Machine learning based estimation of field-scale daily, high resolution, multi-depth soil moisture for the Western and Midwestern United States

Yushu Xia\textsuperscript{1}, Jennifer D. Watts\textsuperscript{1}, Megan B. Machmuller\textsuperscript{2} and Jonathan Sanderman\textsuperscript{1}

\textsuperscript{1}Woodwell Climate Research Center, Falmouth, Massachusetts, United States
\textsuperscript{2}Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado, United States

ABSTRACT

Background: High-resolution soil moisture estimates are critical for planning water management and assessing environmental quality. In-situ measurements alone are too costly to support the spatial and temporal resolutions needed for water management. Recent efforts have combined calibration data with machine learning algorithms to fill the gap where high resolution moisture estimates are lacking at the field scale. This study aimed to provide calibrated soil moisture models and methodology for generating gridded estimates of soil moisture at multiple depths, according to user-defined temporal periods, spatial resolution and extent.

Methods: We applied nearly one million national library soil moisture records from over 100 sites, spanning the U.S. Midwest and West, to build Quantile Random Forest (QRF) calibration models. The QRF models were built on covariates including soil moisture estimates from North American Land Data Assimilation System (NLDAS), soil properties, climate variables, digital elevation models, and remote sensing-derived indices. We also explored an alternative approach that adopted a regionalized calibration dataset for the Western U.S. The broad-scale QRF models were independently validated according to sampling depths, land cover type, and observation period. We then explored the model performance improved with local samples used for spiking. Finally, the QRF models were applied to estimate soil moisture at the field scale where evaluation was carried out to check estimated temporal and spatial patterns.

Results: The broad-scale QRF model showed moderate performance ($R^2 = 0.53$, RMSE = 0.078 $m^3/m^3$) when data points from all depth layers (up to 100 cm) were considered for an independent validation. Elevation, NLDAS-derived moisture, soil properties, and sampling depth were ranked as the most important covariates. The best model performance was observed for forest and pasture sites ($R^2 > 0.5$; RMSE < 0.09 $m^3/m^3$), followed by grassland and cropland ($R^2 > 0.4$; RMSE < 0.11 $m^3/m^3$). Model performance decreased with sampling depths and was slightly lower during the winter months. Spiking the national QRF model with local samples improved model performance by reducing the RMSE to less than 0.05 $m^3/m^3$ for grassland sites. At the field scale, model estimates illustrated more accurate temporal...
trends for surface than subsurface soil layers. Model estimated spatial patterns need to be further improved and validated with management data.  

**Conclusions:** The model accuracy for top 0–20 cm soil depth (R$^2 > 0.5$, RMSE < 0.08 m$^3$/m$^3$) showed promise for adopting the methodology for soil moisture monitoring. The success of spiking the national model with local samples showed the need to collect multi-year high frequency (e.g., hourly) sensor-based field measurements to improve estimates of soil moisture for a longer time period. Future work should improve model performance for deeper depths with additional hydraulic properties and use of locally-selected calibration datasets.

Subjects  Agricultural Science, Soil Science  
Keywords  Soil moisture downscaling, Field scale, Soil climate analysis network (SCAN), U.S. Climate Reference Network (USCRN), North American Land Data Assimilation System (NLDAS), Digital soil mapping, Environmental covariates, Spiking, Grassland, Remote sensing

**INTRODUCTION**

The monitoring of soil moisture is vital for maintaining ecosystem services such as plant and soil health (Lowery et al., 1997; Jiang et al., 2021), erosion and pollution control (Jebellie, Prasher & Clemente, 1996; Wei, Zhang & Wang, 2007), biodiversity, and habitat protection (Sylvain & Wall, 2011). Because of the close association between soil moisture and ecosystem processes including C and nutrient cycling, water and solution movement, and soil biological activities, soil moisture is often used as a key model input to estimate crop and forage productivity, C sequestration potentials, greenhouse gas emissions, and environmental risks (Dominati, Patterson & Mackay, 2010; Grassini et al., 2015; Kuang et al., 2019; Udawatta & Gantzer, 2022). Accurate estimation of soil moisture is therefore considered necessary to support ecosystem assessments and management decisions.

Due to the high variability in weather conditions, soil hydraulic properties, and water movement (Bormann & Klaassen, 2008; Dari et al., 2019), dense spatiotemporal sampling is required to delineate the spatial variability and temporal dynamics of soil moisture when relying solely on *in-situ* measurements which include direct, lab-based gravimetric method, and indirect, field sensor-based methods that rely on neutron probes, electromagnetic techniques, cosmic-ray neutrons, optical reflectance, tensiometers, and thermal dissipation (Stafford, 1988; Walker, Willgoose & Kalma, 2004; Dorigo et al., 2011). Field and lab-based *in-situ* measurements can be costly and time-consuming, making it difficult to characterize soil moisture at higher resolutions needed to inform management. Consequently, soil moisture has been commonly estimated through geostatistical interpolation, proximal and remote sensing (RS), and empirical and process-based modeling (Robinson et al., 2008; Srivastava, 2017; Kathuria, Mohanty & Katzfuss, 2019; Babaeian et al., 2019). In particular, a covariate-based digital soil mapping (DSM) framework (McBratney, Mendonça Santos & Minasny, 2003), which considers soil properties to be a function of soil-forming factors, has been applied at various scales to empirically derive soil moisture estimates where machine learning (ML) models were built.
upon crucial soil, climate, topographic, and biotic covariates (Ahmad, Kalra & Stephen, 2010; Martínez-Murillo, Hueso-González & Ruiz-Sinoga, 2017; Zhang et al., 2020; Adab et al., 2020). Advancements in RS and increased availability of survey-based datasets has improved covariate datasets in recent years which would enable soil moisture to be empirically estimated at finer resolutions with higher accuracy.

Even though ML-based models are known for simplicity and predictive power, they are often criticized for a lack of transparency. It is well acknowledged that ML-based models should be thoroughly evaluated with independent validation datasets to ensure reasonable model selection and transferability (Lamichhane, Kumar & Wilson, 2019; Khaledian & Miller, 2020). Since models are usually only transferable within the inference space delineated by the calibration dataset (Miller et al., 2016), it is crucial to use large datasets spanning multiple climatic, soil, and vegetation gradients for model training in order to cover the data range needed for prediction. For soil moisture estimates within the contiguous U.S., national databases with wide spatial coverages (Table 1) such as Soil Climate Analysis Network (SCAN) (Schaefer, Cosh & Jackson, 2007) and U.S. Climate Reference Network (USCRN) (Diamond et al., 2013) that compiled a large number of in-situ measurements, present an important opportunity for model training and evaluation.

The national soil moisture datasets can serve as baseline maps for land managers but are usually low in spatial resolution (>10 km) and unevenly distributed in space, which requires further data interpolation and harmonization (Table 1). Previous efforts have also relied on interpreting soil moisture directly from RS products (e.g., NASA’s Soil Moisture Active-Passive (SMAP) Mission), which can cover a wide spatial range with relatively fine temporal resolution but is still lacking the finer spatial resolution details needed for decision making (Moran et al., 2000; Hunt et al., 2003; Velpuri, Senay & Morisette, 2016). Combining high-quality RS datasets, modeling techniques, and in-situ soil moisture measurements from national databases through data fusion can help generate high resolution gridded soil moisture products needed for agricultural monitoring (Peng et al., 2017; Cosh et al., 2021); however, land managers are still faced with the difficulty of retrieving and processing data derived from RS products without spending extensive resources on obtaining advanced technical support and expert knowledge. Soil moisture monitoring is particularly challenging for rangeland management considering that ranches commonly encompass a great number of land cover types such as grasses, hay, pasture, and shrubs which are associated with high heterogeneity in soil moisture dynamics. Some ranch properties even have proportions of areas in forest, bare ground, and crops. Under rapidly changing climate conditions (Derner et al., 2018), there is a growing need to rethink grazing management for U.S. rangelands to adapt to and potentially help mitigate climate change, and provide positive ecosystem services such as C sequestration and food security (Schuman, Janzen & Herrick, 2002; Havstad et al., 2007; Briske et al., 2015), where accurate soil moisture information can be critical for drought monitoring and productivity estimates (Richardson & Everitt, 1992; Sohrabi et al., 2015).

