn, B., Entin, J.K., 2013. NASA's soil moisture active passive (SMAP) mission and opportunities for applications users. Bull. Am. Meteorol. Soc. 94, 1125–1128.
Caldwell, T.G., Bongiovanni, T., Cosh, M.H., Jackson, T.J., Colliander, A., Abolt, C.J., Casteel, R., Larson, T., Scanlon, B.R., Young, M.H., 2019. The Texas soil observation network: a comprehensive soil moisture dataset for remote sensing and land surface model validation. Vadose Zone J. 18, 1–20.
Chawla, I., Karthikeyan, L., Mishra, Ashok, 2020. A review of remote sensing applications for water security: Quantity, quality, and extremes. Journal of Hydrology 585, 124826. https://doi.org/10.1016/j.jhydrol.2020.124826.
Chemura, A., Rwasoka, D., Mutanga, O., Dube, T., Mushore, T., 2020. The impact of land-use/land cover changes on water balance of the heterogeneous Buzi sub-catchment, Zimbabwe. Remote Sens. Appl. 18, 100292.
Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Crow, W.T., Berg, A.A., Cosh, M.H., Loew, A., Mohanty, B.P., Panciera, R., de Rosnay, P., Ryu, D., Walker, J.P., 2012. Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. Rev. Geophys 50.
Dari, J., Quintana-Seguí, P., Escorihuela, M.J., Stefan, V., Brocca, L., Morbidelli, R., 2021. Detecting and mapping irrigated areas in a Mediterranean environment by using remote sensing soil moisture and a land surface model. J. Hydrol. 596, 126129.
Das, N.N., Entekhabi, D., Njoku, E.G., 2010. An algorithm for merging SMAP radiometer and radar data for high-resolution soil-moisture retrieval. IEEE Trans. Geosci. Remote Sens. 49, 1504–1512.
Das, N.N., Entekhabi, D., Njoku, E.G., Shi, J.J., Johnson, J.T., Colliander, A., 2013. Tests of the SMAP combined radar and radiometer algorithm using airborne field campaign observations and simulated data. IEEE Trans. Geosci. Remote Sens. 52, 2018–2028.
Das, N.N., Entekhabi, D., Dunbar, R.S., Colliander, A., Chen, F., Crow, W., Jackson, T.J., Berg, A., Bosch, D.D., Caldwell, T., 2018. The SMAP mission combined active-passive soil moisture product at 9 km and 3 km spatial resolutions. Remote Sens. Environ. 211, 204–217.
Das, N.N., Entekhabi, D., Dunbar, R.S., Chaubell, M.J., Colliander, A., Yueh, S., Jagdhuber, T., Chen, F., Crow, W., O'Neill, P.E., 2019. The SMAP and Copernicus Sentinel 1A/B microwave active-passive high resolution surface soil moisture product. Remote Sens. Environ. 233, 111380.
Didan, K., 2015. MOD13A2 MODIS/Terra vegetation indices 16-Day L3 Global 1km SIN Grid V006 [NDVI, EVI]. *NASA EOSDIS LP DAAC*. https://doi.org/10.5067/MODIS/MOD13A2, 6.
Dirmeyer, P.A., Halder, S., 2016. Sensitivity of numerical weather forecasts to initial soil moisture variations in CFSv2. Weather Forecast. 31, 1973–1983.
Dong, J., Crow, W., Reichle, R., Liu, Q., Lei, F., Cosh, M.H., 2019. A global assessment of added value in the SMAP level 4 soil moisture product relative to its baseline land surface model. Geophys. Res. Lett. 46, 6604–6613.
Dorigo, W., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., 2011. International soil moisture network: a data hosting facility for global in situ soil moisture measurements. Hydrol. Earth Syst. Sci. 15.
Dumedah, G., Walker, J.P., Merlin, O., 2015. Root-zone soil moisture estimation from assimilation of downscaled soil moisture and ocean salinity data. Adv. Water Resour. 84, 14–22.
Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybernet. 4, 95–104.
Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., 2010a. The soil moisture active passive (SMAP) mission. Proc. IEEE *98*, 704–716.
Entekhabi, D., Reichle, R.H., Koster, R.D., Crow, W.T., 2010b. Performance metrics for soil moisture retrievals and application requirements. J. Hydrometeorol. 11, 832–840.
