remote sensing

MDPI

Article

Estimating Rootzone Soil Moisture by Fusing Multiple Remote Sensing Products with Machine Learning

Shukran A. Sahaar and Jeffrey D. Niemann *

Department of Civil and Environmental Engineering, Colorado State University, Campus Delivery 1372, Fort Collins, CO 80523-1372, USA; shukran.sahaar@colostate.edu
* Correspondence: jeffrey.niemann@colostate.edu; Tel.: +1-970-491-3517

Abstract: This study explores machine learning for estimating soil moisture at multiple depths (0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm) across the coterminous United States. A framework is developed that integrates soil moisture from Soil Moisture Active Passive (SMAP), precipitation from the Global Precipitation Measurement (GPM), evapotranspiration from the Ecosystem Spaceborne Thermal Radiometer Experiment on Space Station (ECOSTRESS), vegetation data from the Moderate Resolution Imaging Spectroradiometer (MODIS), soil properties from gridded National Soil Survey Geographic (gNATSGO), and land cover information from the National Land Cover Database (NLCD). Five machine learning algorithms are evaluated including the feed-forward artificial neural network, random forest, extreme gradient boosting (XGBoost), Categorical Boosting, and Light Gradient Boosting Machine. The methods are tested by comparing to in situ soil moisture observations from several national and regional networks. XGBoost exhibits the best performance for estimating soil moisture, achieving higher correlation coefficients (ranging from 0.76 at 0–5 cm depth to 0.86 at 0–100 cm depth), lower root mean squared errors (from 0.024 cm^3/cm^3 at 0–100 cm depth to 0.039 cm^3/cm^3 at 0–5 cm depth), higher Nash–Sutcliffe Efficiencies (from 0.551 at 0–5 cm depth to 0.694 at 0–100 cm depth), and higher Kling–Gupta Efficiencies (0.511 at 0–5 cm depth to 0.696 at 0–100 cm depth). Additionally, XGBoost outperforms the SMAP Level 4 product in representing the time series of soil moisture for the networks. Key factors influencing the soil moisture estimation are elevation, clay content, aridity index, and antecedent soil moisture derived from SMAP.

Keywords: rootzone soil moisture; machine learning; SMAP; GPM; ECOSTRESS; artificial neural network; random forest; CatBoost; LightGBM; XGBoost

Citation: Sahaar, S.A.; Niemann, J.D. Estimating Rootzone Soil Moisture by Fusing Multiple Remote Sensing Products with Machine Learning. *Remote Sens.* **2024**, *16*, 3699. https://doi.org/10.3390/rs16193699

Academic Editors: Luca Brocca, David Fairbairn and Bertrand Bonan

Received: 12 August 2024
Revised: 18 September 2024
Accepted: 22 September 2024
Published: 4 October 2024

1. Introduction

Accurate knowledge of soil moisture is crucial for numerous applications, including agricultural water management [1,2], water resources sustainability [3], weather forecasting [4,5], climate modeling [6,7], wildfire prediction [8,9], and monitoring of floods and droughts [10,11]. Rootzone soil moisture is particularly important because it significantly influences plant growth [12], water availability [13], and ecological processes [14,15]. However, obtaining reliable rootzone soil moisture data at fine spatial resolutions (10–100 m grid cells) across large regions (10–100 km extents) remains challenging.

Several microwave satellite missions provide soil moisture nearly globally, including Soil Moisture Active Passive (SMAP) [16], the Advanced Scatterometer [17], Soil Moisture and Ocean Salinity (SMOS) [18], and the Advanced Microwave Scanning Radiometer for the Earth Observing System—Eos (AMSR-E) [19]. However, these datasets have limitations, including coarse spatial resolutions (often ranging from 9 to 60 km) and shallow depths of measurement (around 5 cm) [20]. SMAP also provides rootzone soil moisture estimates by merging the remote sensing information with modeling, but the spatial resolution remains coarse. Downscaling techniques can improve the spatial resolution of microwave soil moisture products. For example, Tagesson et al. [21] used the land surface temperature

and vegetation dryness index to disaggregate SMOS soil moisture from ~40 km resolution to ~5 km resolution in West Africa. Das et al. [22] used the Sentinel-1A and Sentinel-1B synthetic aperture radar data to disaggregate SMAP L-band radiometer measurements from ~40 km to 3 km and 1 km. Wei et al. [23] utilized the Moderate Resolution Imaging Spectroradiometer (MODIS) and a digital elevation model (DEM) with a gradient boosting decision tree to downscale SMAP soil moisture estimates from a 36 km to 1 km resolution across the Tibetan Plateau. Nuñez et al. [24] used MODIS products, sand fraction, and elevation to downscale AMSR2 soil moisture estimates from 25 km to 1 km in Puerto Rico. Vergopolan et al. [25] used a hyper-resolution land surface model, a radiative transfer model, and a Bayesian scheme to merge and downscale SMAP 36 km soil moisture to 30 m, and evaluated the results using four watersheds in the United States. Fischer [26] used topographic attributes and vegetation indices to downscale to 30 m, 10 m, and 3 m resolutions at Maxwell Ranch in Colorado. Fewer soil moisture downscaling methods have considered soil moisture beyond the top 5 cm. Dumedah et al. [27] used Disaggregation based on Physical and Theoretical scale Change (DisPATCh) to downscale satellite soil moisture data and estimate rootzone soil moisture at a spatial resolution of 1 km and depths of 0–30 cm, 30–60 cm, and 60–90 cm.

Soil moisture can also be estimated using optical and thermal data from satellites such as Landsat and the MODIS. These methods typically characterize the relationship between soil moisture and water-stressed vegetation using the visible and near-infrared bands, and they use the thermal infrared band to derive the relationship between rootzone soil moisture and soil thermal properties [28]. The methods provide soil moisture estimates at spatial resolutions (30 m to 1 km) over large spatial extents (185 km to 2330 km). The methods include the triangle and trapezoid methods [29–33], drought index method [34–38], thermal inertia method [39], single optical methods [40,41], energy balance methods [42–46], and synergistic optical/thermal and microwave methods [47–49]. Although acceptable accuracies have been reported, optical and thermal remote sensing methods have limitations. The triangle method, for instance, requires a flat surface, a large number of pixels, and a wide range of vegetation and moisture conditions, making it less effective in non-flat terrain and somewhat subjective in determining the warm edge and vegetation limits [29]. The drought index method calculates soil moisture retroactively and neglects temperature and rainfall effects on vegetation [28]. The thermal inertia method assumes consistent soil properties in both horizontal and vertical directions [28]. Optical methods can be precise for soil samples but are influenced by various factors like vegetation, atmospheric conditions, and topography, and they rely on empirical relationships [28].

An alternative soil moisture estimation approach fuses microwave remote sensing, optical and thermal remote sensing products, and ancillary datasets [50–56]. The ancillary datasets include variables that can impact soil moisture including antecedent soil moisture [57,58], landcover [59,60], soil properties [61,62], meteorological variables such as precipitation and land surface temperature [63,64], and topographic indices [65–68]. Many data fusion methods use machine learning algorithms [69–74], which are data-driven approaches that learn patterns and relationships from data without making assumptions about the processes that govern soil water dynamics [75,76]. Machine learning can merge large volumes of data from various sources, including in situ measurements, meteorological variables, and remote sensing datasets [69,77]. Machine learning algorithms can also perform feature selection, automatically identifying the inputs that are most relevant for estimating soil moisture [78]. Machine learning models have shown strong correlations between in situ soil moisture observations and the predicted soil moisture values [76,79]. For example, Abowarda et al. [70] employed a random forest (RF) model to produce surface soil moisture at a 30 m resolution for the Haihe Basin in northern China. SMAP Level 4 surface soil moisture was incorporated as the background field, and Landsat and MODIS data were used to determine the Normalized Difference Vegetation Index (NDVI), surface albedo, and land surface temperature. Precipitation and soil texture were also used as model inputs. The study reported root mean squared error (RMSE) values ranging from

0.031 to 0.050 cm^3/cm^3. Singh and Gaurav [80] used a feed-forward artificial neural network (ANN) with nine input variables derived from the Sentinel-1 and Sentinel-2 satellites as well as topographic characteristics from a DEM to estimate surface soil moisture at a spatial resolution of 60 m over the Kosi Fan in the Himalayan Foreland in the north Bihar plain, India. The study reported an RMSE value of 0.04 cm^3/cm^3. They also compared the ANN results to ten other machine learning methods and found that the ANN was most accurate. Zhao et al. [74] downscaled SMAP passive surface soil moisture (SSM) (0–5 cm depth) from 36 km to 1 km using a random forest (RF) method, reporting an R above 0.95 and an RMSE of 0.022 cm^3/cm^3. Fathololoumi et al. [73] also employed an RF method to downscale the Advanced Scatterometer (ASCAT) Soil Water Index (SWI) for the top 5 cm depth from a 10 km to 30 m resolution across three diverse field sites in the USA, France, and Iran, achieving RMSE values ranging from 0.072 to 0.172 cm^3/cm^3. Fewer studies have explored machine learning methods for rootzone soil moisture estimation. Fuentes et al. [81] used a deep learning approach to fuse SMAP, Sentinel-1, MODIS products (surface reflectance, land surface temperature, and land cover), and gridded soil properties to estimate soil moisture at a 90 m resolution for multiple depths. They used a multilayer perceptron model for the surface (0–10 cm) soil moisture and recurrent neural network model for 0–30 cm and 30–60 cm soil moisture at the Ozflux and Oznet networks across Australia. The study reported RMSE values of 0.073 cm^3/cm^3 for 0–10 cm and 0.070 cm^3/cm^3 for 0–30 cm and 30–60 cm. Karthikeyan and Mishra [82] used extreme gradient boosting (XGBoost) to estimate soil moisture at 5, 10, 20, 50, and 100 cm depths at a 1 km resolution and reported an unbiased root mean squared error (ubRMSE) of less than 0.040 cm^3/cm^3 for most locations.

Despite these recent advancements in using data fusion to estimate soil moisture, research gaps remain. Many studies have focused on surface soil moisture rather than rootzone soil moisture, and most studies have considered relatively small spatial extents. Prior studies have also used relatively few features as inputs and employed a single machine learning algorithm [70,80–82] without comparing its performance to other machine learning algorithms.

The primary objective of this study is to estimate soil moisture at five depths (0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm) using five machine learning methods (feed-forward ANN, random forest, XGBoost, Catboost, and LightGBM). The methods fuse microwave soil moisture data from SMAP, optical and thermal evapotranspiration products from the ECOSTRESS, vegetation data from the MODIS, precipitation data from the GPM, soil properties from gNATSGO, and land cover information from NLCD. The methods aim to estimate soil moisture for unobserved locations across the contiguous U.S. (CONUS). The machine learning methods are evaluated across eight in situ soil moisture networks that span arid to humid regions. This research also assesses the importance of individual predictor variables on the estimation of soil moisture.

2. Materials and Methods

2.1. Datasets

2.1.1. In Situ Soil Moisture Data

The in situ soil moisture data for training and evaluating the machine learning estimates were obtained from the freely available International Soil Moisture Network (ISMN). Only soil moisture observations for 0–102 cm depth were utilized as very few stations have deeper observations (approximately 0.1%). All datasets are available at the 1 h time step. To combine the datasets, soil moisture was estimated for five consistent depth ranges (0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm) using the weighted average method described by Gao et al. [41] and Liu et al. [83]. Non-uniform depth increments were used to align with the available SMAP products (0–5 cm and 0–100 cm) and to emphasize near surface conditions where more roots and variability occur.

The study period (January 2019 to December 2022) was selected based on the combined availability of all data used. Among the 1430 stations in CONUS, 801 stations are available

for this period and have soil moisture observations from 0 to 102 cm. All soil moisture values flagged by the ISMN [84], including those under frozen conditions and those outside the expected soil moisture range (<0.0 cm^3/cm^3 or >0.6 cm^3/cm^3), were excluded from the analysis. The final in situ dataset includes soil moisture measurements from 731 stations. These stations belong to eight different operational networks (Table 1). A higher density of gages occurs in the western U.S. due to the abundance of SNOTEL sites in that region (Figure 1).

Table 1. Number of stations utilized from each soil moisture network.

Network	Stations	Reference
Atmospheric Radiation Measurement Climate Research Facility (ARM)	17	Cook [85]
Cosmic-Ray Soil Moisture Observing System (COSMOS)	29	Zreda et al. [86]
AMERIFLUX	3	Baldocchi et al. [87]
Roaring Fork Observation Network (iRON)	8	Osenga et al. [88]
Soil Climate Analysis Network (SCAN)	168	Schaefer et al. [89]
Snow Telemetry (SNOTEL)	359	Fleming et al. [90]
Texas Soil Observation Network (TxSON)	40	Caldwell et al. [91]
U.S. Climate Reference Network (USCRN)	107	Bell et al. [92]

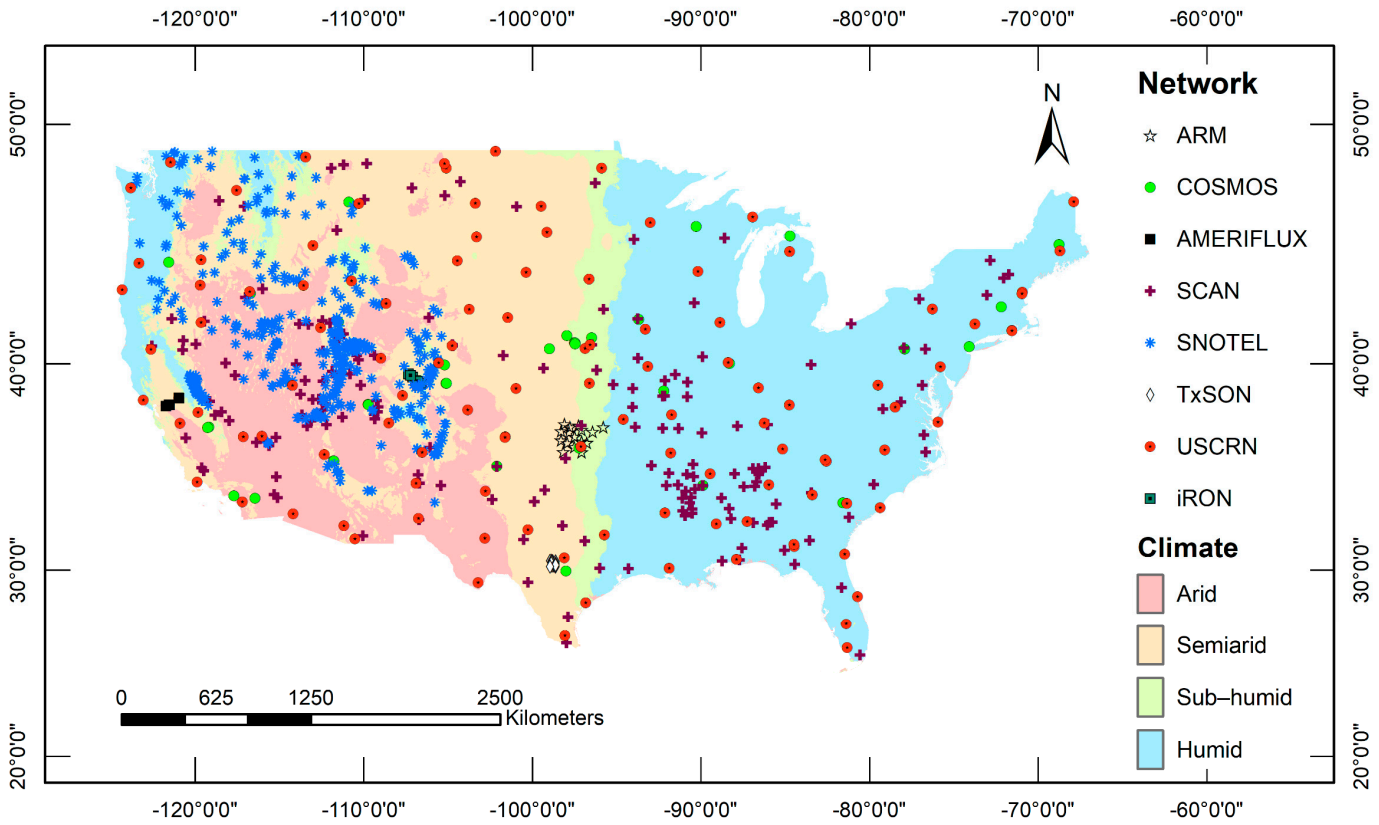

Figure 1. Locations and climates of the in situ soil moisture stations used in this study.