A number of soil moisture products/models developed with data fusion methods may satisfy the need of users in the U.S. even though none of these products are specifically
| Network | Multiple networks | No. sites | Spatial distribution | Temporal coverage | Temporal resolution | Depths | Sensor method | Quality control | Other measured covariates |
|---------|-------------------|-----------|---------------------|------------------|-------------------|--------|---------------|----------------|---------------------------|
| SCAN    | No                | >200      | Uneven and mostly from agricultural lands | Since the 90s     | Hourly            | 5, 10, 20, 51, 102 cm | Dielectric constant | Yes | ppt, soil T, SR, RH, WS and WD, air T, air P, soil pedon info |
| SNOTEL  | No                | >450      | Uneven and mostly from the western states | Since the 30s     | Hourly to daily   | 5, 10, 20, 50, 100 cm | Dielectric constant | Yes | ppt, soil T, SR, RH, WS and WD, air T, air P |
| NASMD   | Yes               | >1,800 in U.S. and Canada | Uneven but were harmonized to gridded scale | Vary by network   | Vary by network and harmonized to daily | Vary by network but harmonized to standard depths | Automated procedure established | Land cover, texture, BD, hydraulic conductivity, elevation |
| Ameriflux | No              | >400 sites | Uneven | Vary by site but mostly after 2000 | Mostly half-hourly | Vary by site | Vary by site | Vary by sites | LULC, ppt, soil T, VPD, SR, RH, WS and WD, air T, air P, C fluxes |
| USCRN   | No                | >100      | Relatively even     | Since 2000        | Hourly and daily  | 5, 10, 20, 50, 100 cm | Dielectric constant | Yes | ppt, soil T, SR, RH, WS, air T |
| COSMOS  | No                | >50       | Uneven             | Mostly after 2010 | Hourly           | Estimated as effective measurement depths | Cosmic-ray neutrons | Standard procedure established | SOC |
| ISMN    | Yes               | >1,400 globally | Uneven and more from the western states | Since the 50s     | Vary by network and harmonized to hourly | Vary by network | Automated procedure established | LULC, ppt, soil T, climate classification |

Notes:

- Ameriflux (Baldocchi et al., 2001; Law, 2005); COSMOS, COsmic-ray Soil Moisture Observing System (Zreda et al., 2012); ISMN, International Soil Moisture Network (Dorigo et al., 2011); NASMD, North American Soil Moisture Database (Quiring et al., 2016); SCAN, Soil Climate Analysis Network (Schaefer, Cosh & Jackson, 2007); SNOTEL, SNOwpack TELemetry (Schaefer & Paetzold, 2000); USCRN, U.S. Climate Reference Network (Diamond et al., 2013).

- Air P, air pressure; air T, air temperature; BD, bulk density; ppt, precipitation; LULC, land use land cover type; RH, relative humidity; SOC, soil organic carbon; soil T, soil temperature; SR, solar radiation; VPD, vapor pressure deficit; WD, wind direction; WS, wind speed.
designed for rangeland and therefore may lack parameters specifically calibrated for different land use/land cover (LULC) types in ranch properties (Table S1). Moreover, the majority of the RS-driven models were focused only on the top soil layer (0–5 cm) probably due to the strength of RS products in predicting surface soil moisture but not root zone moisture. However, the latter is of great importance to plant distribution and growth is usually under-emphasized (Schmura et al., 2010). In addition, the current moisture products have limited spatial resolution (≥1 km) and/or temporal coverage (only available after 2015 because of their model reliance on Sentinel or SMAP products) that is needed for decision making based on spatially-explicit information and long-term records. Even with the limitations, some of the studies (Table S1) which have provided associated data products or codes, such as Adaptive High-Resolution Soil Moisture Map (AHRSM) (Huang et al., 2020), PYthon Sentinel-1 Soil-Moisture Mapping Toolbox (PYSMM) (Greifeneder et al., 2021), and SMAP-Hydroblocks (SMAP-HB) (Vergopolan et al., 2021), are examples for a move towards developing automatic algorithms that are user-friendly and supported by background data processing. Built upon DSM and data fusion frameworks, this study aimed to provide models calibrated with the national databases to allow extraction of long-term and multi-depth soil moisture estimates at high spatial (e.g., 30 m) and temporal resolutions (daily). Model covariates were carefully selected to represent ecosystem processes and to capture the temporal range needed for soil moisture monitoring. We evaluated model performances in relation to calibration datasets and then provided an estimation of resources needed to generate soil moisture layers for site-based management.

MATERIALS AND METHODS

Extract in-situ soil moisture measurements from national databases

For broad-scale model calibration and validation datasets, we extracted in-situ soil moisture measurements from SCAN (Schaefer et al., 2007) and USCRN (Diamond et al., 2013) because they have similar measurement depths and sensor types which makes it easier for data harmonization (Table 1). In addition, the datasets have been quality-controlled and have spatial and temporal coverages that are suited for soil moisture modeling for the contiguous U.S. In this study, we provide soil moisture estimates for the Western or the Midwestern U.S. states (Reeves & Mitchell, 2012; Spangler et al., 2020) so only stations that fall within the defined spatial domain were retained (Fig. 1).

Data from the retained stations were further screened based on the following quality-control criteria adapted from the North American Soil Moisture Database (Liao et al., 2019): (1) records that did not report sampling depths were excluded; (2) records showing soil moisture below or equal to 0, or higher than 0.6 m/m³ were excluded; (3) records were joined with soil temperature records to exclude frozen soils. Since our study is interested in modeling longer term records of soil moisture, we further selected data measured between 2002 and 2019. We then retained records measured from the surface up to 100 cm to include deeper depth records. The retained records measured from 5, 10, 20, 50, and 100 cm depths were then processed to represent 993,771 site-depth-date
combinations from 120 stations. The major land use land cover (LULC) classes identified for the stations (Fig. 1) include shrubs (33%), grassland (23%), cropland (17%), and forest (17%) according to the National Land Cover Database (NLCD) data layer (Homer et al., 2004). The data extraction and processing steps were carried out in R (R Core Team, 2021). All codes for this and the following steps are provided in our GitHub repository (RCTM-soil-moisture) and are available through the link: https://github.com/xiayushu/RCTM-soil-moisture. Table S2 contains processing steps numbered to represent the corresponding scripts in the repository.

The coordinates corresponding to retained SCAN and USCRN stations were extracted and used as sampling points to create 90 m buffers for the subsequent extraction of Land surface model (LSM)-based soil moisture estimates and environmental covariates (Fig. 2).
Process LSM-based soil moisture estimates

We extracted LSM-based soil moisture estimates from the North American Land Data Assimilation System (NLDAS) database (Xia et al., 2012a) for the contiguous U.S. as the baseline soil moisture map for this study. The NLDAS hourly soil moisture is provided at a gridded scale (0.125 degrees) since 1979 that so far has the highest spatial resolution for the contiguous U.S. that is publicly available for the past 20 years. The NLDAS soil moisture is estimated with the Noah LSM which simulates soil freeze-thaw, heating-cooling, evaporation, transpiration, infiltration, and runoff processes based on model inputs including climate variables (precipitation, temperature, humidity, air pressure, wind speed, solar radiation) and surface parameters (vegetation type, green vegetation fraction, soil texture, roughness, albedo, slope factor) (Ek et al., 2003; Xia et al., 2012a). A list of URLs for available soil moisture estimates was downloaded from the EARTHDATA portal using the subset tools (Xia et al., 2012b). Then, Google Colaboratory (Google Inc., Menlo Park, CA, USA) was used to extract images from URLs corresponding to data from 2002 to 2019 before images representing three depth layers (0–10, 10–40, and 40–100 cm) were processed to daily averages and uploaded to the Google Earth Engine.

Figure 2 Flowchart showing the datasets and processes for model calibration, validation, and prediction. The national datasets used for model calibration and validation are Soil Climate Analysis Network (SCAN) and U.S. Climate Reference Network (USCRN). Covariates used in this study include North American Land Data Assimilation System (NLDAS), Land Surface Temperature (LST), Normalized Difference Wetness Index (NDWI), Enhanced Vegetation Index (EVI), Gross Primary Productivity (GPP), soil texture, bulk density (BD), soil organic carbon (SOC), precipitation, temperature (Temp), vapor pressure deficit (VPD), digital elevation model derived variables and indices, land use land cover (LULC), and tree cover percentage.
(GEE) (Gorelick et al., 2017) platform as an asset. Finally, point buffer-based averages of NLDAS-estimated soil moisture were extracted on a site-date-depth basis (Fig. 2), where we matched the median depth of the NLDAS layer with the sampling depth of observations from the national databases.

**Extract and process environmental covariates**

We selected covariates representing four classes including soil properties, climate conditions, biotic factors, and topography, extracted from digital elevation models (DEM) (Table 2), because of their known importance for soil moisture status (Verstraeten, Veroustraete & Feyen, 2008; Vereecken et al., 2014). Our selected covariates largely overlapped with previous work but encompassed a wider range since most of the previous work only utilized two or three covariate classes, with climate and biotic factors being the most common choices (Table S1). In addition, our framework differed from previous efforts in that we considered covariates that are available for at least 20 years (since 2002) so that long-term moisture records can be simulated using our model.