Famiglietti, J.S., Rudnicki, J.W., Rodell, M., 1998. Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas. J. Hydrol. 210, 259–281.
Famiglietti, J.S., Ryu, D., Berg, A.A., Rodell, M., Jackson, T.J., 2008. Field observations of soil moisture variability across scales. Water Resour. Res. 44.
Fan, Y., Miguez-Macho, G., Jobbágy, E.G., Jackson, R.B., Otero-Casal, C., 2017. Hydrologic regulation of plant rooting depth. Proc. Natl. Acad. Sci. *114*, 10572–10577.
Fang, K., Shen, C., 2020. Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. J. Hydrometeorol. 21, 399–413.
Fang, B., Lakshmi, V., Bindlish, R., Jackson, T.J., 2018. AMSR2 soil moisture downscaling using temperature and vegetation data. Remote Sens. 10, 1575.
Fang, B., Kansara, P., Dandridge, C., Lakshmi, V., 2021. Drought monitoring using high spatial resolution soil moisture data over Australia in 2015–2019. J. Hydrol. 594, 125960.
Felfelani, F., Pokhrel, Y., Guan, K., Lawrence, D.M., 2018. Utilizing SMAP soil moisture data to constrain irrigation in the community land model. Geophys. Res. Lett. 45, 12,892–812,902.
Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., Obersteiner, M., 2019. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. Agric. For. Meteorol. 264, 1–15.

Ford, T.W., Quiring, S.M., 2019. Comparison of contemporary in situ, model, and satellite remote sensing soil moisture with a focus on drought monitoring. Water Resour. Res. 55, 1565–1582.
Ford, T., Harris, E., Quiring, S., 2014. Estimating root zone soil moisture using near-surface observations from SMOS. Hydrol. Earth Syst. Sci. 18, 139–154.
Fujii, H., Koike, T., Imaoka, K., 2009. Improvement of the AMSR-E algorithm for soil moisture estimation by introducing a fractional vegetation coverage dataset derived from MODIS data. J. Remote Sens. Soc. Jpn. 29, 282–292.
Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci. Data 2, 1–21.
Gaur, N., Mohanty, B.P., 2016. Land-surface controls on near-surface soil moisture dynamics: traversing remote sensing footprints. Water Resour. Res. 52, 6365–6385.
Gavahi, K., Abbaszadeh, P., Moradkhani, H., Zhan, X., Hain, C., 2020. Multivariate assimilation of remotely sensed soil moisture and evapotranspiration for drought monitoring. J. Hydrometeorol. 21, 2293–2308.
Gruber, A., De Lannoy, G., Albergel, C., Al-Yaari, A., Brocca, L., Calvet, J.-C., Colliander, A., Cosh, M., Crow, W., Dorigo, W., 2020. Validation practices for satellite soil moisture retrievals: what are (the) errors? Remote Sens. Environ. 244, 111806.
Hirschi, M., Mueller, B., Dorigo, W., Seneviratne, S.I., 2014. Using remotely sensed soil moisture for land–atmosphere coupling diagnostics: the role of surface vs. root-zone soil moisture variability. Remote Sens. Environ. 154, 246–252.
Hoeben, R., Troch, P.A., 2000. Assimilation of active microwave observation data for soil moisture profile estimation. Water Resour. Res. 36, 2805–2819.
Jacobs, J.M., Mohanty, B.P., Hsu, E.-C., Miller, D., 2004. SMEX02: field scale variability, time stability and similarity of soil moisture. Remote Sens. Environ. 92, 436–446.
Joshi, C., Mohanty, B.P., 2010. Physical controls of near-surface soil moisture across varying spatial scales in an agricultural landscape during SMEX02. Water Resour. Res. 46.
Karl, T., Koss, W.J., 1984. Regional and National Monthly, Seasonal, and Annual Temperature Weighted by Area, 1895–1983.
Karthikeyan, L., Kumar, D.N., 2016. A novel approach to validate satellite soil moisture retrievals using precipitation data. J. Geophys. Res.-Atmos. 121, 11,516–511,535.
Karthikeyan, L., Pan, M., Wanders, N., Kumar, D.N., Wood, E.F., 2017a. Four decades of microwave satellite soil moisture observations: part 1. A review of retrieval algorithms. Adv. Water Resour. 109, 106–120.