2.1.2. Satellite Soil Moisture Data

The seventh version of the SMAP Level-4 (SPL4SMGP.007) [93] surface soil moisture (SSM) (0–5 cm), rootzone soil moisture (RZSM) (0–100 cm), and profile soil moisture (PSM) (0-bedrock depth) products were used as inputs to the machine learning algorithms (Table 2). These products are generated using a land data assimilation system that combines satellite-based L-band brightness temperature measurements, precipitation observations, and land surface modeling [94]. The SMAP Level-4 products were used because they provide

complete spatial and temporal coverage and include soil moisture values throughout the rootzone.

Table 2. Datasets used in this study and their spatial and temporal resolutions. Most of the satellite data were obtained and processed using the Land Processes Distributed Active Archive Center (LPDAAC) (https://lpdaac.usgs.gov/, accessed on 1 June 2023) and Google Earth Engine (GEE) [95,96].

Data Type	Variables	Product/Access	Spatial Resolution	Temporal Resolution	Reference
In Situ Soil Moisture	0–5 cm 0–10 cm 0–20 cm 0–50 cm 0–100 cm	Soil moisture/ ISMN	Point or field measurements	Hourly	Dorigo et al. [97]
Satellite Soil Moisture	SSM (0–5 cm) RZSM (0–100 cm) PSM (0–bedrock)	SMAP SPL4SMGP.007/ GEE	9 km	3 h	Reichle et al. [93]
Landcover and Vegetation	NLCD	NLCD 2019/GEE	30 m	Static	Dewitz [98,99]
	NDVI EVI	MOD13Q1.061/ LPDAAC	250 m	16 Days	Didan [100]
	LAI fPAR	MCD15A3H.061/ LPDAAC	500 m	4 Days	Myneni et al. [101]
Soil	Sand Silt Clay Organic Matter Bulk Density Electrical Conductivity pH Depth to Restrictive Layer	Soil layers/ USDA-NRCS (gNATSGO)	30 m	Static	Soil Survey Staff [102]
Weather and Climate	Precipitation Measurement	GPM_3IMERGHH/ GEE	11 km	Half-hourly	Huffman et al. [103]
	Instantaneous LST	ECO_L2_LSTE v002/LPDAAC	70 m	Varies	Hook and Hulley [104]
	Instantaneous ET	ECO3ETPTJPLv001/ LPDAAC	70 m	Varies	Hook and Fisher [105]
	Instantaneous ESI, Instantaneous PET	ECO4ESIPTJPLv001/ LPDAAC	70 m	Varies	Hook and Fisher [105]
	Aridity Index AI	Climate Database v3/ CGIAR-CSI	1 km	Static	Zomer et al. [106]
Topography	Elevation Slope Aspect Hillshade	SRTMGL1 v003/ GEE	30 m	Static	NASA-JPL [107]
	mTPI	Global SRTM mTPI/GEE	270 m	Static	Theobald et al. [108]

Higher antecedent soil moisture leads to slower initial infiltration rates and more rapid declines in infiltration rates through time [109]. Thus, antecedent moisture may be predictive of current moisture. The SMAP surface, root zone, and profile soil moisture data were used to calculate antecedent soil moisture values for each hour using a moving average with windows of 1 day, 3 days, 7 days, and 14 days.

2.1.3. Land Cover and Vegetation Data

The 2019 National Land Cover Database (NLCD) was used to characterize the land cover type. The 2019 dataset was used because it aligns with the start of the study period and contained the most recent data available when the analysis was performed. NLCD contains 16 landcover classes at a spatial resolution of 30 m (https://www.mrlc.gov/data/nlcd-2019-land-cover-conus, accessed on 1 June 2023) [99]. To characterize the density

and photosynthetic activity (greenness) of the vegetation cover, the following indices were used: Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Leaf Area Index (LAI), and Fraction of Photosynthetically Active Radiation (fPAR). The NDVI and EVI were obtained from MODIS Vegetation Indices (MOD13Q1) Version 6.1 (Table 2). Higher NDVI and EVI values indicate thicker and/or greener vegetation. The NDVI uses the red and near infrared bands, while the EVI also uses the blue band. The EVI is less sensitive to atmospheric conditions than the NDVI and is therefore preferred if aerosol content is high or soil/background influences are significant [110]. The LAI and fPAR were obtained from MODIS Vegetation Indices (MCD15A3H) Version 6.1 (Table 2). The LAI is the total one-sided green leaf surface area per unit ground area, while the fPAR is the fraction of photosynthetically active radiation that is absorbed by vegetation [111]. The fPAR is an indicator of the water, energy, and carbon balance that plants require for photosynthesis [112]. The LAI typically ranges from 0 for no vegetation to >5 for dense forests. The fPAR ranges between 0 for no vegetation and 1 for dense, healthy vegetation [111].

2.1.4. Soil Data

The Gridded National Soil Survey Geographic (gNATSGO) dataset was used to obtain percent sand, silt, and clay, organic matter, bulk density, electrical conductivity (EC), pH, and depth to restrictive layer (Table 2). EC is an indicator of soil salinity, which can hinder water uptake by plants, and soil pH can affect nutrient availability and thus ET [113,114]. All soil properties except the depth to restrictive layer were calculated for the five depth ranges: 0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm. gNATSGO is a composite of the Soil Survey Geographic (SSURGO) dataset (mostly 1:24,000 scale), State Soil Geographic 2 (STATSGO2) dataset (1:250,000 scale), and the detailed Raster Soil Surveys (RSS) dataset [115]. The gNATSGO dataset was obtained from the Natural Resources Conservation Service (https://www.nrcs.usda.gov/resources/data-and-reports/gridded-national-soil-survey-geographic-database-gnatsgo, accessed on 1 June 2023).

2.1.5. Weather and Climate Data

The weather and climate were characterized using precipitation, land surface temperature (LST), evapotranspiration (ET), the evaporative stress index (ESI), potential evapotranspiration (PET), and the aridity index (AI). Precipitation data were obtained from Integrated Multi-Satellite Retrievals for GPM (IMERG) Final Precipitation Level 3 Half Hourly (GPM_3IMERGHH) (Table 2). The GPM was chosen because it has been applied in other soil moisture studies [116–118] and has high temporal resolution [103], which helps quantify the short-term precipitation effects on soil moisture. The GPM data were used in moving averages to calculate antecedent precipitation for 6 h, 1 day, 3 days, 7 days, and 14 days (denoted as P6H, P1DAY, P3DAY, P7DAY, and P14DAY, respectively).

The LST, ET, ESI, and PET were obtained from the ECOSTRESS (Table 2). The LST product is derived from five thermal infrared bands. The actual ET product estimates instantaneous ET using the Priestly–Taylor Jet Propulsion Laboratory algorithm [105], which uses a series of eco-physiological scaling functions to reduce the potential ET to the actual ET [119]. The ESI is the ratio of ET and PET, which is an indicator of plant water stress [119].

The AI was obtained from the Global AI and PET Database v3 (Global-AI_PET_v3), which was developed by Trabucco and Zomer [120] (Table 2). The AI is defined as the ratio of the mean annual precipitation to mean annual PET. The PET is based on the FAO Penman–Monteith reference crop evapotranspiration equation [121]. The data were obtained from CGIAR-Consortium for Spatial Information (CGIAR-CSI, https://csidotinfo.wordpress.com, accessed on 1 June 2023).

2.1.6. Topographic Data

The elevation, slope, aspect, and hillshade values were determined from the Shuttle Radar Topography Mission Global 1-arc second (SRTMGL1) DEM in Google Earth Engine (GEE). Higher elevations tend to have lower temperatures, which reduce ET and increase soil moisture [68]. Lower slopes can reduce lateral hydraulic gradients, promoting higher rootzone soil moisture [122]. Aspect affects insolation and thus ET and soil moisture [66,67,123]. Hillshade describes the relative shading of a location and depends on the variations in elevation across the landscape as well as the sun's azimuth and altitude angles. Higher hillshade values indicate more shading [124], which can affect soil moisture [125–127]. For simplicity and consistency, 270° (due south) and 45° were used for the azimuth and altitude angles, respectively, for all locations. The Shuttle Radar Topography Mission multi-scale Topographic Position Index (SRTM_mTPI) was used to measure the elevation of a location relative to its surrounding area (Table 2). mTPI is calculated by subtracting the mean elevation of a 3 × 3 pixel neighborhood from the elevation of the central point (the point of interest). It distinguishes peaks, ridges, plains, and valleys [108]. Other topographic attributes such as the drainage area and topographic wetness index have been shown to influence soil moisture [26,127] but were not included because they were not available in GEE.

2.1.7. Data Preprocessing

All remote sensing datasets were projected to the NAD_1983_2011 CONUS Albers projection and resampled to 70 m to match the ECOSTRESS resolution. For example, 30 m topographic attributes were resampled to 70 m using inverse distance weighting. The 9 km SMAP data were resampled to 70 m using the nearest neighbor method. The nearest neighbor approach retains the same soil moisture value across each subdivided 9 km grid cell. Then, the values from the resampled products were obtained at each in situ soil moisture location. The temporal resolution was also based on the ECOSTRESS. The ECOSTRESS had the most missing data, so only dates with available ECOSTRESS data were considered in the study. The missing values in the other datasets were estimated by linearly interpolating in time. For example, the LAI is available every four days, so it was linearly interpolated to create hourly values and these hourly values were resampled to match the ECOSTRESS timestamps. The final dataset has 42 columns. The columns are associated with the 41 predictor (input) variables including a categorical depth column that shows the depth range associated with the in situ soil moisture, and the in situ soil moisture at that depth (the dependent variable). This structure allows the depth range to have its own set of input variables in the machine learning models. The dataset has 72,233 rows, where each row represents a station and time.

The Python 3.11 library sklearn.preprocessing was used to preprocess the data [128]. The categorical depth columns (0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm) and NLCD (landcover type) were encoded using LabelEncoder and OneHotEncoder, respectively [129]. LabelEncoder assigns a unique numerical value to categorical data, and OneHotEncoder creates a new binary feature for each possible value of the categorical feature.

2.2. Machine Learning Algorithms

Five machine learning methods were used: feed-forward ANN, RF, XGBoost, Catboost, and LightGBM. The methods were utilized as regression tools (they can also be used as classification tools). These machine learning methods were selected because they are well suited for modeling complex nonlinear relationships, handling high-dimensional data, processing large datasets efficiently, and capturing variable interactions [130–132]. All five methods have been used previously for estimating soil moisture in some manner [82,133–137]. The RF, XGBoost, Catboost, and LightGBM methods are used to determine the importance of each predictor variable to the prediction [128,129,138,139].

2.2.1. Feed-Forward Artificial Neural Network (ANN)

ANNs [137] are inspired by the structure and functioning of the human brain. An ANN consists of interconnected nodes or neurons that are organized into layers. An input layer receives preprocessed data, one or more hidden layers process the data through weighted connections, and an output layer generates the predictions. In the feed-forward ANN, the information flows from the input layer to the output layer without any feedback loops. Activation functions, which are applied to each neuron's output, introduce nonlinearity to the network, enabling it to capture more complex relationships in the data. L2 regularization is used to prevent overfitting in the machine learning model [138]. During the training process, the weights of the connections between neurons are adjusted iteratively through a backpropagation process to minimize the discrepancy between the predicted outputs and the actual target values [139]. The optimization is achieved using a gradient descent algorithm, which updates the weights to reduce the prediction error as quantified by the loss function.

2.2.2. Random Forest (RF)

RF uses decision trees to make predictions [131]. The algorithm first selects a random sample of the training data with replacement. This means that some observations may be included multiple times in the sample, while others may not be included at all. A decision tree is then grown on each sample. When growing a decision tree, the algorithm randomly selects a subset of the predictor variables to consider at each node. This helps to ensure that the trees are diverse and not too correlated with each other. The predictions of the individual trees are then averaged to produce the final prediction. The averaging process improves the robustness and generalization of the model as it reduces the impact of individual tree outliers or overfitting [131]. An importance score (indicating the relative importance of a given predictor variable) is determined by averaging the reduction in impurity or the decrease in accuracy when the feature values are permuted across all trees [131].

2.2.3. Extreme Gradient Boosting (XGBoost)

Like RF, XGBoost is an ensemble learning algorithm that utilizes decision trees. However, XGBoost employs a gradient boosting framework where the decision trees are sequentially trained with each tree aiming to correct the errors made by the preceding ones. This sequential training process, combined with regularization techniques, focuses on minimizing the residual errors and enhancing predictive performance. Thus, unlike RF, XGBoost trees are not constructed independently, and strong interplay occurs between the trees [132]. The importance score for each predictor variable is calculated based on the number of times a feature is used to split the data and the associated gain in model accuracy [132].

2.2.4. Categorical Boosting (CatBoost)

CatBoost [140] is a gradient boosting decision tree algorithm like XGBoost, but it differs from XGBoost in its approach to training weak learners. CatBoost uses a greedy algorithm (a greedy algorithm iteratively makes the most optimal local decision at each step with the objective of eventually converging to the global optimum) to effectively combine categorical features and their interactions, and it utilizes a prior value to reduce noise from infrequent categories [140,141]. This allows CatBoost to learn more complex relationships between categorical inputs and the target variable [140]. CatBoost also employs ordered boosting, which trains a model for each sample in the training dataset to estimate the gradient of the loss function [140]. These gradient estimates are then aggregated to construct the final model. Ordered boosting improves gradient estimation precision and reduces the risk of overfitting [141]. The importance score is determined using the change in the loss function when features are included or permuted [140].

2.2.5. Light Gradient Boosting Machine (LightGBM)

LightGBM [142] is a gradient boosting decision tree algorithm that is designed to be fast and memory-efficient. It is similar to XGBoost but incorporates techniques to enhance performance. One of the key features of LightGBM is gradient-based one-side sampling, which selectively excludes data instances with small gradients to reduce the sample size and computational demands [142]. LightGBM also integrates exclusive feature bundling, which groups highly correlated features together, reducing the number of input variables that need to be considered during decision tree growth. This further enhances computational efficiency [142]. The importance score is determined by calculating the overall improvement in accuracy from all splits that involve each feature across all trees in the model [142].

2.3. Model Training and Evaluation

All five machine learning methods are provided the same 41 input (predictor) variables. The predictor variables include the satellite soil moisture, landcover/vegetation, soil, weather/climate, and topographic variables in Table 2. The outputs for each method are the predicted soil moisture (θ_{pred}) for the five depths: 0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm. The models are trained and evaluated using the in situ soil moisture networks.

The in situ dataset was divided into 70% for training/validation and 30% for testing. Of the 70% used for training/validation, a further split was made into training (~80%) and validation (~20%) datasets. Additionally, a 5-fold cross-validation technique was employed to ensure robust model evaluation and prevent overfitting. The training dataset is used to train the machine learning models, while the validation dataset is used to fine-tune the machine learning algorithm's hyperparameters (i.e., parameters that control the development of the machine learning models). The testing dataset is not used for model development, so it is used to assess the performance of the machine learning algorithms when they are applied to unobserved conditions. The divisions were determined based on stratified splitting [143], which ensures that all in situ networks are represented in each division and that the models are trained on representative samples. Divisions were based only on location (not time), so the entire record of a given in situ soil moisture station occurs in a single division of the dataset. Thus, the testing dataset evaluates the ability of the machine learning algorithms to estimate the soil moisture at unobserved locations.

RMSE was used as the evaluation metric (loss function) for all the machine learning methods. Random search was used for hyperparameter optimization, and the range of hyperparameters was defined based on each method's documentation and the literature reviews of similar applications. Table 3 summarizes the optimized hyperparameter values for each machine learning algorithm, and Appendix A describes the roles of the main hyperparameters.

Table 3. Optimal hyperparameters for five machine learning models used in this study.