Soil covariates including soil texture (clay and sand contents), soil organic C (SOC), and bulk density (BD) were extracted from the SoilGrid+ product which maps soil properties at multiple depths from the contiguous U.S. at a 100 m resolution (Ramcharan et al., 2018). We retained the data layers representing standard depths of 5, 15, 30, 60, and 100 cm. In order to match the estimates with the in-situ soil moisture records, the first and second, second and third, and third and fourth depth layers were averaged to create predictions that approximate soil properties at depths of 10, 20, and 50 cm, respectively. The extraction for soil covariates was carried out on a site-depth basis (Table 2).

We extracted climate covariates, including daily precipitation, air temperature, and vapor deficit pressure (VPD) from the DAYMET v4 product which is available at a spatial resolution of 1 km (Thornton et al., 2020). Since the DAYMET data is generally not up-to-date on GEE, we provided optional, supplemental code for using the Parameter elevation Regression on Independent Slopes Model (PRISM) product (Daly et al., 2008) which has a coarser spatial resolution (4.6 km) but longer temporal coverage. Our code was developed to extract climate covariates, and make soil moisture predictions, according to site locations and dates specified by the user. Accordingly, using PRISM in the model instead of DAYMET is only needed for soil moisture predictions beyond 2021 (Table 2).

Biotic covariates consist of RS-derived indices and LULC classes. We extracted Land Surface Temperature (LST), Normalized Different Wetness Index (NDWI), and Enhanced Vegetation Index (EVI) on a site-date basis from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Huete, Didan & Leeuwen, 1999; Platnick et al., 2003) satellite imagery. The MODIS-derived Gross Primary Productivity (GPP) was available every 8 days at a spatial resolution of 250 m (Robinson et al., 2018); so the values were extracted based on data from the nearest available date (Table 2). We also used MODIS product (Townsend & DiMiceli, 2015) to extract tree cover % on a site-year basis. The LULC classes were extracted from the NLCD product (Homer et al., 2004) on a yearly basis where information from the closest year was used when data is not available for the specific year.

Xia et al. (2022), PeerJ, DOI 10.7717/peerj.14275
We extracted elevation, slope, and aspect from the Shuttle Radar Topography Mission (SRTM)-based DEMs. Further, the mean, vertical, and horizontal curvatures were processed from SRTM using the “TAGEE” package (Safanelli et al., 2020).

The topographic wetness index (TWI) was extracted from Hydrological data and maps based on Shuttle Elevation Derivatives at multiple Scales (HydroSHEDS) (Lehner, Verdin & Jarvis, 2008) and the surface roughness was determined using the Geomorpho90m.
The extraction for DEM-based topographic covariates was carried out on a site basis (Table 2).

The covariate extraction step was implemented with the High-Performance Computing (HPC) platforms (Table S2). Soil, LULC, and topographic covariates were extracted in GEE while biotic and climate covariates were extracted in Google Colaboratory to avoid computation time-out issues. Finally, separate covariate files were combined in R for subsequent quality control and model evaluation.

Quality-control and reduction of the calibration dataset
We screened the NLDAS-based estimates to exclude outliers (soil moisture estimates >0.6 m³/m³). GEE-based MODIS bitmasks were used to exclude pixels with lower quality caused by cloud cover, geometry problems, or emissivity issues. This step produced a dataset that only contains records associated with RS pixels labeled as “good quality”. This quality-controlled dataset contains 586,337 observations from 114 stations and will be referred to as the “full” dataset hereafter. We examined the Pearson correlation between the extracted modeling covariates and soil moisture observations for the full dataset (Table S3).

For LULC class, we used a rangeland-specific LULC product, the Rangeland Analysis Platform (RAP) (Jones et al., 2018) as an alternative modeling approach, to account for the land cover percentage of common vegetation types (annual herbs, perennial herbs, shrubs, trees, litter, and barren) in rangelands in addition to the NLCD product. In this case, the finer resolution tree cover% extracted from RAP (30 m) was used to replace MODIS-derived tree cover which is provided at a coarser resolution (250 m) (Table 2). The RAP layers are only available for the western U.S., so the quality-controlled dataset was further reduced. This step generated a reduced dataset with 500,119 observations from 94 stations which will be referred to as the “regionalized” dataset hereafter.

Model calibration and model validation
We first generated calibration datasets by randomly dividing the study sites into the calibration (70%) and validation (30%) sets. Since model performance can vary with sites used for independent validation, we carried out the random division five times to get different sets of calibration and validation sites for subsequent model building and evaluation. The Quantile Random Forest (QRF) models adapted from AHRSM (Huang et al., 2020) were built with our calibration datasets in R using the “caret” and “quantregForest” packages (Kuhn, 2008). The QRF model (Meinshausen, 2006) combines the concept of both quartile regression (Koenker & Bassett, 1978) and random forest (RF) (Breiman, 2001). It therefore has the advantages of tree-based models including producing higher accuracy estimates and being less sensitive to model overfitting (Belgiu & Drăguț, 2016; Hengl et al., 2018). However, QRF is different from the conventional RF model in that median instead of mean is used for point prediction. In particular, the QRF model gives non-parametric estimates for high-dimensional variables and is considered to be more consistent and less influenced by extreme values (Meinshausen, 2006; Vaysse & Lagacherie, 2017; Gyamerah, Ngare & Ikpe, 2020).
The QRF models for the full model were tested with different tree numbers (10, 15, 20, 25, 30, 35, 40, 45, and 50) and node sizes (5, 10, 15, and 20) to reduce overfitting, which is achieved by comparing model accuracy between the calibration and the validation dataset. The model was finally established using 15 trees and five nodes based on the 5-fold cross-validation results. To further examine the potential of model overfitting and the utility of different covariate classes, we also facilitated additional model runs using reduced covariate datasets as shown in Table S4.

Calibration models were built and validated for both the full and the regionalized datasets. Furthermore, separate models were built on subsets divided based on sampling depths, land use classes, and sampling periods (Table 3) to evaluate model efficiency associated with the use of different calibration datasets. Error metrics, including coefficient of determination ($R^2$) (Eq. (1)), root mean square error (RMSE) (Eq. (2)), mean bias error (MBE) (Eq. (3)), and Residual Prediction Deviation (RPD) (Eq. (4)) were derived from independent sites held out for validation. Covariate importance was reported based on QRF’s increase in node purity measure in the calibration models.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$  \hspace{1cm} (1)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$  \hspace{1cm} (2)

$$MBE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n}$$  \hspace{1cm} (3)

$$RPD = \frac{SD_{val}}{RMSEP \sqrt{n/(n-1)}}$$  \hspace{1cm} (4)

where $n$ represents the number of samples, $y_i$ represents observed value of sample $i$, $\hat{y}_i$ represents the model predicted value of sample $i$, $\bar{y}$ represents the mean of observations, $SD_{val}$ represents standard deviation of the validation dataset, and $RMSEP$ represents the RMSE of the validation dataset.

An alternative model calibration scheme was carried out to test if model performance can be improved with the spiking strategy that incorporates local samples for model building. Specifically, the calibration dataset consists of both data from the calibration sites from the national database and local data from 10%, 30%, 50%, 70%, and 90% of the study years at the sites that were held out for validation. Model performance was then evaluated using data from the years that were not used for calibration at each validation site. Summary statistics were calculated based on $R^2$ and RMSE from the validation sites associated with different spiking strategies and then compared with the model performance without spiking. The spiking test was carried out only on sites categorized under grassland according to the NLCD 2019 data layer, which accounts for about 25% of the full dataset (Table 3).
Estimate and evaluate field-scale soil moisture for Colorado sites

We used two sites in Colorado with available in-situ measurements as examples for generating daily, gap-free, high resolution, and multi-depth soil moisture layers. Although our codes were written in a way to allow user-defined area of interest (AOI) and spatial resolution, we set the AOI to our test sites and used a spatial resolution of 30 m to allow case study-based examination of modeling results and computational resources.

The first case study is the Cedaredge ranch, which has an area of 13.1 km² and is located in Western Colorado (Fig. 3). The central and the southwest portion of the ranch property are dominated by irrigated pasture, where there is a mixture of cool-season perennial grasses (tall fescue, smooth brome, and orchard grass), white and red clovers, and alfalfa.