Karthikeyan, L., Pan, M., Wanders, N., Kumar, D.N., Wood, E.F., 2017b. Four decades of microwave satellite soil moisture observations: part 2. Product validation and inter-satellite comparisons. Adv. Water Resourc 109, 236–252.
Karthikeyan, L., Chawla, I., Mishra, A.K., 2020. A review of remote sensing applications in agriculture for food security: crop growth and yield, irrigation, and crop losses. J. Hydrol. 124905.
Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Martinuzzi, J., Font, J., Berger, M., 2001. Soil moisture retrieval from space: the soil moisture and ocean salinity (SMOS) mission. IEEE Trans. Geosci. Remote Sens. 39, 1729–1735.
Kim, S., Zhang, R., Pham, H., Sharma, A., 2019. A review of satellite-derived soil moisture and its usage for flood estimation. Remote Sens. Earth Syst. Sci. 2, 225–246.
Koike, T., Nakamura, Y., Kaihotsu, I., Davaa, G., Matsuura, N., Tamagawa, K., Fujii, H., 2004. Development of an advanced microwave scanning radiometer (AMSR-E) algorithm for soil moisture and vegetation water content. Proc. Hydraul. Eng. *48*, 217–222.
Konapala, G., Mishra, A., 2020. Quantifying climate and catchment control on hydrological drought in the continental United States. Water Resour. Res. *56* e2018WR024620.
Kornelsen, K.C., Coulibaly, P., 2014. Root-zone soil moisture estimation using data-driven methods. Water Resour. Res. 50, 2946–2962.
Korres, W., Reichenau, T., Fiener, P., Koyama, C., Bogena, H.R., Cornelissen, T., Baatz, R., Herbst, M., Diekkrüger, B., Vereecken, H., 2015. Spatio-temporal soil moisture patterns–A meta-analysis using plot to catchment scale data. J. Hydrol. 520, 326–341.
Koster, R.D., Guo, Z., Yang, R., Dirmeyer, P.A., Mitchell, K., Puma, M.J., 2009. On the nature of soil moisture in land surface models. J. Clim. 22, 4322–4335.
Kotsiantis, S.B., 2013. Decision trees: a recent overview. Artif. Intell. Rev. 39, 261–283.
Kovačević, J., Cvijetinović, Ž., Stančić, N., Brodić, N., Mihajlović, D., 2020. New downscaling approach using ESA CCI SM products for obtaining high resolution surface soil moisture. Remote Sens. 12, 1119.
Kranz, W.L., Irmak, S., van Donk, S.J., Yonts, C.D., Martin, D.L., 2008. Irrigation Management for Corn. In: University of Nebraska-Lincoln Extension, Institute of Agriculture and Natural Resources.
Lawston, P.M., Santanello Jr., J.A., Kumar, S.V., 2017. Irrigation signals detected from SMAP soil moisture retrievals. Geophys. Res. Lett. 44, 11,860–811,867.
Li, X., Zhang, L., Weihermüller, L., Jiang, L., Vereecken, H., 2013. Measurement and simulation of topographic effects on passive microwave remote sensing over mountain areas: a case study from the Tibetan plateau. IEEE Trans. Geosci. Remote Sens. 52, 1489–1501.
Lievens, H., Reichle, R.H., Liu, Q., De Lannoy, G., Dunbar, R.S., Kim, S., Das, N.N., Cosh, M., Walker, J.P., Wagner, W., 2017. Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates. Geophys. Res. Lett. 44, 6145–6153.
Liu, D., Mishra, A.K., Yu, Z., 2016. Evaluating uncertainties in multi-layer soil moisture estimation with support vector machines and ensemble Kalman filtering. J. Hydrol. 538, 243–255.
Liu, Y., Jing, W., Wang, Q., Xia, X., 2020a. Generating high-resolution soil moisture by using spatial downscaling techniques: a comparison of six machine learning algorithms. Adv. Water Resour. 103601.
Liu, Y., Xia, X., Yao, L., Jing, W., Zhou, C., Huang, W., Li, Y., Yang, J., 2020b. Downscaling satellite retrieved soil moisture using regression tree-based machine learning algorithms over Southwest France. Earth Space Sci. *7* e2020EA001267.
Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., Quan, J., Meng, X., Shi, C., 2019. Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution. Remote Sens. Environ. 233, 111364.
Long, D., Yan, L., Bai, L., Zhang, C., Li, X., Lei, H., Yang, H., Tian, F., Zeng, C., Meng, X., 2020. Generation of MODIS-like land surface temperatures under all-weather conditions based on a data fusion approach. Remote Sens. Environ. 246, 111863.
Ma, Y., Feng, S., Song, X., 2013. A root zone model for estimating soil water balance and crop yield responses to deficit irrigation in the North China plain. Agric. Water Manag. 127, 13–24.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. Oakland, CA, USA.
Mao, H., Kathuria, D., Duffield, N., Mohanty, B.P., 2019. Gap filling of high-resolution soil moisture for SMAP/Sentinel-1: A two-layer machine learning-based framework. Water Resour. Res. 55, 6986–7009.
Massari, C., Brocca, L., Ciabatta, L., Moramarco, T., Gabellani, S., Albergel, C., De Rosnay, P., Puca, S., Wagner, W., 2015. The use of H-SAF soil moisture products for operational hydrology: flood modelling over Italy. Hydrology 2, 2–22.
McInnes, L., Healy, J., Melville, J., 2018. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint. arXiv:1802.03426*.
Merlin, O., Chehbouni, A., Boulet, G., Kerr, Y., 2006. Assimilation of disaggregated microwave soil moisture into a hydrologic model using coarse-scale meteorological data. J. Hydrometeorol. 7, 1308–1322.
Merlin, O., Al Bitar, A., Walker, J.P., Kerr, Y., 2010. An improved algorithm for disaggregating microwave-derived soil moisture based on red, near-infrared and thermal-infrared data. Remote Sens. Environ. 114, 2305–2316.
Miller, D.A., White, R.A., 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth Interact. 2, 1–26.
Mishra, A.K., Ines, A.V., Das, N.N., Khedun, C.P., Singh, V.P., Sivakumar, B., Hansen, J. W., 2015. Anatomy of a local-scale drought: application of assimilated remote sensing products, crop model, and statistical methods to an agricultural drought study. J. Hydrol. 526, 15–29.
Mishra, A.K., Vu, Tue, Veetil, Anoop Valiya, Entekhabi, Dara, 2017. Drought monitoring with soil moisture active passive (SMAP) measurements. Journal of Hydrology 552, 620–632. https://doi.org/10.1016/j.jhydrol.2017.07.033.
Moghaddam, M., Silva, A., Clewley, D., Akbar, R., Hussaini, S., Whitcomb, J., Devarakonda, R., Shrestha, R., Cook, R., Prakash, G., 2016. Soil Moisture Profiles and Temperature Data from SoilSCAPE Sites. USA. ORNL DAAC.
Montaldo, N., Albertson, J.D., 2003. Multi-scale assimilation of surface soil moisture data for robust root zone moisture predictions. Adv. Water Resour. 26, 33–44.
Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., Liu, J., 2020. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. J. Hydrol. 124901.
O, S., Orth, R., 2020. Global soil moisture from in-situ measurements using machine learning–SoMo. ml. *arXiv preprint. arXiv:2010.02374*.
Omernik, J.M., Griffith, G.E., 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environ. Manag. 54, 1249–1266.
Pal, M., Maity, R., 2019. Development of a spatially-varying statistical soil moisture profile model by coupling memory and forcing using hydrologic soil groups. J. Hydrol. 570, 141–155.
Pan, X., Kornelsen, K.C., Coulibaly, P., 2017. Estimating root zone soil moisture at continental scale using neural networks. JAWRA J. Am. Water Resourc. Assoc. 53, 220–237.
Parinussa, R.M., Holmes, T.R., Wanders, N., Dorigo, W.A., de Jeu, R.A., 2015. A preliminary study toward consistent soil moisture from AMSR2. J. Hydrometeorol. 16, 932–947.
Parinussa, R.M., Lakshmi, V., Johnson, F.M., Sharma, A., 2016. A new framework for monitoring flood inundation using readily available satellite data. Geophys. Res. Lett. 43, 2599–2605.