Model	Hyperparameter	Optimal Value	Default
ANN	Number of hidden layers	3	1
	Hidden layer sizes	100	100
	Activation function	Relu	Relu
	Training algorithm	Adam	Adam
	Regularization term	0.01	0.0001
	Learning rate	0.001	constant
	Maximum iterations	100	200
RF	Number of trees	200	100
	Maximum depth	10	None
	Min. samples for split	5	2
	Min. samples for leaf	2	1
	Max. features at split	sqrt	1
	Split criterion	Squared error	Squared error

Table 3. *Cont.*

Model	Hyperparameter	Optimal Value	Default
XGBoost	Learning rate	0.3	0.3
	Maximum depth	6	6
	Number of trees	500	100
	Subsample for tree	1	1
	Depth sample fraction	1	1
	Min. child weight	0.8	1
CatBoost	Number of trees	1000	1000
	Learning rate	0.05	0.03
	Depth of tree	10	6
	Subsample for iteration	1	1
	Level feature proportion	1	1
	Regularization	3	3
LightGBM	Number of boosting iterations	1000	100
	Learning rate	0.05	0.01
	Number of leaves	31	31
	Maximum depth	10	−1 (unlimited)
	Min. data in leaf	20	20
	Regularization	0.1	0.0

The Pearson correlation coefficient (R), mean bias error (MBE), RMSE, ubRMSE, Nash–Sutcliffe Efficiency (NSE), and KGE were used to evaluate the accuracy of the soil moisture predictions (θ_{pred}) in reproducing the in situ observations (θ_{obs}). These metrics are calculated as follows:

$$R = \frac{\sum_{i=1}^{N}\left(\theta_{obs,i} - \overline{\theta_{obs}}\right)\left(\theta_{pred,i} - \overline{\theta_{pred}}\right)}{\sqrt{\sum_{i=1}^{N}\left(\theta_{obs,i} - \overline{\theta_{obs}}\right)^2 \sum_{i=1}^{N}\left(\theta_{pred,i} - \overline{\theta_{pred}}\right)^2}} \quad (1)$$

$$\text{MBE} = \frac{1}{N}\sum_{i=1}^{N}\left(\theta_{pred,i} - \theta_{obs,i}\right) \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_{pred,i} - \theta_{obs,i}\right)^2} \quad (3)$$

$$\text{ubRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_{pred,i} - \theta_{obs,i}\right)^2} - \text{MBE} \quad (4)$$

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{N}\left(\theta_{obs,i} - \theta_{pred,i}\right)^2}{\sum_{i=1}^{N}\left(\theta_{obs,i} - \overline{\theta_{obs}}\right)^2}\right] \quad (5)$$

$$\text{KGE} = 1 - \sqrt{(R-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \quad (6)$$

The metric R describes the linear correlation between the predicted (i.e., model) and observed values where $\overline{\theta_{pred}}$ and $\overline{\theta_{obs}}$ are the predicted and observed soil moisture means [144]. MBE describes whether a model typically overestimates or underestimates the observed values (positive values indicate overestimations) [145]. ubRMSE considers the error that remains if the bias is removed from the model estimates [146]. NSE compares the squared error to the variance of the observations [147,148]. NSE ranges from $-\infty$ to 1 with a value of 1 indicating a perfect agreement between the model and observations and a value of 0 occurring if the mean of the observations is used as the model. KGE combines three measures of model performance including R, $\beta = \mu_{pred}/\mu_{obs}$, and $\alpha = \sigma_{pred}/\sigma_{obs}$, where μ_{pred} and μ_{obs} are the predicted and observed means and σ_{pred} and σ_{obs} are the predicted and observed standard deviations [149]. KGE ranges from $-\infty$ to 1, with values closer to 1 indicating more accurate model estimates.

3. Results

3.1. Performance of Machine Learning Algorithms

Figure 2 shows the performance metrics for the machine learning methods' soil moisture predictions of the testing dataset across all networks and depth ranges (combined). The results indicate that the RF, XGBoost, and CatBoost models exhibit better testing performance than SMAP as well as the ANN and LightGBM models. In particular, the RF, XGBoost, and CatBoost models typically have lower RMSE and ubRMSE values and higher R, NSE, and KGE values. These same three models also tend to have smaller biases than the SMAP, ANN, and LightGBM models. Overall, the XGBoost method exhibits the best testing accuracy among the methods. In contrast, SMAP shows the lowest accuracy among the methods, primarily due to higher bias. However, this evaluation uses SMAP as a direct estimate of soil moisture at each station. SMAP is expected to have better performance as an estimate of the spatial average soil moisture across its 9 km grid cells.

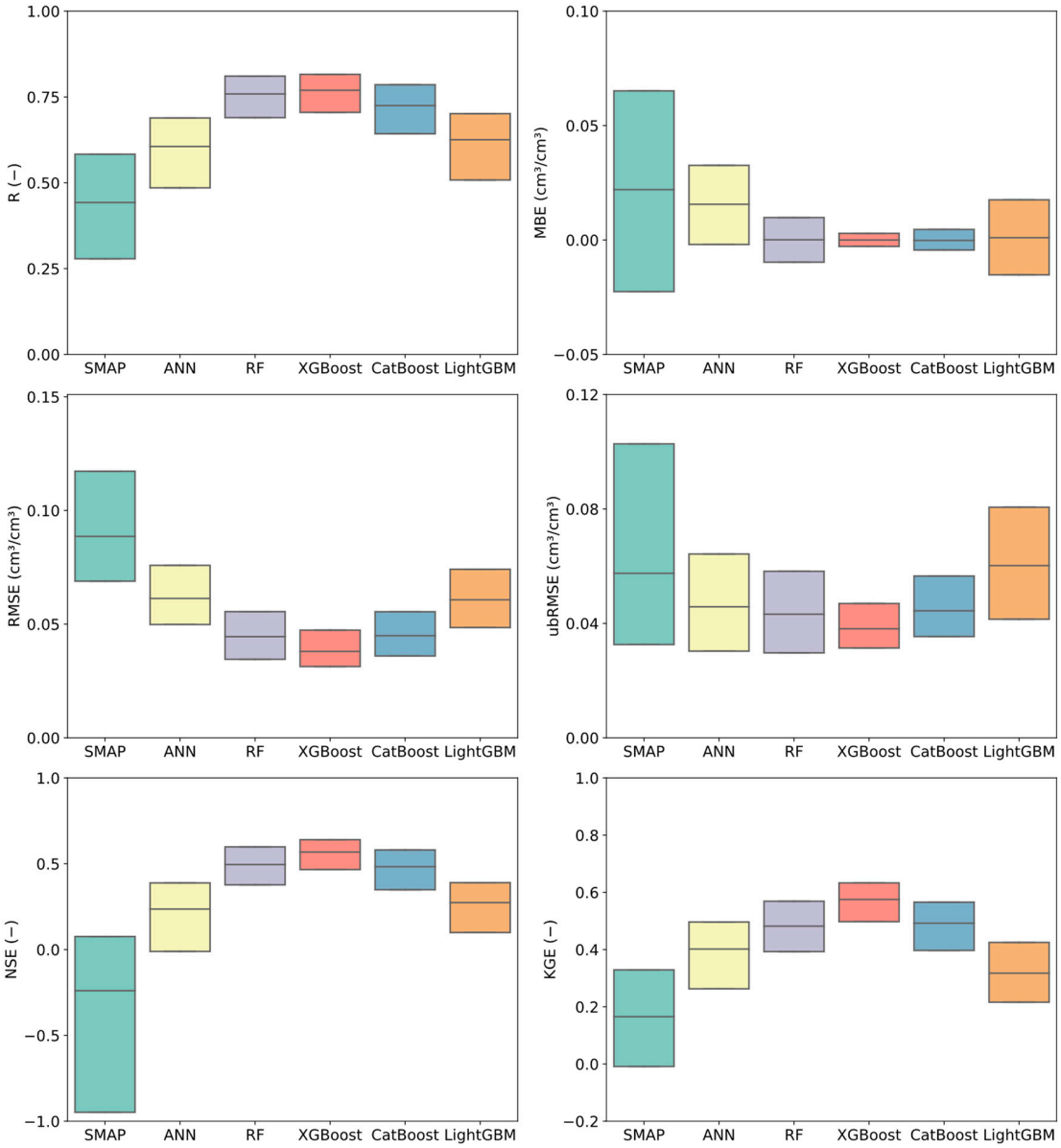

Figure 2. Performance metrics (*R*, MBE, RMSE, ubRMSE, NSE, and KGE) for the soil moisture estimates of the machine learning algorithms when compared to the testing data, including all depths and stations. For each performance metric, the line inside the box indicates the median value and the box represents the interquartile range.

Figure 3a examines the RMSE of the machine learning methods when the testing data are divided according to the in situ soil moisture networks (ARM, COSMOS, AMERIFLUX, SCAN, SNOTEL, TxSON, USCRN, and iRON). The machine learning methods exhibit similar performance across most of the networks. However, the RMSE values are higher for the COSMOS network than the other networks. The poorer performance at the COSMOS stations likely occurs because the spatial support for the cosmic ray neutron measurements (~700 m diameter) [86] is much larger than the support for the in situ probes used in the other networks (centimeters at most). The machine learning methods are trained on all networks simultaneously, so the COSMOS data are inconsistent with the other datasets. No predictor variable allows the machine learning methods to identify whether an in situ measurement is from the COSMOS dataset or the other networks. Furthermore, all predictor variables are represented at a 70 m resolution, so the machine learning methods lack information to characterize much of the spatial support for the COSMOS data.

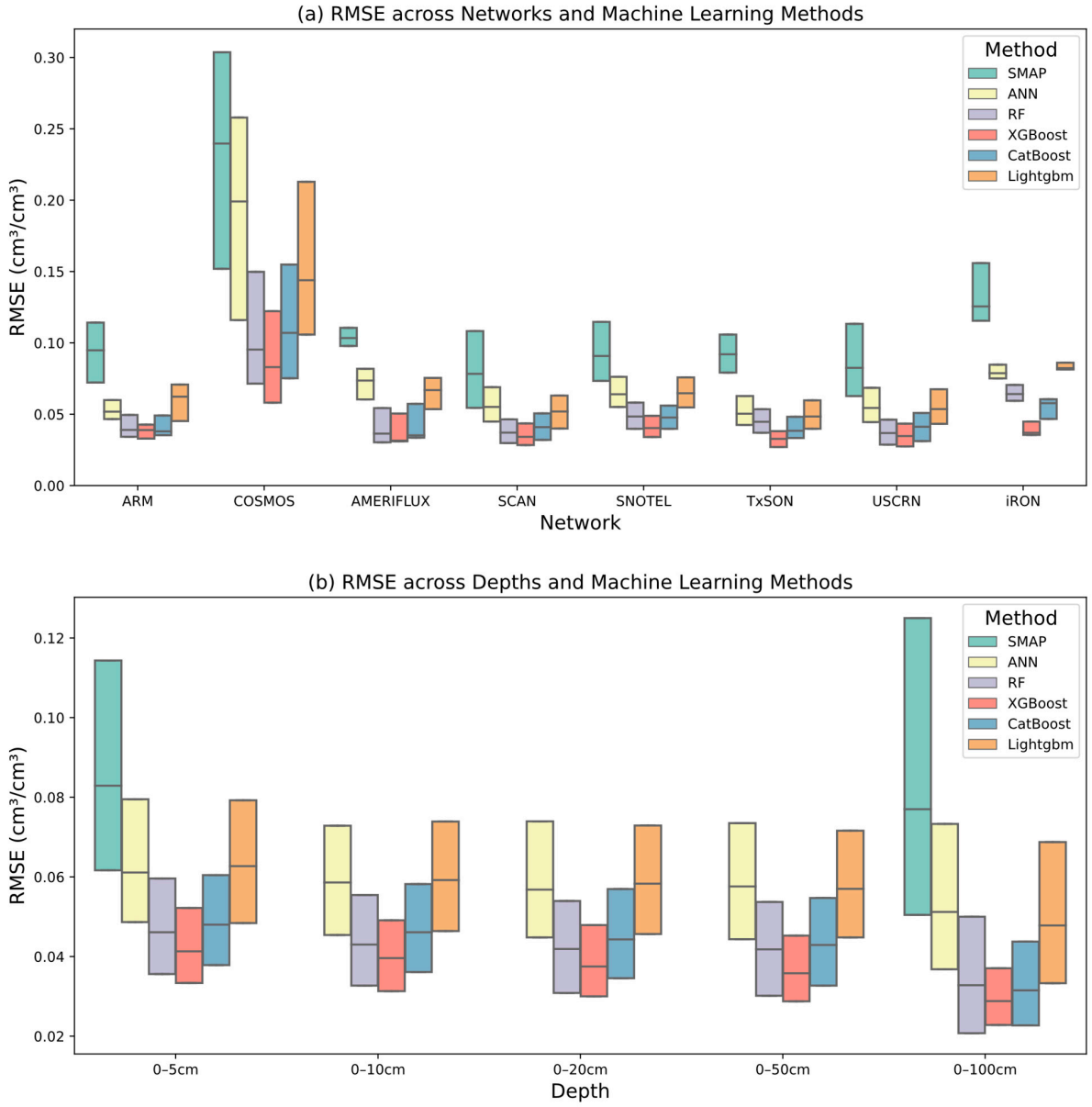

Figure 3. RMSE of the soil moisture estimates from the machine learning algorithms for the testing dataset when the data are divided according to the (**a**) in situ soil moisture networks and (**b**) depths. For each performance metric, the line inside the box indicates the median and the box represents the interquartile range.

Figure 3b compares the performance of the machine learning methods when the testing dataset is divided according to depth. XGBoost consistently has the lowest RMSE values across all depths, followed by RF, CatBoost, LightGBM, ANN, and SMAP. As the depth increases, the accuracy of the methods usually improves. The largest improvement in performance occurs between a 0–50 cm and 0–100 cm depth. The improved performance for deeper layers may occur due to the greater uniformity of soil moisture at greater depths, which facilitates model learning and prediction. Maps of both the surface and rootzone soil moisture using the machine learning methods are provided in the Supplementary Materials (Figures S1 and S2).

Figure 4 compares the XGBoost soil moisture estimates to the in situ soil moisture observations when the testing dataset is divided by climate classification. The climate classification for each location was determined using the UNEP [150] system, which is based on the AI. An arid region has an AI below 0.20, a semiarid region has an AI from 0.20 to 0.50, a sub-humid region has an AI from 0.50 to 0.65, and a humid region has an AI above 0.65. In the dataset, 59 stations are arid, 401 stations are semiarid, 86 stations are sub-humid, and 185 stations are humid. The soil moisture estimates from XGBoost are relatively accurate across all climatic regions with R exceeding 0.8 and RMSE below 0.045 cm^3/cm^3 for all four climates. The weakest performance occurs in the semiarid region, which is the only region where the NSE and KGE values are below 0.70. Lower RMSE values might occur for the semiarid climate because that climate's dataset is more diverse than the others. A wider variety of topography, soil types, vegetation types, and other factors that influence soil moisture occurs within this climatic region. Consequently, the machine learning model might have more difficulty capturing the underlying patterns of soil moisture.

Figures 5 and 6 examine whether XGBoost captures the temporal dynamics of soil moisture for an arid location (USCRN LasCruces20N) and a humid location (USCRN Versailles3NNW), respectively. Both stations are members of the testing dataset and typical for their climatic region. In each figure, the upper part considers the surface soil moisture (0–5 cm), and the lower part considers the rootzone soil moisture (0–100 cm). For the arid location (Figure 5), XGBoost closely follows the in situ soil moisture variations at both depths including responses to individual precipitation events. XGBoost has a small wet bias, but the magnitude of the bias is smaller than the dry bias seen when using the SMAP L4 product at this station. XGBoost provides a more accurate representation of soil moisture dynamics at the arid location than directly using SMAP. XGBoost has correlations of 0.79 for surface and 0.75 for rootzone while SMAP has correlations of 0.64 for surface and 0.70 for rootzone. XGBoost has RMSE values of 0.016 cm^3/cm^3 for surface and 0.017 cm^3/cm^3 for rootzone while SMAP has RMSE values of 0.043 cm^3/cm^3 for surface and 0.054 cm^3/cm^3 for rootzone.

At the humid location (Figure 6), XGBoost also tracks the temporal variations of in situ soil moisture, maintaining high moisture values except in prolonged periods of low precipitation. Again, XGBoost provides a better representation of the time series than directly using the SMAP estimates. XGBoost has small wet biases at this station (MBE of 0.007 cm^3/cm^3 for the surface and 0.013 cm^3/cm^3 for the rootzone), whereas SMAP exhibits substantial wet biases (MBE of 0.107 cm^3/cm^3 for the surface and 0.124 cm^3/cm^3 for the rootzone). XGBoost has correlations of 0.80 for the surface and 0.76 for the rootzone while SMAP has correlations of 0.62 for the surface and 0.69 for the rootzone. XGBoost has RMSE values of 0.043 cm^3/cm^3 for the surface and 0.036 cm^3/cm^3 for the rootzone while SMAP has an RMSE value of 0.124 cm^3/cm^3 for both the surface and the rootzone.