---

Table 3 Summary statistics of the full and regionalized dataset divided by sampling depth, land cover classes, and month of the year.

| Classes       | Full dataset |                  |                  |                  | Regionalized dataset |                  |                  |
|---------------|--------------|------------------|------------------|------------------|----------------------|------------------|------------------|
|               | Obs. (Frequency) | Range (m³/m³) | Mean ± stdev (m³/m³) | Median (m³/m³) | Obs. (Frequency) | Range (m³/m³) | Mean ± stdev (m³/m³) | Median (m³/m³) |
| All           | 586,337 (100%) | 0.001–0.600     | 0.201 ± 0.119     | 0.181           | 500,119 (100%)     | 0.001–0.600     | 0.182 ± 0.109     | 0.163           |
| Sampling depth|              |                  |                  |                  |                      |                  |                  |
| 5 cm          | 124,009 (21%)  | 0.001–0.596     | 0.156 ± 0.106     | 0.136           | 106,422 (21%)      | 0.001–0.596     | 0.137 ± 0.096     | 0.118           |
| 10 cm         | 128,072 (22%)  | 0.001–0.597     | 0.192 ± 0.104     | 0.179           | 109,693 (22%)      | 0.001–0.597     | 0.177 ± 0.097     | 0.165           |
| 20 cm         | 117,817 (20%)  | 0.001–0.599     | 0.213 ± 0.110     | 0.201           | 99,836 (20%)       | 0.001–0.599     | 0.197 ± 0.103     | 0.185           |
| 50 cm         | 108,572 (19%)  | 0.001–0.600     | 0.225 ± 0.123     | 0.211           | 91,503 (18%)       | 0.001–0.600     | 0.207 ± 0.112     | 0.191           |
| 100 cm        | 107,777 (18%)  | 0.001–0.600     | 0.226 ± 0.136     | 0.188           | 92,665 (19%)       | 0.001–0.600     | 0.201 ± 0.122     | 0.169           |
| Land cover class |            |                  |                  |                  |                      |                  |                  |
| Grassland     | 145,166 (25%)  | 0.001–0.599     | 0.201 ± 0.108     | 0.188           | 145,166 (29%)      | 0.001–0.599     | 0.201 ± 0.108     | 0.201           |
| Cropland      | 133,636 (23%)  | 0.001–0.600     | 0.242 ± 0.126     | 0.229           | 105,640 (21%)      | 0.001–0.600     | 0.217 ± 0.119     | 0.205           |
| Forest        | 80,017 (14%)   | 0.001–0.585     | 0.229 ± 0.123     | 0.221           | 46,045 (9%)        | 0.001–0.577     | 0.181 ± 0.109     | 0.169           |
| Pasture       | 47,919 (8%)    | 0.001–0.600     | 0.264 ± 0.130     | 0.256           | 23,669 (5%)        | 0.001–0.600     | 0.238 ± 0.119     | 0.227           |
| Shrubs        | 177,816 (30%)  | 0.001–0.598     | 0.144 ± 0.085     | 0.129           | 177,816 (36%)      | 0.001–0.598     | 0.141 ± 0.085     | 0.129           |
| Month of the year |        |                  |                  |                  |                      |                  |                  |
| 1             | 20,324 (3.5%)  | 0.001–0.600     | 0.228 ± 0.125     | 0.207           | 16,576 (3%)        | 0.001–0.600     | 0.199 ± 0.113     | 0.180           |
| 2             | 19,449 (3%)    | 0.001–0.599     | 0.228 ± 0.124     | 0.211           | 16,154 (3%)        | 0.001–0.599     | 0.202 ± 0.111     | 0.182           |
| 3             | 33,708 (6%)    | 0.001–0.600     | 0.241 ± 0.121     | 0.236           | 28,531 (6%)        | 0.001–0.600     | 0.219 ± 0.111     | 0.213           |
| 4             | 44,622 (8%)    | 0.001–0.600     | 0.238 ± 0.121     | 0.235           | 37,687 (8%)        | 0.001–0.600     | 0.218 ± 0.112     | 0.213           |
| 5             | 51,472 (9%)    | 0.001–0.599     | 0.222 ± 0.120     | 0.209           | 44,106 (9%)        | 0.001–0.599     | 0.204 ± 0.111     | 0.191           |
| 6             | 70,984 (12%)   | 0.001–0.600     | 0.196 ± 0.117     | 0.176           | 62,588 (13%)       | 0.001–0.600     | 0.182 ± 0.109     | 0.163           |
| 7             | 74,441 (13%)   | 0.001–0.600     | 0.184 ± 0.113     | 0.163           | 64,702 (13%)       | 0.001–0.600     | 0.171 ± 0.105     | 0.153           |
| 8             | 77,699 (13%)   | 0.001–0.599     | 0.177 ± 0.112     | 0.156           | 67,583 (13%)       | 0.001–0.597     | 0.164 ± 0.104     | 0.147           |
| 9             | 77,274 (13%)   | 0.001–0.599     | 0.181 ± 0.114     | 0.159           | 65,618 (13%)       | 0.001–0.598     | 0.164 ± 0.104     | 0.146           |
| 10            | 59,761 (10%)   | 0.001–0.600     | 0.190 ± 0.113     | 0.170           | 50,701 (10%)       | 0.001–0.600     | 0.172 ± 0.104     | 0.153           |
| 11            | 36,496 (6%)    | 0.001–0.600     | 0.201 ± 0.117     | 0.183           | 29,635 (6%)        | 0.001–0.600     | 0.173 ± 0.104     | 0.153           |
| 12            | 20,107 (3.5%)  | 0.001–0.600     | 0.221 ± 0.126     | 0.202           | 16,238 (3%)        | 0.001–0.600     | 0.190 ± 0.113     | 0.169           |

Notes:
a The full dataset contains records from Midwestern and Western U.S. states.
b The regionalized dataset contains records from Western U.S. states. The dataset was created according to the spatial domain of the Rangeland Analysis Platform (RAP) dataset.

---

Xia et al. (2022), PeerJ, DOI 10.7717/peerj.14275
The northeast portion of the property is predominantly open canopy forest while the rest of the ranch consists of a mixture of grass, legumes (alfalfa, clover), and shrubs. For this case study, we simulated soil moisture at a median depth of 7.5 cm to represent the 0–15 cm depth layer, and at a median depth of 80 cm to represent the 60–100 cm depth layer. Soil moisture was simulated for the year 2021 in order to compare against measurements from field campaigns carried out during the growing season of the year which aimed to capture field scale spatial variability. During the sampling campaigns carried out in May (5/24–5/26) and July (7/20–7/21) in 2021, soil moisture was determined for 41 sampling locations across the irrigated portion of the site for the 0–15 cm depth (Fig. S1) using the gravimetric method (Eq. (5)) based on soil cores (D = 5.25 cm) collected with field core samplers. Model estimates of 0–15 cm depth were also generated for the year 2019 due to the slight difference in model input source between 2019 and 2021 caused by data availability (Table 2).

Gravimetric soil moisture (%) = \frac{\text{Weight of wet soil (g)} - \text{Weight of dry soil (g)}}{\text{Weight of dry soil (g)}} \times 100\% \quad (5)

We used the Central Plains Experimental Range (CPER), which has an area of 92.8 km² and is located in north-central Colorado, as the second case study (Fig. 3). The CPER site is dominated by rangeland plant communities including blue grama, plains prickly pear, sand dropseed, red three-awn, western wheatgrass, and four-wing saltbush (Steinert & Morgan, 2016). Since the site is included in the National Ecological Observatory Network (NEON) network, sensor-based soil moisture and temperature measures are available for
public use (Metzger et al., 2019). We downloaded data from the Ameriflux data portal (https://ameriflux.lbl.gov/sites/siteinfo/US-xCP) before processing them to daily resolution and excluding frozen soils in R. Five sensors were installed approximately 30 m apart at the CPER site (Fig. 3) with measurements mostly available from 2019 and 2020 at multiple depths (Table S5). However, the data was not continuous in time due to sensor calibration requirements. In addition, the sensors covered different depth layers. Therefore, we retained observations from three depth layers (7, 46, and 66 cm) that had the most overlap among the five sensors for further analysis. Temporal trends of soil moisture were examined for 2019 with sensor averages during the time period when continuous measurements were available, mostly from the growing season to winter, for each depth layer.