Peng, J., Loew, A., Zhang, S., Wang, J., Niesel, J., 2015. Spatial downscaling of satellite soil moisture data using a vegetation temperature condition index. IEEE Trans. Geosci. Remote Sens. 54, 558–566.
Peng, J., Loew, A., Merlin, O., Verhoest, N.E., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Rev. Geophys. 55, 341–366.
Peng, J., Albergel, C., Balenzano, A., Brocca, L., Cartus, O., Cosh, M.H., Crow, W.T., Dabrowska-Zielinska, K., Dadson, S., Davidson, M.W., 2020. A roadmap for high-resolution satellite soil moisture applications–confronting product characteristics with user requirements. Remote Sens. Environ. 112162.
Piles, M., Entekhabi, D., Camps, A., 2009. A change detection algorithm for retrieving high-resolution soil moisture from SMAP radar and radiometer observations. IEEE Trans. Geosci. Remote Sens. 47, 4125–4131.
Piles, M., Sánchez, N., Vall-llossera, M., Camps, A., Martínez-Fernández, J., Martínez, J., González-Gambau, V., 2014. A downscaling approach for SMOS land observations: evaluation of high-resolution soil moisture maps over the Iberian Peninsula. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 7, 3845–3857.
Qiu, Y., Fu, B., Wang, J., Chen, L., 2001. Spatial variability of soil moisture content and its relation to environmental indices in a semi-arid gully catchment of the Loess Plateau, China. J. Arid Environ. 49, 723–750.

Rahman, M., Di, L., Yu, E., Lin, L., Zhang, C., Tang, J., 2019. Rapid flood progress monitoring in cropland with NASA SMAP. Remote Sens. 11, 191.
Rebetez, M., 2001. Changes in daily and nightly day-to-day temperature variability during the twentieth century for two stations in Switzerland. Theor. Appl. Climatol. 69, 13–21.
Reichle, R.H., Liu, Q., Koster, R.D., JV, A., Colliander, A., Crow, W., De Lannoy, G.J., & Kimball, J., 2018. Soil Moisture Active Passive (SMAP) project assessment report for version 4 of the L4_SM data product. In: NASA Technical Report Series on Global Modeling and Data Assimilation. National Aeronautics and Space Administration, Goddard Space Flight Center, p. 67.
Reichle, R.H., Liu, Q., Koster, R.D., Crow, W.T., De Lannoy, G.J., Kimball, J.S., Ardizzone, J.V., Bosch, D., Colliander, A., Cosh, M., 2019. Version 4 of the SMAP Level-4 soil moisture algorithm and data product. J. Adv. Model. Earth Syst. 11, 3106–3130.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204.
Running, S., Mu, Q., Zhao, M., 2015. MOD17A2H MODIS/Terra Gross Primary Productivity 8-Day L4 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
Sabaghy, S., Walker, J.P., Renzullo, L.J., Jackson, T.J., 2018. Spatially enhanced passive microwave derived soil moisture: capabilities and opportunities. Remote Sens. Environ. 209, 551–580.
Sabater, J.M., Jarlan, L., Calvet, J.-C., Bouyssel, F., De Rosnay, P., 2007. From near-surface to root-zone soil moisture using different assimilation techniques. J. Hydrometeorol. 8, 194–206.
Sahoo, A.K., De Lannoy, G.J., Reichle, R.H., Houser, P.R., 2013. Assimilation and downscaling of satellite observed soil moisture over the Little River Experimental Watershed in Georgia, USA. Adv. Water Resour. 52, 19–33.
Schnur, M.T., Xie, H., Wang, X., 2010. Estimating root zone soil moisture at distant sites using MODIS NDVI and EVI in a semi-arid region of southwestern USA. Ecol. Informat. 5, 400–409.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: A review. Earth Sci. Rev. 99, 125–161.
Shellito, P.J., Small, E.E., Colliander, A., Bindlish, R., Cosh, M.H., Berg, A.A., Bosch, D.D., Caldwell, T.G., Goodrich, D.C., McNairn, H., 2016. SMAP soil moisture drying more rapid than observed in situ following rainfall events. Geophys. Res. Lett. 43, 8068–8075.
Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. 54, 8558–8593.