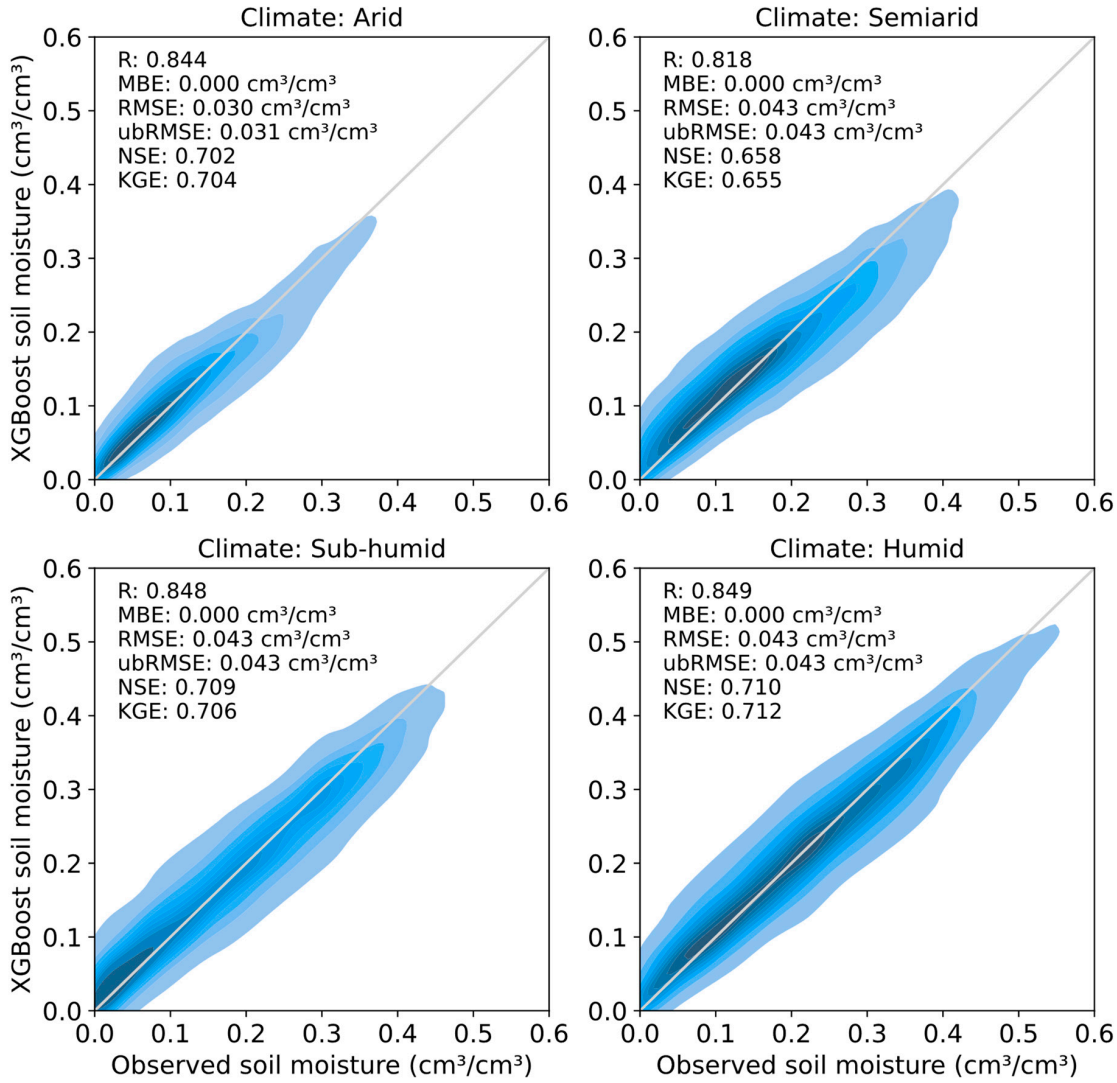

Figure 4. Density plots comparing the observed and XGBoost estimates of soil moisture for each depth using the testing datasets for each climate. Darker blues represent higher concentrations of data, while lighter blues represent lower concentrations.

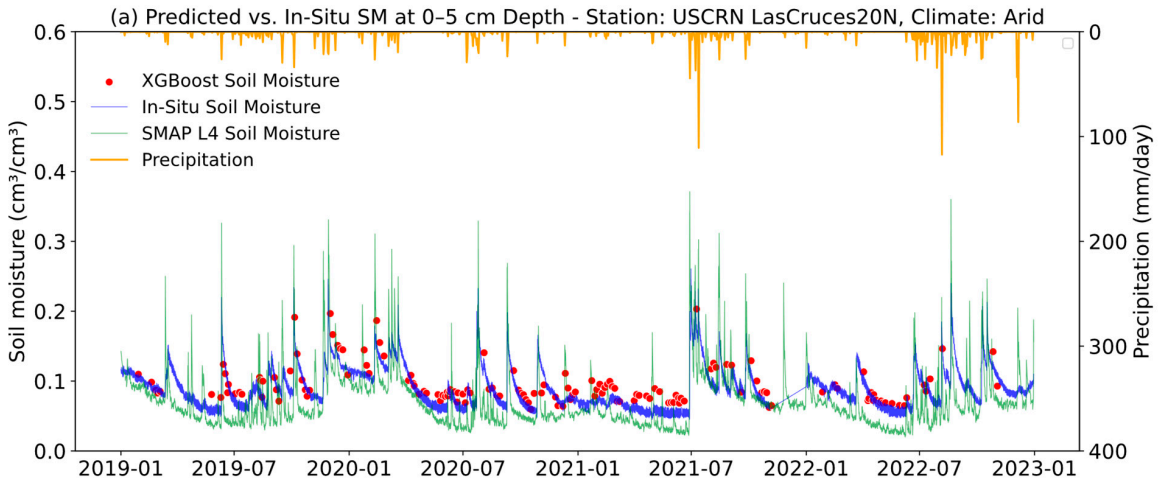

Figure 5. *Cont.*

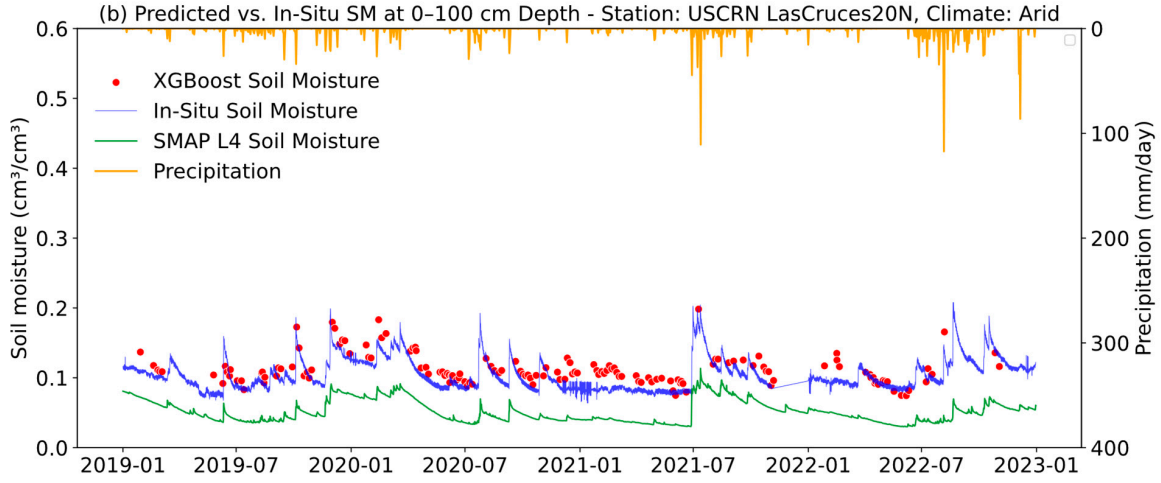

Figure 5. Time series of soil moisture at (**a**) 0–5 cm and (**b**) 0–100 cm depths at the arid USCRN Las Cruces 20N station (a member of the testing dataset). The plotted soil moisture data include hourly in situ measurements, estimates from the XGBoost model, and 3 h SMAP L4 soil moisture estimates. Daily GPM precipitation data at the site are also shown.

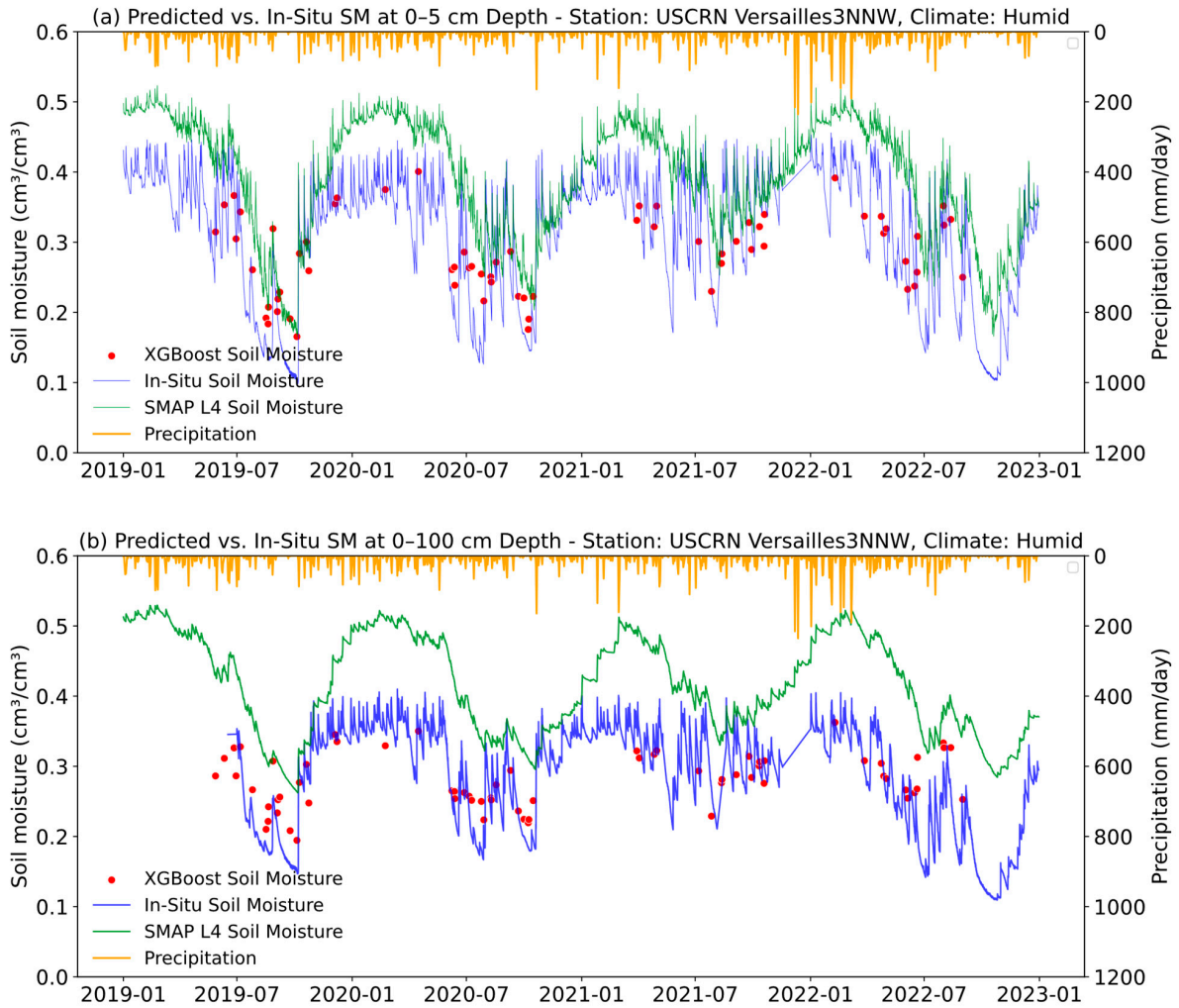

Figure 6. Time series of soil moisture at (**a**) 0–5 cm and (**b**) 0–100 cm depths at the humid USCRN Versailles 3NNW station (a member of the testing dataset). The plotted soil moisture data include hourly in situ measurements, estimates from the XGBoost model, and 3 h SMAP L4 soil moisture estimates. Daily GPM precipitation data for the site are also shown.

3.2. Importance of Predictor Variables

Figure 7 presents the correlations between all the predictor variables and the in situ soil moisture at different depths for the complete dataset. The correlations indicate that the in situ soil moisture at a given depth is most related to the soil moisture in nearby layers. It also shows that the in situ soil moisture values for 0–100 cm are most different from the in situ soil moisture at the other depths.

Soil moisture from SMAP is highly correlated with in situ soil moisture values (Figure 7). Unexpectedly, the SMAP surface soil moisture exhibits higher correlations than the SMAP rootzone and profile soil moisture products with the deep in situ soil moisture. This behavior likely occurs because the surface soil moisture is inferred more directly from the satellite sensors, while the rootzone and profile soil moisture products rely in part on models. Both the SMAP soil moisture values from the same date and the antecedent SMAP soil moisture values exhibit similar correlations to the in situ soil moisture.

Vegetation indices such as the NDVI and EVI exhibit moderate correlations to the in situ soil moisture, with the EVI displaying the highest correlations. The EVI is less sensitive to variations in canopy structure and background factors such as soil conditions and atmospheric influences, and it is more sensitive to changes in canopy chlorophyll content, which makes it a better indicator of plant health and moisture stress than the NDVI [151–153]. Among the vegetation indices, the LAI usually exhibits the weakest correlations with soil moisture, perhaps because it is less indicative of the degree of soil surface shading than the NDVI or EVI.

In situ soil moisture exhibits positive correlations with clay and silt content, and negative correlations with sand content. Higher sand content (and lower clay and silt content) is expected to increase drainage and reduce soil moisture. Wang et al. [154] found that the spatial pattern of soil moisture is greatly influenced by soil factors, such as the sand and clay fractions. Bulk density, organic matter, and EC exhibit low correlations with soil moisture. Higher bulk density and lower organic matter are expected to reduce porosity, which would reduce the allowable range for soil moisture. However, errors in the estimated bulk density and organic matter values may cause the low correlations seen in Figure 7. pH exhibits moderate negative correlations with in situ soil moisture. Soil pH plays a role in determining microbial activity, nutrient availability, and soil structure, which can indirectly influence the soil moisture [155].

Positive correlations occur between in situ soil moisture and precipitation. The correlations are strongest for 14-day antecedent precipitation and weaken as the time period shortens to 6 h. This result highlights soil moisture's memory (i.e., its ability to retain information) about past precipitation events [156–158]. The ESI and ET are positively correlated with soil moisture, meaning that higher soil moisture allows higher ET rates. The LST exhibits a weak negative correlation with soil moisture. Lower soil moisture reduces the latent heat flux and warms the land surface. Moreover, the AI is positively correlated with soil moisture, reflecting the greater availability of water in more humid climates.

The topographic variables exhibit a range of correlations with in situ soil moisture. Elevation displays a strong negative correlation with soil moisture. Rong et al. [159] reported that soil moisture at low elevation is often supplemented by surface runoff and subsurface flow from higher elevation points. This leads to a negative correlation between soil moisture and elevation. However, within smaller spatial extents (Reynolds Creek watershed in southern Idaho), Cowley et al. [68] found that soil moisture has a positive correlation with elevation due to increased precipitation at higher elevations. Slope exhibits a negative correlation, likely because it promotes lateral outflow of moisture [60,160]. Aspect, hillside, and mTPI exhibit only weak correlations in this dataset.