For both case studies, covariates including soil properties, topography, and NLCD-based LULC classes that are used for model building were extracted from GEE with a gap-filling algorithm being carried out to replace missing data using the nearest neighbors (Table S2). In Google Colaboratory, we generated both the NLDAS data layer and the layer containing climate and biotic covariates at a daily time step. For biotic covariates, MODIS-based bitmasks were imposed to mask out poor-quality pixels, which were then replaced with temporal averages. In the case where data is missing or unavailable on a daily basis, a record from the closest date was used for temporal gap-filling. This simple interpolation method on time-series data ensures fast extraction of covariates. However, more sophisticated spatio-temporal gap-filling approach using information of neighboring pixels and moving windows coupled with geostatistical regression may be used in the future to further improve the accuracy of data from the missing dates (Weiss et al., 2014; Siabi, Sanaeinejad & Ghahrama, 2022). The MODIS-based tree cover% and RAP-based LULC class% were extracted for the year of interest with the closest year being used when data of the specific year is unavailable. The layers containing extracted covariates were then combined and deposited to the Google Cloud Storage Bucket (GCS) (Google Inc., Menlo Park, CA, USA) to allow data import to an R virtual machine implemented with the Google Cloud Computing Platform (GCP) (Google Inc., Menlo Park, CA, USA). The final step was to apply the QRF model derived from the full dataset to predict soil moisture based on the defined date range, depth layers, and spatial resolution for the AOI. The modeled 30 m resolution soil moisture estimates at the two sites were examined for their spatial and temporal patterns at the defined depth layers.

RESULTS

Model performance influenced by calibration datasets and strategy

The model built upon the full dataset had a better fit with $R^2 = 0.53$, RMSE = 0.078 m$^3$/m$^3$, RPD = 1.43) compared to the regionalized dataset that only contained Western U.S. records ($R^2 = 0.41$, RMSE = 0.096 m$^3$/m$^3$, RPD = 1.30). The model performance was improved when only the top 0–5 cm layer soils were considered (full model: $R^2 = 0.58$, RPD = 1.53; regionalized model: $R^2 = 0.53$, RPD = 1.44) (Fig. 4). Model predicted soil moisture generally had narrower ranges than the observed values. Our model which encompassed a number of soil, climate, biotic, and topographic covariates as inputs
outperformed those built only on NLDAS-derived soil moisture (NLDAS-SM), which showed model fits of $R^2 = 0.31$ and $R^2 = 0.29$, respectively, using the full and regionalized datasets for all the depths (Fig. S2). The NLDAS-only model also had lower predictive power than ours for the 0–5 cm depth (full model: $R^2 = 0.43$; regionalized model: $R^2 = 0.42$). The models built upon reduced covariate datasets had performance varying from very low ($R^2 = 0.33$, RPD = 1.14) using only NLDAS-SM, sampling depth, LULC and remote sensing covariates, to close to the full model ($R^2 = 0.52$, RPD = 1.43) when the top 15 most influential covariates are retained (Table S4). Our comparison illustrated

Figure 4 Soil moisture model performance reported as error metrics according to the independent validation sites. The model performance were shown for (A) all depths and (B) surface 5 cm depth samples of the full dataset and for (C) all depths and (D) surface 5 cm depth samples of the regionalized dataset. The full dataset contains observations from the Midwestern and Western U.S. states while the regionalized dataset only contains observations from the Western U.S. states due to the use of covariates from the rangeland analysis platform data layers.
that estimates of NLDAS-SM can be improved by adding modeling covariates but the covariates may be selected based on their variable importance to reduce computational costs.

The model performance decreased greatly with sampling depth (Fig. 5A and Fig. S3A) and was the greatest for forest and pasture ($R^2 > 0.5$, $RPD > 1.3$), followed by grassland and cropland ($R^2 > 0.4$, $RPD > 1.25$), and was the lowest for shrublands ($R^2 = 0.25$, $RPD = 1.2$). Slight decrease in model performance ($R^2$ between 0.4 and 0.45) was suggested for the winter months compared to the growing season ($R^2 > 0.5$). This might be because there is generally a 60% decrease in data availability in the calibration dataset during winter than the rest of the year (Table 3). However, model performance evaluated for the winter months is associated with a larger variance than those reported for the growing season (Fig. 5C).
The model performance represented by $R^2$ had a range of 0.05–0.61 with a mean of $R^2 = 0.36$ for grassland sites used for independent validation (Fig. 6A). Using data measured from 10% of the years to spike the national model and the remaining 90% of the years for validation, model performance was improved by 37% on average. Using 30% of the years for spiking and 70% of the years for validation further improved model predictive power where an averaged performance was calculated to be $R^2 = 0.53$ with a range of 0.24 to 0.78. In addition to a more than 50% increase in model $R^2$, this spiking strategy decreased model RMSE to 0.053 m$^3$/m$^3$, which is approximately 40% lower compared to not using any spiking strategies (0.092 m$^3$/m$^3$; Fig. 6B). Model performance (averaged $R^2 = 0.56$ and RMSE = 0.051 m$^3$/m$^3$) was slightly improved by dividing the dataset evenly into spiking and validation sets for the study years but did not improve further when more than 50% of the samples were used for spiking (Fig. 6).

**Variable importance of calibration models**

Both the full and the regional models ranked soil texture (sand and clay contents), BD, SOC, NLDAS-SM, sampling depth, and elevation as the most influential covariates (Figs. 7A and 7B). This is consistent with the correlations computed between observed soil moisture and modeling covariates showing the strongest predictors as NLDAS-SM, soil properties, and elevation (Table S3). LULC-based covariates were in general less influential than soil covariates but more influential than RS-derived indices. Compared to the model built on the full dataset, the model built on the regionalized dataset ranked NLCD-derived LULC with slightly lower importance (Fig. 7B) probably because that...
information overlapped with the RAP-derived LULC class%. However, ranking between NLCD- and RAP-derived LULC vary among class-based models shown in Fig. S4 and this might be associated with certain LULC classes being more explanatory of the moisture variance within the category. The climate covariates generally ranked in the last one thirds in terms of their importance in building the QRF models regardless of their high temporal variability.

Elevation and NLDAS-SM were determined to be most influential for moisture model built on soils from the surface sampling depth (5 cm) (Fig. 8A). Besides NLDAS-SM and elevation, soil texture was also determined to have a predominant influence on soil moisture estimated from the deeper depth (100 cm) (Fig. 8B). Interestingly, sand% was ranked with higher importance than clay% in the 5 cm model while clay% was ranked as the most important covariate for the 100 cm model. This might illustrate the difference in water drainage and storage as controlling factors at different soil depths. The RS-derived variables such as LST and GPP, together with RAP-based estimates of LULC class percentages, were ranked with higher importance for surface compared to the deeper depth soil moisture models. When restricting the training data to the regionalized dataset, NLCD-derived general LULC classes were more influential than RAP classes for the deeper depth (Fig. S4B). Topographical covariates which reflect spatial rather than temporal differences were generally estimated to have higher importance in the deeper depth soil moisture modeling.

Figure 7 Variable importance for (A) the full dataset and (B) the regionalized dataset ranked according to the increase in node purity of the Quantile Random Forest model. The full dataset contains observations from the Midwestern and Western U.S. states while the regionalized dataset only contains observations from the Western U.S. states due to the use of covariates from the rangeland analysis platform (RAP) data layers. The covariates include soil sand (Sand) and clay (Clay) contents, soil bulk density (BD), soil organic carbon (SOC), sampling depth (depth), NLDAS-derived soil moisture (SM), precipitation (ppt), temperature (T), vapor pressure deficit (VPD), Gross Primary Productivity (GPP), Enhanced Vegetation Index (EVI), Modis-based tree cover% (Tree%), land use land cover (LULC), Land Surface Temperature (LST), Normalized Difference wetness Index (NDWI), elevation (EL), slope (SL), aspect (AS), mean (mcurv), vertical (vcurv), and horizontal (hcurv) curvatures, Topographic wetness Index (TWI), surface roughness (SRG), and RAP-based estimates of annual herbs% (AFGC%), perennial herbs (PFGC%), bare ground (BG%), tree (TREE%), litter (LTR%), and shrub (SHB%) covers. Model performance was presented in the panels for comparison.

Xia et al. (2022), PeerJ, DOI 10.7717/peerj.14275
Variable importance was comparable for models built on soils from grasslands (Fig. 8C) and croplands (Fig. 8D) with soil properties, NLDAS-SM, and elevation ranked the highest, while climate variables were considered to be less influential. When the model is restricted to the grassland category which is most relevant to ranch managers, the most influential RAP estimates were perennial herbaceous cover (Fig. S4C). Both models built for January (winter) and July (summer) identified soil properties, NLDAS-SM, elevation, and tree cover to be the most influential covariates while precipitation was ranked with low importance (Figs. 8E and 8F). Sampling depth was considered to have higher importance in summer than winter which is associated with the seasonal variation of soil moisture of different depth layers.
Model application for case studies: time and resources needed for implementation

Model predictions for daily soil moisture were made at a spatial resolution of 30 m for the Cedaredge ranch and the CPER site and can be flexibly adapted by future users for their AOI and defined time frame using the following steps: (1) extraction of temporally constant soil and topographic covariates, (2) extraction of temporally dynamic climate and biotic covariates, (3) extraction of NLDAS soil moisture layers, (4) generate estimates, and (5) quality checking. The total time required for carrying out all steps for the Cedaredge ranch was about 5 h (Table S6) using a standard laptop computing machine with eight cores and 32 GB RAM. The computation time can be shortened if multiple steps, such as (2) and (3) were carried out simultaneously or if a virtual machine with a higher number of cores was set up to speed up the most time-consuming step (4). The total storage required for completing the task was about 150 MB. For the CPER site which is about 8 times larger than the Cedaredge site, the computation time was about twice (9.5 h) and the storage requirement tripled.