Shi, Y., Wu, P., Zhao, X., Li, H., Wang, J., Zhang, B., 2014. Statistical analyses and controls of root-zone soil moisture in a large gully of the Loess Plateau. Environ. Earth Sci. 71, 4801–4809.
Shiff, S., Helman, D., Lensky, I.M., 2021. Worldwide continuous gap-filled MODIS land surface temperature dataset. Sci. Data 8, 1–10.
Stefan, V.-G., Indrio, G., Escorihuela, M.-J., Quintana-Seguí, P., Villar, J.M., 2021. High-resolution SMAP-derived root-zone soil moisture using an exponential filter model calibrated per land cover type. Remote Sens. 13, 1112.
Teuling, A.J., Troch, P.A., 2005. Improved understanding of soil moisture variability dynamics. Geophys. Res. Lett. 32.
Tobin, K.J., Torres, R., Crow, W.T., Bennett, M.E., 2017. Multi-decadal analysis of root-zone soil moisture applying the exponential filter across CONUS. Hydrol. Earth Syst. Sci. 21, 4403–4417.
Ulaby, F., Long, D., 2015. Microwave Radar and Radiometric Remote Sensing. Artech House.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9.
Velpuri, N.M., Senay, G.B., Morisette, J.T., 2016. Evaluating new SMAP soil moisture for drought monitoring in the rangelands of the US high plains. Rangelands 38, 183–190.
Vereecken, H., Kamai, T., Harter, T., Kasteel, R., Hopmans, J., Vanderborght, J., 2007. Explaining soil moisture variability as a function of mean soil moisture: a stochastic unsaturated flow perspective. Geophys. Res. Lett. 34.
Wan, Z., Hook, S., Hulley, G., 2015. MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. 2015, distributed by NASA EOSDIS Land Processes DAAC. In.
Wang, X., Xie, H., Guan, H., Zhou, X., 2007. Different responses of MODIS-derived NDVI to root-zone soil moisture in semi-arid and humid regions. J. Hydrol. 340, 12–24.
Wang, T., Franz, T.E., Li, R., You, J., Shulski, M.D., Ray, C., 2017. Evaluating climate and soil effects on regional soil moisture spatial variability using EOF s. Water Resour. Res. 53, 4022–4035.
Wickham, J., Stehman, S.V., Sorenson, D.G., Gass, L., Dewitz, J.A., 2021. Thematic accuracy assessment of the NLCD 2016 land cover for the conterminous United States. Remote Sens. Environ. 257, 112357.
Wigneron, J.-P., Li, X., Frappart, F., Fan, L., Al-Yaari, A., De Lannoy, G., Liu, X., Wang, M., Le Masson, E., Moisy, C., 2021. SMOS-IC data record of soil moisture and L-VOD: historical development, applications and perspectives. Remote Sens. Environ. 254, 112238.
Wu, X., Walker, J.P., Rüdiger, C., Panciera, R., Gao, Y., 2016. Intercomparison of alternate soil moisture downscaling algorithms using active–passive microwave observations. IEEE Geosci. Remote Sens. Lett. 14, 179–183.
Wu, D., Wang, T., Di, C., Wang, L., Chen, X., 2020. Investigation of controls on the regional soil moisture spatiotemporal patterns across different climate zones. Sci. Total Environ. 138214.
Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 13, 841–847.
Zhang, H., Yang, Q., Shao, J., Wang, G., 2019a. Dynamic streamflow simulation via online gradient-boosted regression tree. J. Hydrol. Eng. 24, 04019041.
Zhang, R., Kim, S., Sharma, A., 2019b. A comprehensive validation of the SMAP enhanced Level-3 soil moisture product using ground measurements over varied climates and landscapes. Remote Sens. Environ. 223, 82–94.
Zhang, J., Liu, K., Wang, M., 2021a. Downscaling groundwater storage data in China to a 1-km resolution using machine learning methods. Remote Sens. 13, 523.
Zhang, R., Kim, S., Sharma, A., Lakshmi, V., 2021b. Identifying relative strengths of SMAP, SMOS-IC, and ASCAT to capture temporal variability. Remote Sens. Environ. 252, 112126.
Zhao, W., Sánchez, N., Lu, H., Li, A., 2018. A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression. J. Hydrol. 563, 1009–1024.