Categories	Variable	Depth (cm)	Soil moisture				
			0–5	0–10	0–20	0–50	0–100
In-situ Soil moisture	In-Situ Soil Moisture	0–5	1.000	0.915	0.822	0.852	0.432
		0–10	0.915	1.000	0.908	0.858	0.460
		0–20	0.822	0.908	1.000	0.853	0.451
		0–50	0.852	0.858	0.853	1.000	0.537
		0–100	0.432	0.460	0.451	0.537	1.000
Antecedent Soil Moisture	SMAP surface soil moisture (0–5 cm) (SSM)	0–5	0.492	0.527	0.531	0.476	0.390
	SMAP Root zone soil moisture (0–100 cm) (RZSM)	0–100	0.400	0.438	0.455	0.406	0.357
	SMAP profile soil moisture (0–bedrock depth) (PSI	--	0.379	0.418	0.432	0.392	0.340
	Antecedent 1-Day SMAP SSM	0–5	0.498	0.534	0.536	0.482	0.394
	Antecedent 1-Day SMAP RZSM	0–100	0.400	0.439	0.455	0.407	0.358
	Antecedent 1-Day SMAP PSM	--	0.379	0.418	0.432	0.392	0.341
	Antecedent 3-Day SMAP SSM	0–5	0.501	0.538	0.542	0.488	0.398
	Antecedent 3-Day SMAP RZSM	0–100	0.399	0.437	0.454	0.406	0.358
	Antecedent 3-Day SMAP PSM	--	0.377	0.416	0.431	0.391	0.341
	Antecedent 7-Day SMAP SSM	0–5	0.498	0.536	0.543	0.490	0.400
	Antecedent 7-Day SMAP RZSM	0–100	0.394	0.432	0.450	0.404	0.356
	Antecedent 7-Day SMAP PSM	--	0.373	0.412	0.427	0.388	0.339
	Antecedent 14-Day SMAP SSM	0–5	0.490	0.528	0.538	0.488	0.398
	Antecedent 14-Day SMAP RZSM	0–100	0.387	0.425	0.444	0.398	0.352
	Antecedent 14-Day SMAP PSM	--	0.366	0.405	0.421	0.383	0.336
Vegetation	Normalized Difference Vegetation Index (NDVI)	--	0.149	0.151	0.143	0.136	0.106
	Enhanced Vegetation Index (EVI)	--	0.175	0.193	0.191	0.191	0.179
	Leaf Area Index (LAI)	--	0.091	0.085	0.073	0.090	0.086
	fraction of Photosynthetically Active Radiation	--	0.095	0.094	0.089	0.086	0.089
Soil Properties and Landcover	Landcover (NLCD)	--	0.149	0.195	0.184	0.213	0.297
	Silt	0–5	0.184	0.226	0.217	0.254	0.273
		0–10	0.186	0.230	0.218	0.255	0.272
		0–20	0.191	0.237	0.222	0.256	0.270
		0–50	0.175	0.221	0.197	0.239	0.260
		0–100	0.172	0.214	0.186	0.234	0.245
	Sand	0–5	-0.213	-0.267	-0.258	-0.262	-0.219
		0–10	-0.227	-0.285	-0.277	-0.278	-0.242
		0–20	-0.229	-0.292	-0.288	-0.291	-0.261
		0–50	-0.232	-0.297	-0.306	-0.300	-0.286
		0–100	-0.211	-0.270	-0.288	-0.275	-0.273
	Clay	0–5	0.187	0.229	0.258	0.222	0.224
		0–10	0.188	0.231	0.259	0.223	0.218
		0–20	0.190	0.236	0.266	0.229	0.207
		0–50	0.229	0.275	0.287	0.281	0.242
		0–100	0.246	0.290	0.292	0.288	0.245
	Bulk Density	0–5	-0.019	-0.037	-0.054	-0.050	-0.116
		0–10	-0.013	-0.035	-0.050	-0.040	-0.106
		0–20	-0.011	-0.033	-0.050	-0.038	-0.100
		0–50	-0.017	-0.038	-0.059	-0.046	-0.112
		0–100	-0.022	-0.038	-0.060	-0.051	-0.110
	Organic Matter	0–5	0.039	0.047	0.067	0.048	0.108
		0–10	0.043	0.049	0.065	0.040	0.091
		0–20	0.053	0.055	0.066	0.040	0.078
		0–50	0.048	0.046	0.026	0.042	0.068
		0–100	0.069	0.063	0.024	0.060	0.057
	Electrical Conductivity	0–5	-0.052	-0.050	-0.048	-0.044	-0.038
		0–10	-0.053	-0.051	-0.050	-0.041	-0.031
		0–20	-0.073	-0.066	-0.064	-0.044	-0.019
		0–50	-0.080	-0.067	-0.063	-0.035	0.002
		0–100	-0.079	-0.064	-0.060	-0.029	0.012
	pH	0–5	-0.136	-0.126	-0.099	-0.118	-0.100
		0–10	-0.142	-0.134	-0.102	-0.124	-0.114
		0–20	-0.144	-0.137	-0.106	-0.129	-0.127
		0–50	-0.137	-0.129	-0.098	-0.119	-0.119
		0–100	-0.123	-0.115	-0.086	-0.103	-0.109
	Depth to Restrictive Layer	--	0.022	0.021	-0.010	0.017	0.015
Precipitation and Climate	Antecedent 6-hour Precipitation	--	0.037	0.036	0.033	0.033	0.031
	Antecedent 1-Day Precipitation	--	0.104	0.111	0.104	0.096	0.088
	Antecedent 3-Day Precipitation	--	0.227	0.239	0.225	0.212	0.198
	Antecedent 7-Day Precipitation	--	0.311	0.331	0.315	0.297	0.278
	Antecedent 14-Day Precipitation	--	0.365	0.393	0.380	0.361	0.336
	Land Surface Temperature	--	-0.109	-0.108	-0.083	-0.063	0.041
	Evaporative Stress Index (ESI)	--	0.248	0.267	0.286	0.232	0.180
	Potential Evapotranspiration	--	-0.044	-0.048	-0.041	-0.029	0.007
	Evapotranspiration (ET)	--	0.148	0.160	0.177	0.148	0.132
	Aridity Index	--	0.166	0.176	0.173	0.185	0.185
Topography	Elevation	--	-0.226	-0.259	-0.296	-0.278	-0.365
	Slope	--	-0.169	-0.189	-0.207	-0.214	-0.181
	Aspect	--	-0.025	-0.023	-0.031	-0.046	-0.075
	Hillshade	--	0.005	0.011	0.014	0.015	-0.001
	Multi-Scale Topographic Position Index (mTPI)	--	0.039	0.042	0.055	0.055	0.055

Figure 7. Correlations between predictor variables and in situ soil moisture at different depths. Positive correlations are shown in blue and negative correlations are shown in red.

Figure 8 presents the predictor variable importance scores for RF, XGBoost, CatBoost, and LightGBM. These scores differ substantially between the methods, which likely occurs because some of the input variables contain similar information. Thus, different variables can be selected depending on the methods' learning architectures. Among topographic variables, elevation is most important to the model predictions. However, the remaining topographic factors all typically have substantial importance (all are in the top half of the table in terms of average importance). Elevation above sea level does not directly influence soil moisture but serves as a proxy for other environmental factors. As the elevation increases, temperatures decrease, reducing ET and potentially increasing soil moisture. Precipitation patterns also shift, with higher elevations typically receiving more precipitation and more snow. Among soil texture variables, percent clay is most important for three of the four models (LightGBM relies more heavily on silt). The reliance on clay is interesting because percent sand is more correlated with soil moisture (Figure 7). The reliance on clay suggests either that clay has a nonlinear relationship with soil moisture or that its role is more independent from other predictor variables than that of sand. Soil depth, organic matter, and bulk density are also somewhat important. Vegetation plays a smaller role than topography and soil. Among vegetation variables, the EVI is most important, followed by the NDVI and LAI. This result aligns with Figure 7, where the EVI shows the strongest relationships with soil moisture among the vegetation variables. Land cover classification provides little value in predicting soil moisture (Figure 8). The AI is the most important climate/weather variable for the machine learning methods. Among temporally varying weather variables, 14-day antecedent precipitation is the most important. Although recent rainfall directly impacts soil moisture, a 14-day window provides a more complete indication of moisture additions. The other meteorological variables (shorter antecedent precipitation as well as PET, the ESI, and ET) have only moderate importance. Considering the SMAP products, the most important predictor variables are all related to surface soil moisture. All rootzone and profile soil moisture variables have low importance. The most important SMAP variables are usually the 3-day and 14-day antecedent soil moisture. However, the current SMAP surface soil moisture is used heavily by the RF model.

Predictor Variable	Importance				
	RF	XGBoost	CatBoost	LightGBM	Average
Elevation	0.94	0.58	1.00	0.79	0.83
Aridity Index (AI)	0.67	0.33	0.84	0.72	0.64
Clay	0.66	0.28	0.79	0.81	0.64
Depth	0.47	0.25	0.90	0.86	0.62
Antecedent 3-Day SMAP SSM	0.88	1.00	0.13	0.47	0.62
Antecedent 14-Day SMAP SSM	0.75	0.36	0.59	0.72	0.60
Sand	0.74	0.36	0.51	0.80	0.60
Silt	0.65	0.27	0.52	0.93	0.59
Multi-Scale Topographic Position Index (mTPI)	0.54	0.26	0.78	0.68	0.56
Organic Matter	0.51	0.18	0.61	0.95	0.56
Aspect	0.49	0.23	0.68	0.84	0.56
Antecedent 7-Day SMAP SSM	1.00	0.22	0.38	0.49	0.52
Hillshade	0.46	0.17	0.62	0.60	0.46
Bulk Density	0.50	0.18	0.41	0.76	0.46
Antecedent 14-Day Precipitation	0.42	0.10	0.29	1.00	0.45
Enhanced Vegetation Index (EVI)	0.29	0.08	0.27	0.95	0.40
fraction of Photosynthetically Active Radiation (fPAR)	0.31	0.08	0.25	0.92	0.39
Antecedent 14-Day SMAP SSM	0.78	0.14	0.22	0.42	0.39
Normalized Difference Vegetation Index (NDVI)	0.31	0.11	0.18	0.93	0.38
pH	0.43	0.15	0.34	0.59	0.38
Slope	0.37	0.12	0.61	0.38	0.37
Evaporative Stress Index (ESI)	0.26	0.07	0.23	0.81	0.34
Antecedent 7-Day Precipitation	0.26	0.06	0.27	0.76	0.34
Leaf Area Index (LAI)	0.26	0.07	0.22	0.76	0.33
SMAP surface soil moisture (0-5 cm)	0.68	0.05	0.08	0.48	0.32
Land Surface Temperature (LST)	0.19	0.06	0.20	0.71	0.29
Depth to Restrictive Layer	0.24	0.08	0.42	0.29	0.26
Antecedent 6-hour Precipitation	0.03	0.01	0.03	0.85	0.23
Antecedent 14-Day SMAP RZSM	0.31	0.05	0.09	0.46	0.23
Antecedent 3-Day Precipitation	0.15	0.04	0.11	0.61	0.23
Antecedent 14-Day SMAP PSM	0.29	0.04	0.11	0.46	0.22
SMAP Root zone soil moisture (0-100 cm)	0.38	0.05	0.05	0.36	0.21
Evapotranspiration (ET)	0.16	0.03	0.06	0.55	0.20
Potential Evapotranspiration (PET)	0.13	0.02	0.09	0.52	0.19
Antecedent 1-Day Precipitation	0.07	0.01	0.04	0.62	0.19
SMAP profile soil moisture (0-bedrock depth)	0.34	0.03	0.05	0.32	0.19
Antecedent 7-Day SMAP RZSM	0.38	0.02	0.09	0.25	0.18
Antecedent 1-Day SMAP RZSM	0.49	0.01	0.06	0.15	0.18
Antecedent 3-Day SMAP RZSM	0.44	0.03	0.04	0.19	0.17
Antecedent 7-Day SMAP PSM	0.26	0.02	0.05	0.23	0.14
Antecedent 1-Day SMAP PSM	0.30	0.01	0.06	0.16	0.13
Antecedent 3-Day SMAP PSM	0.28	0.01	0.03	0.19	0.13
NLCD Shrub, scrub	0.11	0.02	0.32	0.02	0.12
NLCD Cultivated crops	0.09	0.07	0.16	0.03	0.09
NLCD Evergreen forest	0.06	0.02	0.14	0.02	0.06
NLCD Grassland, herbaceous	0.04	0.02	0.06	0.04	0.04
NLCD Deciduous forest	0.03	0.01	0.04	0.03	0.03
NLCD Developed, open space	0.03	0.01	0.02	0.02	0.02
NLCD Pasture, hay	0.02	0.00	0.01	0.02	0.01
NLCD Mixed forest	0.02	0.01	0.00	0.02	0.01
NLCD Emergent herbaceous wetlands	0.01	0.01	0.01	0.01	0.01
NLCD Woody wetlands	0.03	0.00	0.00	0.01	0.01
NLCD Developed, low intensity	0.01	0.00	0.00	0.00	0.00
NLCD Developed, medium intensity	0.00	0.00	0.00	0.00	0.00
NLCD Barren land (rock, sand, clay)	0.00	0.00	0.00	0.00	0.00

Figure 8. Relative importance of each predictor variable in the RF, XGBoost, CatBoost, and LightGBM models and the average importance among the four models.

4. Discussion

Overall, this study suggests that machine learning algorithms can provide accurate estimates of rootzone soil moisture at unobserved locations within CONUS. It also suggests

that XGBoost provides the most accurate estimates among the algorithms tested (XGBoost produced an overall RMSE of 0.042 cm^3/cm^3). The results are generally consistent with the findings of similar studies conducted in other regions. For example, Kornelsen et al. [133] obtained an RMSE for soil moisture of around 0.07 cm^3/cm^3 using an ANN. Senyurek et al. [161] reported an RMSE of approximately 0.052 cm^3/cm^3 for surface soil moisture using an RF. Liu et al. [133] reported an RMSE of 0.048 cm^3/cm^3 using a GBM. However, the present study considered a larger region than prior studies, such as those by Singh et al. [80] and Abowarda et al. [70]. The larger spatial extent likely includes more heterogeneity, which may impact the algorithms' performance. In the present study, the performance was weakest in the semiarid region, which is consistent with Ren et al. [162] and Jamie et al. [135], who reported challenges in semiarid regions due to diverse topographic, soil, and vegetation characteristics. Future studies could consider more advanced machine learning techniques, such as convolutional neural networks and recurrent neural networks. Ensemble models could also be developed to combine the predictions from multiple machine learning algorithms.

The performance of machine learning models usually improved with depth, which supports their suitability for estimating rootzone soil moisture. Smaller errors likely occurred at greater depths because the soil moisture is steadier with less dependence on individual precipitation events or recent evaporation.

The present study considered more predictor variables (including vegetation indices, soil characteristics, weather and climate variables, and topographic features) than prior studies. For example, Du et al. [163] and Park et al. [164] primarily used vegetation indices and climate data. Using more predictor variables likely improves performance, but it also increases the effort required for data preparation and the time needed to train and apply the machine learning algorithms. Future studies could consider more refined feature selection. Landcover classifications and ECOSTRESS products typically have low importance in the trained models, so excluding these variables may simplify the models while having little effect on the results. However, this study only considered soil moisture at locations with long-term monitoring where landcover has been stable through time. If predictions are made at other locations where landcover has changed, landcover (and landcover changes) may play a more important role. The limited temporal data coverage of ECOSTRESS data (due to clouds) is a significant limitation. It allowed for only a small fraction (1%) of the available hourly soil moisture data to be utilized for the study. Other remote sensing data with similar spectral bands, such as Sentinel-1, Sentinel-2, and Landsat, could potentially be fused to enhance temporal coverage. The selection of topographic indices in this study was based on their availability in GEE and their relevance in previous soil moisture studies. Including other relevant indices such as the drainage area and topographic wetness index could enhance the soil moisture predictions, especially in regions with complex terrain.

The predictor variables used in this study were represented at a 70 m resolution, which produces soil moisture estimates with the same nominal spatial resolution. This resolution is finer than most prior studies. Jamei et al. [135] considered a 9 km resolution for rootzone soil moisture, Senyurek et al. [161] considered a 3 km resolution for surface soil moisture, and Huang et al. [165] and Yang et al. [166] considered a 1 km resolution covering various depths. Sun et al. [136] and Singh and Gaurav [80] considered 500 m and 60 m resolutions, respectively. Greifeneder et al. [77] and Abowarda et al. [70] considered 50 m and 30 m for surface soil moisture, respectively, and Nguyen et al. [167] and Meyer et al. [78] considered 10 m for surface soil moisture. In the present study, the models were trained by comparing to point measurements of soil moisture (aside from the COMOS data), which implicitly assumes that the point measurements are representative of the average soil moisture over the 70 m grid cell. Because the in situ soil moisture data are widely spaced, the soil moisture estimates from the machine learning methods are not expected to fully capture fine-scale spatial variations in soil moisture. Future studies could consider the accuracy of these estimates when compared to more closely spaced in situ soil moisture observations from sub-regions.

5. Conclusions

This study evaluated the accuracy of five machine learning algorithms (feed-forward ANN, RF, XGBoost, CatBoost, and LightGBM) for estimating soil moisture at multiple depths (0–5 cm, 0–10 cm, 0–20 cm, 0–50 cm, and 0–100 cm) at unobserved locations across CONUS. The machine learning methods operated at a 70 m spatial resolution and were trained and tested by comparing to several in situ soil moisture networks.