Evaluation of model estimates for case studies

The model estimated soil moisture, when averaged at the site level, was much more temporally dynamic for the top 0–15 cm compared to the deeper layer (60–100 cm), especially during the growing season for the Cedaredge ranch (Fig. 9A). The temporal trends were modeled to be more similar between the top and deeper soil depth layers for CPER (Fig. 9B) that is likely associated with differences in soil properties at these two sites. The spikes of top layer soil moisture typically matched with site-level precipitation patterns and when irrigation occurs during the growing season. Winter soil moisture was less dynamic and more representative of saturated levels. The averaged soil moisture level was modeled to be higher during the growing season in the deeper depth than the surface depth for both sites. Temporal dynamics of soil moisture were comparable at the site level regardless of the calibration datasets being used (Fig. S5). Spatially, the correlation between model estimated top layer (0–15 cm) soil moisture and that measured through the gravimetric method at the Cedaredge ranch during the growing season was significant but only moderate (R = 0.4, P < 0.01), with the measured soil moisture at the higher end (>20 m^3/m^3) being underestimated (Fig. S1). This might be due to the fact that soil moisture samples were taken across the irrigated portion of the ranch, but moisture level influenced by irrigation and grazing activities are mostly reflective from the RS-based inputs in the model that are only ranked with medium importance in the model (Fig. 7A). Moreover, the model estimates generated at the 30 m resolution would have a scale mismatch with the field samples taken at the point scale.

A wide range of correlations between modeled and measured temporal soil moisture (R between 0.28 and 0.57) with a mean of R = 0.42 (Table S5) was observed from the five sensors installed in CPER, which is independent of the broad-scale calibration dataset. Temporal validation at the site showed R = 0.64 for the top depth (7 cm) when sensor measurements were averaged (Fig. 10A). Noticeable mismatches in magnitude were observed in early July and mid-September where model estimates were lower than
measured. Since the sudden spikes in measured soil moisture did not seem to be solely associated with precipitation events, they might be caused by management factors such as irrigation that are difficult to capture with the current model input. The correlation was observed to be moderate for the medium depth (46 cm: $R = 0.87$) and lowest for the deep depth (66 cm: $R = 0.37$). However, it should be mentioned that the magnitude of soil moisture was estimated to be consistently lower with model prediction compared to field measures. Moreover, sensor-based variation was observed to be much wider than modeled at deeper soil depths (Figs. 10B and 10C) which reflects the difference in footprint size for sensor measurements and model estimates.

Even though NLDAS-SM, which is used as the underlying dataset for downscaling the results, only consisted of two pixels for the extent of both of the case studies and therefore could not mirror spatially-explicit moisture dynamics (Figs. S6 and S7), the spatial pattern for model estimated soil moisture was able to illustrate much more spatially-refined patterns including the summer drying period for the irrigated pasture area of the Cedaredge ranch (Fig. S6). For the CPER site, the underlying NLDAS-SM caused a sudden change in moisture estimates within neighboring pixels (Fig. S7).

**DISCUSSION**

**Depth-based model performance compared with previous work**

Previous work found RS-based soil moisture estimates to be moderately correlated with field-measured soil moisture for the surface but not deeper depth layers (*Velpuri, Senay & Morisette, 2016*). This aligned with validation results by depth shown in Fig. 5A. Our model performed similarly or slightly better ($R^2$ between 0.5 and 0.6, RMSE between 0.5 and 0.6) compared to previous efforts.
0.07 and 0.075 m$^3$/m$^3$) than previous data fusion-based 0–5 cm soil moisture estimates that used independent sites as the validation set. The 12 studies that adopted independent validation scheme summarized in Table S1 reported a model fit (R$^2$) between 0.22 and 0.66, with 58% of the studies reporting R$^2 = 0.5$ for the best-fitted model and 33% of the studies reporting an R$^2$ between 0.5 and 0.6 like ours. The RMSE reported from these studies generally fell within the range of 6–8% which is comparable to our study. Jing, Zhang & Zhao (2018) achieved better performance (R$^2 = 0.66$) but predicted soil moisture at coarser spatial (0.05 m$^3$/m$^3$) and temporal (monthly) resolutions. The other studies which outperformed ours used alternative model calibration and validation strategies (Table S1). For example, random division of calibration and validation datasets was adopted by studies that reported more promising model fits (R$^2 > 0.8$) (Liu et al., 2020; Greifeneder, Notarnicola & Wagner, 2021; Zhang et al., 2021). It should be noted that validation against randomly selected data points, even if stratified by study years, can generate much better model fits compared to independent sites because of the use of autocorrelated time-series data for both model training and validation (Meyer et al., 2018; Ploton et al., 2020). Our model fit for the validation dataset was calculated to be R$^2 > 0.95$ (data not shown) if the random division method were to be adopted. However, such model validation schemes cannot guarantee model performance against independent sites which may differ significantly from the calibration dataset.

Karthikeyan & Mishra (2021) reported model performance for soil moisture estimated up to 100 cm and their conclusions are akin to ours showing a drastic decrease in model performance for the 50 and 100 cm depths (R$^2 < 0.3$) compared to surface layers.
Our model outperformed theirs slightly but still had relatively poor fits for deeper depth layers ($R^2 = 0.29$ for 100 cm and $R^2 = 0.40$ for 50 cm, Fig. 5A) which cautions the use of soil moisture estimates at the root zone relying solely on the ML-based method. Deeper depth soil moisture tends to vary more by soil hydraulic properties, soil texture, and LULC compared to surface layers (Xiao et al., 2014; Karandish & Šimůnek, 2016; Das & Mohanty, 2006; Li et al., 2019). Even though our model did include LULC, LSM-derived estimates, and soil properties as covariates which were ranked with moderate to high importance for deeper soil layers (Fig. 7B), soil hydraulic properties were not directly included except that they could be inferred from soil texture, BD, and SOC (Saxton & Rawls, 2006). In the future, soil hydraulic maps (Hengl & Gupta, 2019) may be used to improve the modeling of soil moisture at deeper depths. Additionally, more adequate delineation of soil moisture variability of deeper depths may rely on techniques such as vadose zone and root growth modeling (Das & Mohanty, 2006; Yadav, Mathur & Siebel, 2009) that can better account for mechanisms associated with soil water flow and plant water uptake.

**Model performance influenced by land cover and season**

The model predictive powers for soil moisture were moderate for grassland, cropland, forest, and pasture ($R^2 > 0.4$, Fig. 5B). By considering all depth layers, forest had the best prediction ($R^2 = 0.56$, RMSE = 0.068 m$^3$/m$^3$) among the LULC types in our work, which is at odds with the study of Tavakol et al. (2019) that reported the highest error in estimating forest soil moisture using RS-based method because of the difficulty posed by a higher intensity of vegetation. The reason for this might be that NLDAS-SM correlated moderately ($R = 0.65$, Table S3) with observed soil moisture for forest and that our model incorporated soil and climate information in addition to the RS-based indices that are more sensitive to canopy covers. Moreover, the RMSE showed a very high variation among model runs using different sets of independent validation sites, meaning that model accuracy can vary greatly with site factors for forest prediction (Fig. 5B).

The model fit was better for grassland ($R^2 = 0.45$, RMSE = 0.088 m$^3$/m$^3$) than cropland ($R^2 = 0.4$, RMSE = 0.103 m$^3$/m$^3$) in our study, which is in line with the finding that SMAP-downscaled products generated lower RMSE for grassland due to its homogeneity compared to cropland (Wu et al., 2015). Unfortunately, model performance was relatively poor for shrublands and this could explain that the regionalized model using RAP covariates did not outperform the full model probably because of the higher proportion of sites being categorized under shrubs (36%) for the Western U.S. (Table 3). It could be that the average soil moisture contents are much lower in shrublands than in others and so the variation within a smaller data range was harder to capture. However, mixed results were found in other studies comparing model performance of shrublands with other LULC types (Liu et al., 2020; Senyurek et al., 2020; Fuentes, Padarian & Vervoort, 2022). The difference in soil moisture dynamics tied to LULC types could be explained by vegetation-associated evaporation and precipitation infiltration, soil BD and rooting depth, and initial soil moisture contents (Niu, Musa & Liu, 2015). Building models to
better refine soil moisture estimates for diversified systems under various LULC types may therefore consider incorporating additional soil and biotic factors in the future.