1. For the dataset considered, XGBoost provides more accurate soil moisture estimates than the other machine learning models considered. XGBoost also provides better accuracy than directly using SMAP as an estimate of point soil moisture. XGBoost achieves the lowest mean RMSE of 0.042 cm^3/cm^3 compared to random forest (0.048 cm^3/cm^3), CatBoost (0.050 cm^3/cm^3), ANN (0.067 cm^3/cm^3), LightGBM (0.066 cm^3/cm^3), and SMAP (0.101 cm^3/cm^3) for the testing locations. XGBoost produces the best accuracy when considering the entire testing dataset, and it produces the best accuracy when separately considering each in situ soil moisture network and each depth.
2. All the machine learning algorithms perform more poorly when comparing to data from the COSMOS network than to the other in situ data networks. The COSMOS measurements have a larger footprint (~700 m diameter) than the point measurements in the other networks. This inconsistency as well as the inconsistency between the COSMOS footprint and the 70 m resolution used to represent the predictor variables likely contributes to the poorer performance at the COSMOS sites.
3. The machine learning algorithms typically provide more accurate estimates as the depth of the estimate increases. For example, XGBoost produces a median RMSE of around 0.041 cm^3/cm^3 for a 0–5 cm depth and 0.029 cm^3/cm^3 for a 0–100 cm depth. The accuracy may improve with depth because deeper soil moisture varies more gradually and predictably than surface soil moisture.
4. Although XGBoost exhibits similar accuracy for arid, semiarid, sub-humid, and humid regions, the accuracy is lowest for the semiarid region (semiarid is the only region with NSE and KGE values below 0.70). The accuracy might be lower in semiarid regions due to more complex topographic, soil, and vegetation characteristics. XGBoost can reproduce the typical dynamics of soil moisture in arid regions, where soil moisture remains low except during responses to precipitation events. It can also reproduce the typical behavior in humid regions, where soil moisture remains high except for prolonged periods with low precipitation.
5. Feature importance analysis identified elevation as the most important topographic variable when the machine learning models are applied to CONUS, and percent clay is typically the most important soil characteristic. Vegetation plays a lesser role in the models, with EVI being the most important vegetation variable. Land cover classification provides little value to the machine learning algorithms. Among SMAP soil moisture products, surface soil moisture is the most important, with rootzone and profile products having lower importance. Furthermore, 3-day and 14-day antecedent soil moisture variables are more important to the algorithms than the current soil moisture.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs16193699/s1.

Author Contributions: Conceptualization, S.A.S. and J.D.N.; methodology, S.A.S.; software, S.A.S.; validation, S.A.S. and J.D.N.; formal analysis, S.A.S.; investigation, S.A.S.; resources, S.A.S. and J.D.N.; data curation, S.A.S.; writing—original draft preparation, S.A.S.; writing—review and editing, S.A.S. and J.D.N.; visualization, S.A.S.; supervision, S.A.S. and J.D.N.; project administration, S.A.S. and J.D.N.; funding acquisition, J.D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the USDA National Institute for Food and Agriculture, Hatch grant 1000065-COL00797, and the National Science Foundation grant number 2312319.

Data Availability Statement: The original contributions presented in the study are included in the article and Supplementary Materials. Further inquiries can be directed to the corresponding author.

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

Appendix A

For the ANN, the number of hidden layers and neurons controls the model's capacity to capture complex patterns, while the activation function determines how neurons process inputs and introduce nonlinearity. The training algorithm affects how efficiently the model learns from data, and the L2 regularization term helps prevent overfitting by penalizing large weights. The learning rate influences the speed of convergence during training, and the maximum number of iterations ensures sufficient learning time.

For the RF model, the number of trees can improve accuracy but increases computational cost, and the maximum depth limits tree complexity to prevent overfitting. The minimum samples to split a node and minimum samples in a leaf ensure nodes are split only when there are enough data, while the maximum features per split balances between randomness and model robustness. The split criterion defines how the quality of a potential split is evaluated.

For XGBoost, the learning rate balances the model's learning speed and accuracy, while the maximum depth and the number of trees control model complexity and learning capacity. The sampling fractions introduce randomness to reduce overfitting, and the minimum sum of instance weight in a leaf prevents over-partitioning with insufficient data.

For CatBoost, the number of trees and step size directly affect the model's learning dynamics, while the maximum depth controls the granularity of learned patterns. Data and feature proportions introduce variability to enhance generalization, and the L2 regularization parameter prevents overfitting by penalizing large model coefficients.

For LightGBM, the number of boosting iterations and learning rate affect the convergence rate and final model accuracy, while the number of leaves and maximum depth influence the model's ability to capture intricate patterns. The minimum data points in a leaf prevent overfitting by requiring sufficient data in terminal nodes, and the regularization term helps to generalize by penalizing overly complex models.

References

1. Martínez-Fernández, J.; González-Zamora, A.; Sánchez, N.; Gumuzzio, A.; Herrero-Jiménez, C.M. Satellite Soil Moisture for Agricultural Drought Monitoring: Assessment of the SMOS Derived Soil Water Deficit Index. *Remote Sens. Environ.* **2016**, *177*, 277–286. [CrossRef]
2. Foster, T.; Mieno, T.; Brozović, N. Satellite-Based Monitoring of Irrigation Water Use: Assessing Measurement Errors and Their Implications for Agricultural Water Management Policy. *Water Resour. Res.* **2020**, *56*, e2020WR028378. [CrossRef]
3. Petropoulos, G.; Srivastava, P.; Piles, M.; Pearson, S. Earth Observation-Based Operational Estimation of Soil Moisture and Evapotranspiration for Agricultural Crops in Support of Sustainable Water Management. *Sustainability* **2018**, *10*, 181. [CrossRef]
4. Zhan, X.; Zheng, W.; Fang, L.; Liu, J.; Hain, C.; Yin, J.; Ek, M. A Preliminary Assessment of the Impact of SMAP Soil Moisture on Numerical Weather Forecasts from GFS and NUWRF Models. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5229–5232.
5. Koster, R.D.; Suarez, M.J. Impact of Land Surface Initialization on Seasonal Precipitation and Temperature Prediction. *J. Hydrometeorol.* **2003**, *4*, 408–423. [CrossRef]
6. Koster, R.D.; Suarez, M.J. Soil Moisture Memory in Climate Models. *J. Hydrometeorol.* **2001**, *2*, 558–570. [CrossRef]
7. Brocca, L.; Ciabatta, L.; Massari, C.; Camici, S.; Tarpanelli, A. Soil Moisture for Hydrological Applications: Open Questions and New Opportunities. *Water* **2017**, *9*, 140. [CrossRef]
8. Bartsch, A.; Balzter, H.; George, C. The Influence of Regional Surface Soil Moisture Anomalies on Forest Fires in Siberia Observed from Satellites. *Environ. Res. Lett.* **2009**, *4*, 045021. [CrossRef]
9. Hou, X.; Orth, R. Observational Evidence of Wildfire-Promoting Soil Moisture Anomalies. *Sci. Rep.* **2020**, *10*, 11008. [CrossRef]
10. Kim, S.; Zhang, R.; Pham, H.; Sharma, A. A Review of Satellite-Derived Soil Moisture and Its Usage for Flood Estimation. *Remote Sens. Earth Syst. Sci.* **2019**, *2*, 225–246. [CrossRef]

11. Xu, L.; Abbaszadeh, P.; Moradkhani, H.; Chen, N.; Zhang, X. Continental Drought Monitoring Using Satellite Soil Moisture, Data Assimilation and an Integrated Drought Index. *Remote Sens. Environ.* **2020**, *250*, 112028. [CrossRef]
12. Ruff, M.S.; Krizek, D.T.; Mirecki, R.M.; Inouye, D.W. Restricted Root Zone Volume: Influence on Growth and Development of Tomato. *J. Am. Soc. Hortic. Sci.* **1987**, *112*, 763–769. [CrossRef]
13. Leenaars, J.G.B.; Claessens, L.; Heuvelink, G.B.M.; Hengl, T.; Ruiperez González, M.; van Bussel, L.G.J.; Guilpart, N.; Yang, H.; Cassman, K.G. Mapping Rootable Depth and Root Zone Plant-Available Water Holding Capacity of the Soil of Sub-Saharan Africa. *Geoderma* **2018**, *324*, 18–36. [CrossRef]
14. Seneviratne, S.I.; Lüthi, D.; Litschi, M.; Schär, C. Land–Atmosphere Coupling and Climate Change in Europe. *Nature* **2006**, *443*, 205–209. [CrossRef] [PubMed]
15. Rodríguez-Iturbe, I.; Porporato, A. *Ecohydrology of Water-Controlled Ecosystems*; Cambridge University Press: Cambridge, UK, 2005; ISBN 9780521819435.
16. 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.; et al. The Soil Moisture Active Passive (SMAP) Mission. *Proc. IEEE* **2010**, *98*, 704–716. [CrossRef]
17. Wagner, W.; Hahn, S.; Kidd, R.; Melzer, T.; Bartalis, Z.; Hasenauer, S.; Figa-Saldaña, J.; de Rosnay, P.; Jann, A.; Schneider, S.; et al. The ASCAT Soil Moisture Product: A Review of Its Specifications, Validation Results, and Emerging Applications. *Meteorol. Z.* **2013**, *22*, 5–33. [CrossRef]
18. Kerr, Y.H.; Waldteufel, P.; Wigneron, J.-P.; Martinuzzi, J.; Font, J.; Berger, M. Soil Moisture Retrieval from Space: The Soil Moisture and Ocean Salinity (SMOS) Mission. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 1729–1735. [CrossRef]
19. Njoku, E.G.; Jackson, T.J.; Lakshmi, V.; Chan, T.K.; Nghiem, S.V. Soil Moisture Retrieval from AMSR-E. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 215–229. [CrossRef]
20. Peng, J.; Loew, A.; Merlin, O.; Verhoest, N.E.C. A Review of Spatial Downscaling of Satellite Remotely Sensed Soil Moisture. *Rev. Geophys.* **2017**, *55*, 341–366. [CrossRef]
21. Tagesson, T.; Horion, S.; Nieto, H.; Zaldo Fornies, V.; Mendiguren González, G.; Bulgin, C.E.; Ghent, D.; Fensholt, R. Disaggregation of SMOS Soil Moisture over West Africa Using the Temperature and Vegetation Dryness Index Based on SEVIRI Land Surface Parameters. *Remote Sens. Environ.* **2018**, *206*, 424–441. [CrossRef]
22. 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.; et al. The SMAP and Copernicus Sentinel 1A/B Microwave Active-Passive High Resolution Surface Soil Moisture Product. *Remote Sens. Environ.* **2019**, *233*, 111380. [CrossRef]
23. Wei, Z.; Meng, Y.; Zhang, W.; Peng, J.; Meng, L. Downscaling SMAP Soil Moisture Estimation with Gradient Boosting Decision Tree Regression over the Tibetan Plateau. *Remote Sens. Environ.* **2019**, *225*, 30–44. [CrossRef]
24. Nuñez-Olivieri, J.; Muñoz-Barreto, J.; Tirado-Corbalá, R.; Lakhankar, T.; Fisher, A. Comparison and Downscale of AMSR2 Soil Moisture Products with In Situ Measurements from the SCAN–NRCS Network over Puerto Rico. *Hydrology* **2017**, *4*, 46. [CrossRef]
25. Vergopolan, N.; Chaney, N.W.; Beck, H.E.; Pan, M.; Sheffield, J.; Chan, S.; Wood, E.F. Combining Hyper-Resolution Land Surface Modeling with SMAP Brightness Temperatures to Obtain 30-m Soil Moisture Estimates. *Remote Sens. Environ.* **2020**, *242*, 111740. [CrossRef]
26. Fischer, S.C. *Assessing the Influence of Model Inputs on Performance of the EMT+VS Soil Moisture Downscaling Model for a Large Foothills Region in Northern Colorado*; Colorado State University: Fort Collins, CO, USA, 2024.
27. Dumedah, G.; Walker, J.P.; Merlin, O. Root-Zone Soil Moisture Estimation from Assimilation of Downscaled Soil Moisture and Ocean Salinity Data. *Adv. Water Resour.* **2015**, *84*, 14–22. [CrossRef]
28. Zhang, D.; Zhou, G. Estimation of Soil Moisture from Optical and Thermal Remote Sensing: A Review. *Sensors* **2016**, *16*, 1308. [CrossRef]
29. Carlson, T. An Overview of the "Triangle Method" for Estimating Surface Evapotranspiration and Soil Moisture from Satellite Imagery. *Sensors* **2007**, *7*, 1612–1629. [CrossRef]
30. Carlson, T.N.; Petropoulos, G.P. A New Method for Estimating of Evapotranspiration and Surface Soil Moisture from Optical and Thermal Infrared Measurements: The Simplified Triangle. *Int. J. Remote Sens.* **2019**, *40*, 7716–7729. [CrossRef]
31. Petropoulos, G.; Carlson, T.N.; Wooster, M.J.; Islam, S. A Review of Ts/VI Remote Sensing Based Methods for the Retrieval of Land Surface Energy Fluxes and Soil Surface Moisture. *Prog. Phys. Geogr. Earth Environ.* **2009**, *33*, 224–250. [CrossRef]
32. Sadeghi, M.; Babaeian, E.; Tuller, M.; Jones, S.B. The Optical Trapezoid Model: A Novel Approach to Remote Sensing of Soil Moisture Applied to Sentinel-2 and Landsat-8 Observations. *Remote Sens. Environ.* **2017**, *198*, 52–68. [CrossRef]
33. Wang, W.; Huang, D.; Wang, X.-G.; Liu, Y.-R.; Zhou, F. Estimation of Soil Moisture Using Trapezoidal Relationship between Remotely Sensed Land Surface Temperature and Vegetation Index. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1699–1712. [CrossRef]
34. Amani, M.; Salehi, B.; Mahdavi, S.; Masjedi, A.; Dehnavi, S. Temperature-Vegetation-Soil Moisture Dryness Index (TVMDI). *Remote Sens. Environ.* **2017**, *197*, 1–14. [CrossRef]
35. Anderson, W.B.; Zaitchik, B.F.; Hain, C.R.; Anderson, M.C.; Yilmaz, M.T.; Mecikalski, J.; Schultz, L. Towards an Integrated Soil Moisture Drought Monitor for East Africa. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 2893–2913. [CrossRef]
36. Feng, H.; Chen, C.; Dong, H.; Wang, J.; Meng, Q. Modified Shortwave Infrared Perpendicular Water Stress Index: A Farmland Water Stress Monitoring Method. *J. Appl. Meteorol. Climatol.* **2013**, *52*, 2024–2032. [CrossRef]
37. Ghulam, A.; Qin, Q.; Zhan, Z. Designing of the Perpendicular Drought Index. *Environ. Geol.* **2007**, *52*, 1045–1052. [CrossRef]