Our validation showed slightly better model performance during the growing season than winter months (Fig. 5C). The greatest proportion of training data was taken from June to September (Table 3), which coincided with the slightly better agreement between the observed and modeled soil moisture. The model performance was lower ($R^2 \leq 0.4$) with high variance between model runs in January and February, but RMSE was also low ($<0.045 \text{ m}^3/\text{m}^3$) because of the lower temporal variation in soil moisture during this time period. Even though RS-based indices were found to have good correlations with soil moisture ($R$ ranged between 0.5 and 0.7) for a winter wheat field (Ren et al., 2022), our model did not emphasize spectral indices as influential covariates (Fig. 8E) probably due to the complexity of LULC and depth classes involved in the training dataset. Winter soil moisture may be better simulated with process-based models that can account for effective pore spaces, heat exchange, and evaporation during the freeze and thaw processes (Kahimba, Ranjan & Mann, 2009; Banimahd & Zand-Parsa, 2013).

Covariate importance for empirical estimates of soil moisture
Soil properties including soil texture, BD, and SOC were consistently identified to be among the most influential covariates for building the ML-based soil moisture models (Figs. 7 and 8). This is not surprising because soil properties, especially soil texture and pore size distribution, are highly influential on soil hydraulic properties (Saxton & Rawls, 2006; Strudley, Green & Ascough, 2008; Cichota et al., 2013; Taghizadehghasab, Safadoust & Mosaddeghi, 2021), which in turn determine lateral and horizontal water movements. On one hand, process-based models such as Century (Parton & Rasmussen, 1994) and EPIC (Izaurralde et al., 2006) have commonly employed pedo-transfer functions to estimate water holding capacity which is then used to estimate soil moisture dynamics through the simple bucket model. On the other hand, the empirical method which is being increasingly used to model or downscale soil moisture has largely relied on DEMs and RS products (Table S1). Recent advances in DSMs and global soil monitoring networks (e.g., GlobalSoilMap) (Hartemink et al., 2010; Chen et al., 2022) should enable the use of gridded, quality-controlled products of soil properties as inputs for both process-based and empirical models.

The NLDAS-derived estimates of soil moisture and soil sampling depth were also ranked with very high importance for all the models (Figs. 7 and 8). Based on variable importance, it might be suggested that the NLDAS data product would be suited for studies carried out at the broad scale where fine-scale soil moisture pattern is not required to drive the modeling of spatial dynamics. This notion is supported by the fact that the direct correlation between NLDAS and measured soil moisture is moderate ($R = 0.5$) for all layers and good ($R = 0.7$) for the top 0–5 cm layer (Table S3). Soil sampling depth could be used either as a covariate like in this study or as a class factor to develop separate models. Considering that models built for surface and deeper depth layer soil moisture estimates had relatively different rankings in covariate importance (Figs. 8A and 8B),
especially with regards to RS-derived covariates, it may be advisable to build separate models for different depths layers by selecting customized sets of covariates.

Elevation was ranked with high importance for all models (Figs. 7 and 8), which mainly reflects site-based differences. In contrast, other DEM-derived covariates such as aspect and curvatures had lower rankings in many instances. When extracting soil moisture data points from national databases, only one coordinate is provided for each site, meaning that field-level spatial variation could not be accounted for. This could explain the discrepancy that topographic covariates are often identified with high importance for field-scale soil moisture estimates (Zhu & Lin, 2011; Lee & Kim, 2019) but not in this study. One way to compensate for the limitation of applying a broad-scale model to the fine spatial scale where there is a lack of topographic explanation on soil moisture is to spike the broad scale model with higher resolution soil moisture maps using topographic covariates. These maps can be derived from field sampling campaigns or well-validated, RS-based data fusion products that focus on delineating spatial rather than temporal soil moisture dynamics. However, it remains to be tested whether the high-resolution maps need to be highly representative of the condition that is similar to the target site in order to make the model prediction successful.

The RS-derived indices exhibit both spatial and temporal variations but were ranked only with medium importance when all records are considered (Fig. 7B). Furthermore, RS indices had relatively weak correlation with observed soil moisture (Table S3), which, combined with our finding that model excluding RS covariates performed similarly to the full model (Table S4), suggests that future modeling work might retain only the most influential RS covariates to reduce computation time and risks of model fitting. It could be that some of the variance provided by RS indices overlapped with those from NLDAS-SM and site-based soil properties. Nevertheless, modeling of surface soil moisture relied more on GPP and LST (Fig. 8A), which makes sense as GPP is closely tied to plant productivity that is known to be influenced by soil moisture conditions (He et al., 2016; Fu et al., 2020) and LST derived from the thermal bands informs soil moisture status through its association with energy balance and hydrology (Sandholt, Rasmussen & Andersen, 2002; Cammalleri & Vogt, 2015). For deeper depth, soil moisture appeared to be best correlated with GPP and tree cover among the RS indices (Table S4), which demonstrated the association between soil moisture and belowground biomass that can be inferred from surface indices. The RAP and NLCD-based LULC covariates were also frequently ranked with medium importance (Fig. 8 and Fig. S4). RAP provided more refined temporal and class percentage information compared to NLCD and so was ranked slightly higher than NLCD in many cases. However, the RAP product is especially oriented toward the rangeland ecosystem and does not include detailed information or threshold values for croplands or different types of forests. It may therefore be advisable to build soil moisture models with combined use of NLCD and RAP for areas with more mixed LULC types.

Climate covariates that reflect both site-based and temporal differences were generally ranked with low importance (Figs. 7 and 8). In view of the feedback between precipitation and soil moisture, and the fact that the former is a main driver of the latter (Salvucci,
it was a bit surprising that precipitation was consistently ranked with lower importance and had poor correlations with observed soil moisture regardless of LULC and time of the year (Table S4). The cause for this might be that there is usually a time lag between rainfall events and changes in soil moisture, in which case the anomalies of soil moisture and precipitation might be better correlated, and an antecedent precipitation index may be used instead of daily precipitation (Wagner, 2003; Koster et al., 2004; Schoener & Stone, 2020). Another explanation is that the climate data have relatively coarse spatial resolution compared to other covariate datasets (Table 2), which would more likely cause a spatial mismatch between the measured data point and the value extracted from the coarse-resolution pixel.

**Model spiking and evaluation of soil moisture estimates for grassland sites**

As illustrated previously, the drastic difference between model performance caused by the validation scheme (random sample selection vs independent sites) demonstrated relatively poor transferability of the soil moisture model from the broad to the fine spatial scale. The spiking test which was meant to alleviate the scale mismatch for model training showed the need to incorporate at least 30% of the local measurements into the national model in order to reduce RMSE to be lower than 0.05 m$^3$/m$^3$ (Fig. 6B), which means that site-based soil moisture dynamics might need to be monitored for three years within a ten-year time window for the grassland sites. That being said, increasing spiking samples to more than half of the study years did not further improve model performance (Figs. 6A and 6B), indicating potential overfitting of temporal soil moisture dynamics that can vary year by year.

The soil moisture estimated for the Cedaredge ranch showed reasonable temporal dynamics (Fig. 9) with winter moisture levels estimated to be close to field capacity due to snowmelt and reduced evapotranspiration, which is followed by a drying period during the growing season especially at the surface depth (Moran et al., 2000; James et al., 2003). The soil moisture estimated for the CPER site demonstrated reasonable temporal matches with sensor measurements for the top 0–15 cm depth in terms of magnitude and direction of change (R = 0.64, Fig. 10). However, some of the spikes observed by the sensors, especially those with >20% change within a week, seem to have been missed or reduced in model simulations (Fig. 10A). The spikes associated with precipitation were better captured which makes sense as climate data was incorporated into the model. Other spikes caused by management practices such as irrigation and grazing are likely harder to model and are mostly reflected by RS-derived indices which are sometimes subject to temporal mismatches and data quality issues. One way to improve the modeling of soil moisture responses to management would be to set up controlled experiments with contrasting management practices in order to explore the explanatory covariates for soil moisture dynamics within a short period of time. The sensor readings also indicated higher temporal and spatial variations for deeper depth layers which were not mirrored by the modeling results (Figs. 10B and 10C). The modeled temporal changes in deeper depth soil moisture may be enhanced by adding covariates reflecting soil hydraulic properties.
and management. Since the sensors were relatively close to each other in the CPER site (Fig. 3B), a lack of difference among sensors from the model simulation is likely related to the input data not being able to represent refined spatial patterns.