38. Hu, T.; Renzullo, L.J.; van Dijk, A.I.J.M.; He, J.; Tian, S.; Xu, Z.; Zhou, J.; Liu, T.; Liu, Q. Monitoring Agricultural Drought in Australia Using MTSAT-2 Land Surface Temperature Retrievals. *Remote Sens. Environ.* **2020**, *236*, 111419. [CrossRef]
39. Van doninck, J.; Peters, J.; De Baets, B.; De Clercq, E.M.; Ducheyne, E.; Verhoest, N.E.C. The Potential of Multitemporal Aqua and Terra MODIS Apparent Thermal Inertia as a Soil Moisture Indicator. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 934–941. [CrossRef]
40. Amazirh, A.; Merlin, O.; Er-Raki, S.; Gao, Q.; Rivalland, V.; Malbeteau, Y.; Khabba, S.; Escorihuela, M.J. Retrieving Surface Soil Moisture at High Spatio-Temporal Resolution from a Synergy between Sentinel-1 Radar and Landsat Thermal Data: A Study Case over Bare Soil. *Remote Sens. Environ.* **2018**, *211*, 321–337. [CrossRef]
41. Gao, X.; Zhao, X.; Brocca, L.; Huo, G.; Lv, T.; Wu, P. Depth Scaling of Soil Moisture Content from Surface to Profile: Multistation Testing of Observation Operators. *Hydrol. Earth Syst. Sci. Discuss.* **2017**, *2017*, 1–25. [CrossRef]
42. Bastiaanssen, W.G.M.; Molden, D.J.; Makin, I.W. Remote Sensing for Irrigated Agriculture: Examples from Research and Possible Applications. *Agric. Water Manag.* **2000**, *46*, 137–155. [CrossRef]
43. Chen, J.M.; Liu, J. Evolution of Evapotranspiration Models Using Thermal and Shortwave Remote Sensing Data. *Remote Sens. Environ.* **2020**, *237*, 111594. [CrossRef]
44. Hain, C.R.; Mecikalski, J.R.; Anderson, M.C. Retrieval of an Available Water-Based Soil Moisture Proxy from Thermal Infrared Remote Sensing. Part I: Methodology and Validation. *J. Hydrometeorol.* **2009**, *10*, 665–683. [CrossRef]
45. Sahaar, S.A.; Niemann, J.D. Impact of Regional Characteristics on the Estimation of Root-Zone Soil Moisture from the Evaporative Index or Evaporative Fraction. *Agric. Water Manag.* **2020**, *238*, 106225. [CrossRef]
46. Scott, C.A.; Bastiaanssen, W.G.M.; Ahmad, M.-D. Mapping Root Zone Soil Moisture Using Remotely Sensed Optical Imagery. *J. Irrig. Drain. Eng.* **2003**, *129*, 326–335. [CrossRef]
47. Chauhan, N.S.; Miller, S.; Ardanuy, P. Spaceborne Soil Moisture Estimation at High Resolution: A Microwave-Optical/IR Synergistic Approach. *Int. J. Remote Sens.* **2003**, *24*, 4599–4622. [CrossRef]
48. Merlin, O.; Rudiger, C.; Al Bitar, A.; Richaume, P.; Walker, J.P.; Kerr, Y.H. Disaggregation of SMOS Soil Moisture in Southeastern Australia. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1556–1571. [CrossRef]
49. Piles, M.; Petropoulos, G.P.; Sánchez, N.; González-Zamora, Á.; Ireland, G. Towards Improved Spatio-Temporal Resolution Soil Moisture Retrievals from the Synergy of SMOS and MSG SEVIRI Spaceborne Observations. *Remote Sens. Environ.* **2016**, *180*, 403–417. [CrossRef]
50. Portal, G.; Vall-Llosscra, M.; Piles, M.; Camps, A.; Chaparro, D.; Pablos, M.; Rossato, L.; Aabouch, K. Microwave and Optical Data Fusion for Global Mapping of Soil Moisture at High Resolution. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 341–344.
51. Bai, L.; Long, D.; Yan, L. Estimation of Surface Soil Moisture with Downscaled Land Surface Temperatures Using a Data Fusion Approach for Heterogeneous Agricultural Land. *Water Resour. Res.* **2019**, *55*, 1105–1128. [CrossRef]
52. Long, D.; Bai, L.; Yan, L.; Zhang, C.; Yang, W.; Lei, H.; Quan, J.; Meng, X.; Shi, C. Generation of Spatially Complete and Daily Continuous Surface Soil Moisture of High Spatial Resolution. *Remote Sens. Environ.* **2019**, *233*, 111364. [CrossRef]
53. Huang, S.; Zhang, X.; Chen, N.; Ma, H.; Zeng, J.; Fu, P.; Nam, W.-H.; Niyogi, D. Generating High-Accuracy and Cloud-Free Surface Soil Moisture at 1 Km Resolution by Point-Surface Data Fusion over the Southwestern U.S. *Agric. For. Meteorol.* **2022**, *321*, 108985. [CrossRef]
54. Owe, M.; de Jeu, R.; Walker, J. A Methodology for Surface Soil Moisture and Vegetation Optical Depth Retrieval Using the Microwave Polarization Difference Index. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 1643–1654. [CrossRef]
55. Leng, P.; Li, Z.-L.; Liao, Q.-Y.; Gao, M.-F.; Duan, S.-B.; Zhang, X.; Shang, G.-F. Determination of All-Sky Surface Soil Moisture at Fine Spatial Resolution Synergistically Using Optical/Thermal Infrared and Microwave Measurements. *J. Hydrol.* **2019**, *579*, 124167. [CrossRef]
56. Mishra, V.; Cruise, J.F.; Hain, C.R.; Mecikalski, J.R.; Anderson, M.C. Development of Soil Moisture Profiles through Coupled Microwave–Thermal Infrared Observations in the Southeastern United States. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 4935–4957. [CrossRef]
57. Wangemann, S.G.; Kohl, R.A.; Molumeli, P.A. Infiltration and Percolation Influenced by Antecedent Soil Water Content and Air Entrapment. *Trans. ASAE* **2000**, *43*, 1517–1523. [CrossRef]
58. Tong, C.; Wang, H.; Magagi, R.; Goïta, K.; Zhu, L.; Yang, M.; Deng, J. Soil Moisture Retrievals by Combining Passive Microwave and Optical Data. *Remote Sens.* **2020**, *12*, 3173. [CrossRef]
59. Zhang, Y.-K.; Schilling, K.E. Effects of Land Cover on Water Table, Soil Moisture, Evapotranspiration, and Groundwater Recharge: A Field Observation and Analysis. *J. Hydrol.* **2006**, *319*, 328–338. [CrossRef]
60. Ranney, K.J.; Niemann, J.D.; Lehman, B.M.; Green, T.R.; Jones, A.S. A Method to Downscale Soil Moisture to Fine Resolutions Using Topographic, Vegetation, and Soil Data. *Adv. Water Resour.* **2015**, *76*, 81–96. [CrossRef]
61. Takagi, K.; Lin, H.S. Temporal Dynamics of Soil Moisture Spatial Variability in the Shale Hills Critical Zone Observatory. *Vadose Zone J.* **2011**, *10*, 832–842. [CrossRef]
62. Crow, W.T.; Berg, A.A.; Cosh, M.H.; Loew, A.; Mohanty, B.P.; Panciera, R.; de Rosnay, P.; Ryu, D.; Walker, J.P. Upscaling Sparse Ground-Based Soil Moisture Observations for the Validation of Coarse-Resolution Satellite Soil Moisture Products. *Rev. Geophys.* **2012**, *50*, RG2002. [CrossRef]
63. Rosenbaum, U.; Bogena, H.R.; Herbst, M.; Huisman, J.A.; Peterson, T.J.; Weuthen, A.; Western, A.W.; Vereecken, H. Seasonal and Event Dynamics of Spatial Soil Moisture Patterns at the Small Catchment Scale. *Water Resour. Res.* **2012**, *48*, W10544. [CrossRef]

64. Evans, J.G.; Ward, H.C.; Blake, J.R.; Hewitt, E.J.; Morrison, R.; Fry, M.; Ball, L.A.; Doughty, L.C.; Libre, J.W.; Hitt, O.E.; et al. Soil Water Content in Southern England Derived from a Cosmic-ray Soil Moisture Observing System—COSMOS-UK. *Hydrol. Process.* **2016**, *30*, 4987–4999. [CrossRef]
65. Famiglietti, J.S.; Ryu, D.; Berg, A.A.; Rodell, M.; Jackson, T.J. Field Observations of Soil Moisture Variability across Scales. *Water Resour. Res.* **2008**, *44*, W01423. [CrossRef]
66. Geroy, I.J.; Gribb, M.M.; Marshall, H.P.; Chandler, D.G.; Benner, S.G.; McNamara, J.P. Aspect Influences on Soil Water Retention and Storage. *Hydrol. Process.* **2011**, *25*, 3836–3842. [CrossRef]
67. Coleman, M.L.; Niemann, J.D. Controls on Topographic Dependence and Temporal Instability in Catchment-scale Soil Moisture Patterns. *Water Resour. Res.* **2013**, *49*, 1625–1642. [CrossRef]
68. Cowley, G.S.; Niemann, J.D.; Green, T.R.; Seyfried, M.S.; Jones, A.S.; Grazaitis, P.J. Impacts of Precipitation and Potential Evapotranspiration Patterns on Downscaling Soil Moisture in Regions with Large Topographic Relief. *Water Resour. Res.* **2017**, *53*, 1553–1574. [CrossRef]
69. Abbaszadeh, P.; Moradkhani, H.; Zhan, X. Downscaling SMAP Radiometer Soil Moisture Over the CONUS Using an Ensemble Learning Method. *Water Resour. Res.* **2019**, *55*, 324–344. [CrossRef]
70. Abowarda, A.S.; Bai, L.; Zhang, C.; Long, D.; Li, X.; Huang, Q.; Sun, Z. 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.* **2021**, *255*, 112301. [CrossRef]
71. Liu, Y.; Jing, W.; Wang, Q.; Xia, X. Generating High-Resolution Daily Soil Moisture by Using Spatial Downscaling Techniques: A Comparison of Six Machine Learning Algorithms. *Adv. Water Resour.* **2020**, *141*, 103601. [CrossRef]
72. Peng, J.; Albergel, C.; Balenzano, A.; Brocca, L.; Cartus, O.; Cosh, M.H.; Crow, W.T.; Dabrowska-Zielinska, K.; Dadson, S.; Davidson, M.W.J.; et al. A Roadmap for High-Resolution Satellite Soil Moisture Applications—Confronting Product Characteristics with User Requirements. *Remote Sens. Environ.* **2021**, *252*, 112162. [CrossRef]
73. Fathololoumi, S.; Karimi Firozjaei, M.; Biswas, A. Improving Spatial Resolution of Satellite Soil Water Index (SWI) Maps under Clear-Sky Conditions Using a Machine Learning Approach. *J. Hydrol.* **2022**, *615*, 128709. [CrossRef]
74. Zhao, W.; Sánchez, N.; Lu, H.; Li, A. A Spatial Downscaling Approach for the SMAP Passive Surface Soil Moisture Product Using Random Forest Regression. *J. Hydrol.* **2018**, *563*, 1009–1024. [CrossRef]
75. Fang, K.; Pan, M.; Shen, C. The Value of SMAP for Long-Term Soil Moisture Estimation with the Help of Deep Learning. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2221–2233. [CrossRef]
76. Ali, I.; Greifeneder, F.; Stamenkovic, J.; Neumann, M.; Notarnicola, C. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *Remote Sens.* **2015**, *7*, 16398–16421. [CrossRef]
77. Greifeneder, F.; Notarnicola, C.; Wagner, W. A Machine Learning-Based Approach for Surface Soil Moisture Estimations with Google Earth Engine. *Remote Sens.* **2021**, *13*, 2099. [CrossRef]
78. Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T. Improving Performance of Spatio-Temporal Machine Learning Models Using Forward Feature Selection and Target-Oriented Validation. *Environ. Model. Softw.* **2018**, *101*, 1–9. [CrossRef]
79. Rani, A.; Kumar, N.; Kumar, J.; Kumar, J.; Sinha, N.K. Machine Learning for Soil Moisture Assessment. In *Deep Learning for Sustainable Agriculture*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 143–168. ISBN 9780323852142.
80. Singh, A.; Gaurav, K. Deep Learning and Data Fusion to Estimate Surface Soil Moisture from Multi-Sensor Satellite Images. *Sci. Rep.* **2023**, *13*, 2251. [CrossRef]
81. Fuentes, I.; Padarian, J.; Vervoort, R.W. Towards near Real-Time National-Scale Soil Water Content Monitoring Using Data Fusion as a Downscaling Alternative. *J. Hydrol.* **2022**, *609*, 127705. [CrossRef]
82. Karthikeyan, L.; Mishra, A.K. Multi-Layer High-Resolution Soil Moisture Estimation Using Machine Learning over the United States. *Remote Sens. Environ.* **2021**, *266*, 112706. [CrossRef]
83. Liu, E.; Zhu, Y.; Lü, H.; Horton, R.; Gou, Q.; Wang, X.; Ding, Z.; Xu, H.; Pan, Y. Estimation and Assessment of the Root Zone Soil Moisture from Near-Surface Measurements over Huai River Basin. *Atmosphere* **2023**, *14*, 124. [CrossRef]
84. Dorigo, W.A.; Xaver, A.; Vreugdenhil, M.; Gruber, A.; Hegyiová, A.; Sanchis-Dufau, A.D.; Zamojski, D.; Cordes, C.; Wagner, W.; Drusch, M. Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network. *Vadose Zone J.* **2013**, *12*, 1–21. [CrossRef]
85. Cook, D.R. *Soil Temperature and Moisture Profile (STAMP) System Handbook*; Technical Reports; DOE Office of Science Atmospheric Radiation Measurement (ARM) Program: USA, 2016. Available online: https://www.osti.gov/biblio/1332724 (accessed on 1 November 2023).
86. Zreda, M.; Desilets, D.; Ferré, T.P.A.; Scott, R.L. Measuring Soil Moisture Content Non-Invasively at Intermediate Spatial Scale Using Cosmic-Ray Neutrons. *Geophys. Res. Lett.* **2008**, *35*, L21402. [CrossRef]
87. Baldocchi, D.; Falge, E.; Gu, L.; Olson, R.; Hollinger, D.; Running, S.; Anthoni, P.; Bernhofer, C.; Davis, K.; Evans, R.; et al. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem–Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. *Bull. Am. Meteorol. Soc.* **2001**, *82*, 2415–2434. [CrossRef]
88. Osenga, E.C.; Vano, J.A.; Arnott, J.C. A Community-Supported Weather and Soil Moisture Monitoring Database of the Roaring Fork Catchment of the Colorado River Headwaters. *Hydrol. Process.* **2021**, *35*, e14081. [CrossRef]
89. Schaefer, G.L.; Cosh, M.H.; Jackson, T.J. The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). *J. Atmos. Ocean. Technol.* **2007**, *24*, 2073–2077. [CrossRef]

90. Fleming, S.W.; Zukiewicz, L.; Strobel, M.L.; Hofman, H.; Goodbody, A.G. SNOTEL, the Soil Climate Analysis Network, and Water Supply Forecasting at the Natural Resources Conservation Service: Past, Present, and Future. *JAWRA J. Am. Water Resour. Assoc.* **2023**, *59*, 585–599. [CrossRef]
91. 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. The Texas Soil Observation Network:A Comprehensive Soil Moisture Dataset for Remote Sensing and Land Surface Model Validation. *Vadose Zone J.* **2019**, *18*, 1–20. [CrossRef]
92. Bell, J.E.; Palecki, M.A.; Baker, C.B.; Collins, W.G.; Lawrimore, J.H.; Leeper, R.D.; Hall, M.E.; Kochendorfer, J.; Meyers, T.P.; Wilson, T.; et al. U.S. Climate Reference Network Soil Moisture and Temperature Observations. *J. Hydrometeorol.* **2013**, *14*, 977–988. [CrossRef]
93. Reichle, R.; De Lannoy, G.; Koster, R.D.; Crow, W.T.; Kimball, J.S.; Liu, Q.; Bechtold, M. *SMAP L4 Global 3-Hourly 9 Km EASE-Grid Surface and Root Zone Soil Moisture Analysis Update, Version 7*; NASA National Snow and Ice Data Center Distributed Active Archive Center: Boulder, CO, USA, 2022.
94. Reichle, R.H.; De Lannoy, G.J.M.; Liu, Q.; Ardizzone, J.V.; Colliander, A.; Conaty, A.; Crow, W.; Jackson, T.J.; Jones, L.A.; Kimball, J.S.; et al. Assessment of the SMAP Level-4 Surface and Root-Zone Soil Moisture Product Using In Situ Measurements. *J. Hydrometeorol.* **2017**, *18*, 2621–2645. [CrossRef]
95. Davis, B.N.; Werpy, J.; Friesz, A.; Impecoven, K.; Quenzer, R.L.; Maiersperger, T.; Meyer, D.J. Interactive Access to LP DAAC Satellite Data Archives Through a Combination of Open-Source and Custom Middleware Web Services. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 8–20. [CrossRef]
96. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
97. Dorigo, W.; Himmelbauer, I.; Aberer, D.; Schremmer, L.; Petrakovic, I.; Zappa, L.; Preimesberger, W.; Xaver, A.; Annor, F.; Ardö, J.; et al. The International Soil Moisture Network: Serving Earth System Science for over a Decade. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 5749–5804. [CrossRef]
98. Dewitz, J. National Land Cover Database (NLCD) 2019 Products (Ver. 2.0, June 2021): U.S. Geological Survey Data Release. 2019. Available online: https://data.usgs.gov/datacatalog/data/USGS:60cb3da7d34e86b938a30cb9 (accessed on 1 June 2023).
99. Dewitz, J.; U.S. Geological Survey. National Land Cover Database (NLCD) 2019 Products (Ver. 3.0, February 2024): U.S. Geological Survey Data Release. 2021. Available online: https://www.sciencebase.gov/catalog/item/5f21cef582cef313ed940043 (accessed on 1 June 2023).
100. Didan, K. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid. V061. 2021. Available online: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD13Q1 (accessed on 15 September 2023).
101. Myneni, R.; Knyazikhin, Y.; Park, T. MODIS/Terra+Aqua Leaf Area Index/FPAR 4-Day L4 Global 500m SIN Grid. V061. 2021. Available online: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD15A3H (accessed on 1 June 2023).
102. Soil Survey Staff Gridded National Soil Survey Geographic (GNATSGO) Database for the Conterminous United States. Available online: https://nrcs.app.box.com/v/soils (accessed on 5 January 2023).
103. Huffman, G.J.; Stocker, E.F.; Bolvin, D.T.; Nelkin, E.J.; Jackson, T. *GPM IMERG Final Precipitation L3 Half Hourly 0.1 Degree x 0.1 Degree V06*; Goddard Earth Sciences Data and Information Services Center (GES DISC): Greenbelt, MD, USA, 2019.
104. Hook, S.; Hulley, G. ECOSTRESS Swath Land Surface Temperature and Emissivity Instantaneous L2 Global 70 m v002 [Data Set]. NASA EOSDIS Land Processes Distributed Active Archive Center. 2022. Available online: https://lpdaac.usgs.gov/products/eco_l2_lstev002/ (accessed on 1 June 2023).
105. Hook, S.; Fisher, J. ECOSTRESS Evapotranspiration PT-JPL Daily L3 Global 70 m V001 [Data Set]. NASA EOSDIS Land Processes Distributed Active Archive Center. 2019. Available online: https://lpdaac.usgs.gov/products/eco3etptjplv001/ (accessed on 1 June 2023).
106. Zomer, R.J.; Xu, J.; Trabucco, A. Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. *Sci. Data* **2022**, *9*, 409. [CrossRef] [PubMed]
107. NASA; JPL. NASA Shuttle Radar Topography Mission Global 1 arc second [Data Set]. NASA EOSDIS Land Processes Distributed Active Archive Center. 2013. Available online: https://lpdaac.usgs.gov/products/srtmgl1v003/ (accessed on 1 June 2023).
108. Theobald, D.M.; Harrison-Atlas, D.; Monahan, W.B.; Albano, C.M. Ecologically-Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning. *PLoS ONE* **2015**, *10*, e0143619. [CrossRef] [PubMed]
109. Wei, L.; Yang, M.; Li, Z.; Shao, J.; Li, L.; Chen, P.; Li, S.; Zhao, R. Experimental Investigation of Relationship between Infiltration Rate and Soil Moisture under Rainfall Conditions. *Water* **2022**, *14*, 1347. [CrossRef]
110. Matsushita, B.; Yang, W.; Chen, J.; Onda, Y.; Qiu, G. Sensitivity of the Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) to Topographic Effects: A Case Study in High-Density Cypress Forest. *Sensors* **2007**, *7*, 2636–2651. [CrossRef]
111. Xu, B.; Park, T.; Yan, K.; Chen, C.; Zeng, Y.; Song, W.; Yin, G.; Li, J.; Liu, Q.; Knyazikhin, Y.; et al. Analysis of Global LAI/FPAR Products from VIIRS and MODIS Sensors for Spatio-Temporal Consistency and Uncertainty from 2012–2016. *Forests* **2018**, *9*, 73. [CrossRef]
112. Churkina, G.; Running, S.W. Contrasting Climatic Controls on the Estimated Productivity of Global Terrestrial Biomes. *Ecosystems* **1998**, *1*, 206–215. [CrossRef]
113. Machado, R.; Serralheiro, R. Soil Salinity: Effect on Vegetable Crop Growth. Management Practices to Prevent and Mitigate Soil Salinization. *Horticulturae* **2017**, *3*, 30. [CrossRef]

114. Yadav, D.S.; Jaiswal, B.; Gautam, M.; Agrawal, M. Soil Acidification and Its Impact on Plants. In *Plant Responses to Soil Pollution*; Springer: Singapore, 2020; pp. 1–26. ISBN 9789811549649.
115. Rossiter, D.G.; Poggio, L.; Beaudette, D.; Libohova, Z. How Well Does Digital Soil Mapping Represent Soil Geography? An Investigation from the USA. *SOIL* **2022**, *8*, 559–586. [CrossRef]
116. Krishnan, S.; Pradhan, A.; Indu, J. Estimation of High-Resolution Precipitation Using Downscaled Satellite Soil Moisture and SM2RAIN Approach. *J. Hydrol.* **2022**, *610*, 127926. [CrossRef]
117. Farahani, A.; Moradikhaneghahi, M.; Ghayoomi, M.; Jacobs, J.M. Application of Soil Moisture Active Passive (SMAP) Satellite Data in Seismic Response Assessment. *Remote Sens.* **2022**, *14*, 4375. [CrossRef]
118. Beck, H.E.; Pan, M.; Miralles, D.G.; Reichle, R.H.; Dorigo, W.A.; Hahn, S.; Sheffield, J.; Karthikeyan, L.; Balsamo, G.; Parinussa, R.M.; et al. Evaluation of 18 Satellite- and Model-Based Soil Moisture Products Using in Situ Measurements from 826 Sensors. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 17–40. [CrossRef]
119. Fisher, J.B.; Lee, B.; Purdy, A.J.; Halverson, G.H.; Dohlen, M.B.; Cawse-Nicholson, K.; Wang, A.; Anderson, R.G.; Aragon, B.; Arain, M.A.; et al. ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration from the International Space Station. *Water Resour. Res.* **2020**, *56*, e2019WR026058. [CrossRef]
120. Trabucco, A.; Zomer, R.J. Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database V2. *CGIAR Consort. Spat. Inf.* **2019**, *10*, m9. [CrossRef]
121. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
122. Famiglietti, J.S.; Rudnicki, J.W.; Rodell, M. Variability in Surface Moisture Content along a Hillslope Transect: Rattlesnake Hill, Texas. *J. Hydrol.* **1998**, *210*, 259–281. [CrossRef]
123. Western, A.W.; Grayson, R.B.; Blöschl, G.; Willgoose, G.R.; McMahon, T.A. Observed Spatial Organization of Soil Moisture and Its Relation to Terrain Indices. *Water Resour. Res.* **1999**, *35*, 797–810. [CrossRef]
124. Burrough, P.A.; McDonnell, R.A. *Principles of Geographical Information Systems*; Oxford University Press: Oxford, UK, 1998; ISBN 0198233655.
125. Celik, M.F.; Isik, M.S.; Yuzugullu, O.; Fajraoui, N.; Erten, E. Soil Moisture Prediction from Remote Sensing Images Coupled with Climate, Soil Texture and Topography via Deep Learning. *Remote Sens.* **2022**, *14*, 5584. [CrossRef]
126. Luo, P.; Song, Y.; Huang, X.; Ma, H.; Liu, J.; Yao, Y.; Meng, L. Identifying Determinants of Spatio-Temporal Disparities in Soil Moisture of the Northern Hemisphere Using a Geographically Optimal Zones-Based Heterogeneity Model. *ISPRS J. Photogramm. Remote Sens.* **2022**, *185*, 111–128. [CrossRef]
127. Iverson, L.R.; Prasad, A.M.; Rebbeck, J. A Comparison of the Integrated Moisture Index and the Topographic Wetness Index as Related to Two Years of Soil Moisture Monitoring in Zaleski State Forest, Ohio. In Proceedings of the 14th Central Hardwood Forest conference, Wooster, OH, USA, 16–19 March 2004.
128. Bisong, E. Introduction to Scikit-Learn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*; Apress: Berkeley, CA, USA, 2019; pp. 215–229.
129. Jiang, D.; Lin, W.; Raghavan, N. A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques. *IEEE Access* **2020**, *8*, 197885–197895. [CrossRef]
130. Abiy, A.Z.; Wiederholt, R.P.; Lagerwall, G.L.; Melesse, A.M.; Davis, S.E. Multilayer Feedforward Artificial Neural Network Model to Forecast Florida Bay Salinity with Climate Change. *Water* **2022**, *14*, 3495. [CrossRef]
131. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
132. Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
133. Kornelsen, K.C.; Coulibaly, P. Root-zone Soil Moisture Estimation Using Data-driven Methods. *Water Resour. Res.* **2014**, *50*, 2946–2962. [CrossRef]
134. Adab, H.; Morbidelli, R.; Saltalippi, C.; Moradian, M.; Ghalhari, G.A.F. Machine Learning to Estimate Surface Soil Moisture from Remote Sensing Data. *Water* **2020**, *12*, 3223. [CrossRef]
135. Jamei, M.; Karbasi, M.; Malik, A.; Jamei, M.; Kisi, O.; Yaseen, Z.M. Long-Term Multi-Step Ahead Forecasting of Root Zone Soil Moisture in Different Climates: Novel Ensemble-Based Complementary Data-Intelligent Paradigms. *Agric. Water Manag.* **2022**, *269*, 107679. [CrossRef]
136. Sun, H.; Zhang, X.; Zhao, X. Series or Parallel? An Exploration in Coupling Physical Model and Machine Learning Method for Disaggregating Satellite Microwave Soil Moisture. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
137. McCulloch, W.S.; Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
138. Demir-Kavuk, O.; Kamada, M.; Akutsu, T.; Knapp, E.-W. Prediction Using Step-Wise L1, L2 Regularization and Feature Selection for Small Data Sets with Large Number of Features. *BMC Bioinform.* **2011**, *12*, 412. [CrossRef] [PubMed]
139. Thirumalaiah, K.; Deo, M.C. River Stage Forecasting Using Artificial Neural Networks. *J. Hydrol. Eng.* **1998**, *3*, 26–32. [CrossRef]
140. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *arXiv* **2017**, arXiv:1706.09516.
141. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for Big Data: An Interdisciplinary Review. *J. Big Data* **2020**, *7*, 94. [CrossRef]
142. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; p. 30.

143. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the Stratification of Multi-Label Data. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 145–158. ISBN 9783642238079.
144. Paulik, C.; Dorigo, W.; Wagner, W.; Kidd, R. Validation of the ASCAT Soil Water Index Using in Situ Data from the International Soil Moisture Network. *Int. J. Appl. Earth Obs. Geoinf.* **2014**, *30*, 1–8. [CrossRef]
145. Morales, P.; Sykes, M.T.; Prentice, I.C.; Smith, P.; Smith, B.; Bugmann, H.; Zierl, B.; Friedlingstein, P.; Viovy, N.; Sabaté, S.; et al. Comparing and Evaluating Process-based Ecosystem Model Predictions of Carbon and Water Fluxes in Major European Forest Biomes. *Glob. Chang. Biol.* **2005**, *11*, 2211–2233. [CrossRef] [PubMed]
146. Entekhabi, D.; Reichle, R.H.; Koster, R.D.; Crow, W.T. Performance Metrics for Soil Moisture Retrievals and Application Requirements. *J. Hydrometeorol.* **2010**, *11*, 832–840. [CrossRef]
147. Li, L.; Dai, Y.; Shangguan, W.; Wei, Z.; Wei, N.; Li, Q. Causality-Structured Deep Learning for Soil Moisture Predictions. *J. Hydrometeorol.* **2022**, *23*, 1315–1331. [CrossRef]
148. Nash, J.E.; Sutcliffe, J.V. River Flow Forecasting through Conceptual Models Part I—A Discussion of Principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
149. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]
150. UNEP (United Nations Environment Programme). *World Atlas of Desertification*, 2nd ed.; UNEP: Nairobi, Kenya, 1997.
151. Jamali, S.; Seaquist, J.W.; Ardo, J.; Eklundh, L. Investigating Temporal Relationships between Rainfall, Soil Moisture and MODIS-Derived NDVI and EVI for Six Sites in Africa. In Proceedings of the 34th International Symposium on Remote Sensing of Environment—The GEOSS Era: Towards Operational Environmental Monitoring, Sydney, Australia, 10–15 April 2011.
152. Santos, W.J.R.; Silva, B.M.; Oliveira, G.C.; Volpato, M.M.L.; Lima, J.M.; Curi, N.; Marques, J.J. Soil Moisture in the Root Zone and Its Relation to Plant Vigor Assessed by Remote Sensing at Management Scale. *Geoderma* **2014**, *221–222*, 91–95. [CrossRef]
153. Méndez-Barroso, L.A.; Vivoni, E.R.; Watts, C.J.; Rodríguez, J.C. Seasonal and Interannual Relations between Precipitation, Surface Soil Moisture and Vegetation Dynamics in the North American Monsoon Region. *J. Hydrol.* **2009**, *377*, 59–70. [CrossRef]
154. Wang, T.; Franz, T.E.; Li, R.; You, J.; Shulski, M.D.; Ray, C. Evaluating Climate and Soil Effects on Regional Soil Moisture Spatial Variability Using EOFs. *Water Resour. Res.* **2017**, *53*, 4022–4035. [CrossRef]
155. Zhang, Y.-Y.; Wu, W.; Liu, H. Factors Affecting Variations of Soil PH in Different Horizons in Hilly Regions. *PLoS ONE* **2019**, *14*, e0218563. [CrossRef]
156. Song, Y.M.; Wang, Z.F.; Qi, L.L.; Huang, A.N. Soil Moisture Memory and Its Effect on the Surface Water and Heat Fluxes on Seasonal and Interannual Time Scales. *J. Geophys. Res. Atmos.* **2019**, *124*, 10730–10741. [CrossRef]
157. Martínez-Fernández, J.; González-Zamora, A.; Almendra-Martín, L. Soil Moisture Memory and Soil Properties: An Analysis with the Stored Precipitation Fraction. *J. Hydrol.* **2021**, *593*, 125622. [CrossRef]
158. Orth, R.; Seneviratne, S.I. Analysis of Soil Moisture Memory from Observations in Europe. *J. Geophys. Res. Atmos.* **2012**, *117*, 15115–15117. [CrossRef]
159. Rong, L.; Duan, X.; Feng, D.; Zhang, G. Soil Moisture Variation in a Farmed Dry-Hot Valley Catchment Evaluated by a Redundancy Analysis Approach. *Water* **2017**, *9*, 92. [CrossRef]
160. Jawson, S.D.; Niemann, J.D. Spatial Patterns from EOF Analysis of Soil Moisture at a Large Scale and Their Dependence on Soil, Land-Use, and Topographic Properties. *Adv. Water Resour.* **2007**, *30*, 366–381. [CrossRef]
161. Senyurek, V.; Lei, F.; Boyd, D.; Kurum, M.; Gurbuz, A.C.; Moorhead, R. Machine Learning-Based CYGNSS Soil Moisture Estimates over ISMN Sites in CONUS. *Remote Sens.* **2020**, *12*, 1168. [CrossRef]
162. Ren, Y.; Ling, F.; Wang, Y. Research on Provincial-Level Soil Moisture Prediction Based on Extreme Gradient Boosting Model. *Agriculture* **2023**, *13*, 927. [CrossRef]
163. Du, J.; Kimball, J.S.; Bindlish, R.; Walker, J.P.; Watts, J.D. Local Scale (3-m) Soil Moisture Mapping Using SMAP and Planet SuperDove. *Remote Sens.* **2022**, *14*, 3812. [CrossRef]
164. Park, S.; Im, J.; Park, S.; Rhee, J. AMSR2 Soil Moisture Downscaling Using Multisensor Products through Machine Learning Approach. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1984–1987.
165. Huang, S.; Zhang, X.; Wang, C.; Chen, N. Two-Step Fusion Method for Generating 1 Km Seamless Multi-Layer Soil Moisture with High Accuracy in the Qinghai-Tibet Plateau. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 346–363. [CrossRef]
166. Yang, Z.; He, Q.; Miao, S.; Wei, F.; Yu, M. Surface Soil Moisture Retrieval of China Using Multi-Source Data and Ensemble Learning. *Remote Sens.* **2023**, *15*, 2786. [CrossRef]
167. Nguyen, T.T.; Ngo, H.H.; Guo, W.; Chang, S.W.; Nguyen, D.D.; Nguyen, C.T.; Zhang, J.; Liang, S.; Bui, X.T.; Hoang, N.B. A Low-Cost Approach for Soil Moisture Prediction Using Multi-Sensor Data and Machine Learning Algorithm. *Sci. Total Environ.* **2022**, *833*, 155066. [CrossRef] [PubMed]