The correlation between modeled and measured soil moisture at 0–15 cm depth during the growing season was found to be only moderate (R = 0.4, RMSE = 0.036 m³/m³) when data from one sampling campaign focusing on spatial variability was used (Fig. S1). The discrepancy is associated with a narrower range of soil moisture predicted compared to captured, which likely reflects the difference in modeled resolution (30 m) and point-scale measurements together with the difficulty to track irrigation-associated change in soil moisture. Additionally, some model inputs such as soil and climate data as well as RS-indices have coarser resolution (≥100 m) which can lead to a reduced modeling power at fine resolution prediction. Although modeled soil moisture was able to delineate summer drying clustered within the irrigated area in the Cedaredge ranch compared to the northeastern open canopy area which had smaller moisture change from winter to summer, the contrast between LULC was likely exaggerated in winter (Fig. S6).

One explanation might be that the winter calibration data was relatively small (Table 3) which made it harder to build more robust and transferable predictions. Another explanation might be associated with the fact that the soil moisture estimates were strongly influenced by the underlying soil maps as evidenced by covariate ranking. Spatial pattern simulated at the CPER site reflected similar patches of high and low soil moisture estimates for summer and winter but was much more restricted to the tile difference of NLDAS-SM (Fig. S7). In this case, interpretation based on field averages or relative differences among pixels within the same NLDAS tile might be more reliable than the interpretation of soil moisture magnitude at the whole site based on modeled spatial patterns. The relatively consistent magnitude of standard deviation in soil moisture estimated from the whole ranch (Fig. 9) reflected the notion of temporal persistence of soil moisture changes in space (Zucco et al., 2014).

Implications, limitations, and next steps

Our work provided a protocol for building ML models at the broad scale before applying them to the AOI with user-defined resolutions (Fig. 2). Even though building a national model using long-term soil moisture records can be time-consuming given the size of the calibration dataset, the actual computation time needed to generate site-based, multi-depth, 30 m, daily soil moisture layers for our case study took less than 10 h for a ~100 km² ranch by utilizing HPC (Table S6), which makes it realistic for users with basic model training to obtain field-scale moisture estimates needed for monitoring purposes or as model inputs. We made our model and codes open source (Table S2) so that users may choose to apply or adapt our model to make predictions for their AOI at the defined time and depths. Additional covariates or those from alternative data sources can be used to rebuild models like ours.

Several alternative soil moisture products derived from LSM and/or RS data, which were used by previous studies (Table S1), such as the Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010), Global Land Data Assimilation System (GLDAS) (Rodell et al,
and Climate Change Initiative (CCI) (Dorigo et al., 2017), may be used as an alternative to NLDAS for model training. In our study, we selected NLDAS as a baseline soil moisture product for downscaling (i.e., used as a covariate) because it has a much wider temporal coverage (since 1979) compared to SMAP (since 2015), and is generated at finer spatial resolution (0.125°) than GLDAS (1°) or CCI (0.25°). Furthermore, although GLDAS and USDA-enhanced SMAP are deposited in GEE and more readily available for HPC users, the NLDAS data layers can be downloaded and processed in HPC to allow both broad-scale and fine-scale data extraction (Table S2). The various soil moisture products have been thoroughly validated against in-situ measurements and compared against each other (Ford & Quiring, 2019; Deng et al., 2020; Jiang et al., 2020). However, future work is needed to generate long-term, fine spatial resolution soil moisture products that would be beneficial to users focusing on field-scale research or management. Currently, users would have to choose between using publicly available but coarser resolution soil moisture datasets or developing their own products which can be costly and time-consuming.

In addition to using alternative soil moisture products as baseline maps, covariates that were not included in this study should be examined in terms of their potential to improve model performance and transferability. For example, the shortwave infrared band (SWIR) which is available from most remote sensors is proposed to be useful for improving soil moisture estimates (Sadeghi et al., 2017). A number of RS-derived indices, including Leaf Area Index (LAI), temperature vegetation condition index (TVDI), and crop water stress index (CWSI) (Patel et al., 2009; Sánchez et al., 2012; Chen, Willgoose & Saco, 2015; Akuraju, Ryu & George, 2021), have also been reported to be suitable covariates for soil moisture modeling. Other commonly used covariates for soil moisture modeling include albedo, solar radiation, potential evapotranspiration, and soil hydraulic properties (Table S1). A model selection procedure may be necessary to ensure model efficiency if model training involves a large number of covariates that are correlated.

The next step is to validate our modeling results at additional sites from the U.S. Western states. The ML-based model can perform reasonably well and is specifically useful for data-poor areas where field-measured parameters such as soil hydraulic properties needed to run process-based models are missing (Carranza et al., 2021). Nonetheless, the ML-based method is limited in transferability. Specifically, the controlling factors of soil moisture at the broad scale (e.g., national level) usually differ from those at the fine-scale (e.g., site level). One way to reduce the potential model overfitting or mismatch at the field level using broad-scale data for training is to adopt a localized calibration method by selecting the most representative calibration samples at the pixel level; however, that procedure would be extremely time-consuming and computationally intensive. A less costly approach would be to select appropriate calibration samples for each site from the national library and spike those with local sensor data. The spatial variation of estimated soil moisture at the field level using the full or a selected national library data combined with different number of local sensor measurements should therefore be compared and verified against field observations. Ideally, samples can be collected from different areas of the field to validate spatial patterns. However, such high spatial resolution
validation may only be practical for several sampling campaigns within a year unless a substantial sensor network is established at the site level.

Field-based calibration and validation would be necessary when field characteristics do not fall within the domain of the calibration datasets and therefore cannot be estimated accurately based on observations from the national databases alone. The spiking method incorporating site-based temporal data can improve model transferability but would require site-based measurements. Using sensors installed at a lower spatial resolution, field-scale model calibration and validation can be carried out temporally such that in-situ measurements of soil moisture from different time periods can be used for model spiking and validation. In this case, it would be critical to select the optimal spiking dataset (Nawar & Mouazen, 2017) needed to achieve a balance between model accuracy and measurement costs. We generated spiking samples based on sampling years as suggested by Abowarda et al. (2021), but future work should investigate the use of data at a coarser temporal resolution or only select those that are representative of the seasons of interest in order to reduce monitoring costs.

**CONCLUSIONS**

This work provided a transparent method with open-source codes to generate daily, high-resolution soil moisture estimates by building machine learning models based on national soil moisture databases (SCAN and USCRN), LSM-derived coarse resolution soil moisture estimates (NLDAS-SM), and environmental covariates including soil properties, climate variables, DEM and RS-derived indices, and LULC-based classes. Among the covariates investigated, NLDAS-SM, soil properties, and DEM-derived indices were identified to be most influential.

The model performed similarly or better to other studies carried out for the contiguous U.S. using independent sites for validation, and greatly outperformed those using NLDAS-SM alone. Our model performance ($R^2 \geq 0.5$, RMSE $< 0.078$ m$^3$/m$^3$, RPD $> 1.4$) at 5 and 10 cm illustrated that the data fusion-based model is effective at simulating surface soil moisture. Field-scale validation showed acceptable matches between observed and modeled soil moisture both temporally ($R = 0.64$) and spatially ($R = 0.4$) at the top 0–15 cm depth layer. It was confirmed that the machine learning-based empirical model is relatively weak at predicting soil moisture from the 50 and 100 cm sampling depths, which calls for the use of additional hydraulic covariates or coupling of mechanistic models for simulating root zone soil moisture. Our study obtained moderate agreement between observed and simulated soil moisture for grassland and cropland but was much weaker at predicting moisture from shrublands which likely contributes to the reduced model performance using a regionalized dataset from the Western U.S. states. Winter soil moisture was predicted with less accuracy caused by a lack of training data and the challenge to account for the freeze-thaw events.

The relatively poor model transferability evidenced by the difference in model performance between random and independent validation emphasized the need to use
techniques to reduce mismatch associated with applying broad-scale calibration models to the fine-scale. Spiking the national model with local samples significantly improved the model’s ability to predict temporal soil moisture patterns ($R^2 > 0.6$ and RMSE $< 0.05 \text{ m}^3/\text{m}^3$) but model accuracy for predicting spatial soil moisture at the site level needs to be further validated and improved with optimized calibration datasets.

**ACKNOWLEDGEMENTS**

We would like to acknowledge Dr. John Kimball for providing insight and expertise that greatly assisted with the method and data development. Special thanks to Dr. Jingyi Huang for providing open-source codes and suggestions which helped us improve the methodology. We appreciate the help from Dr. Edward Ayres in answering questions regarding the CPER sensor data. Our gratitude also extends to Jim Howell and Brandon Dalton for maintaining the Colorado site and examining our site-level modeling results. The communication and data exchange with Dr. Jennie DeMarco has helped to improve the contents of the manuscript. Research Assistant Haydee Hernandez helped with communications and data acquisition in Colorado.

**ADDITIONAL INFORMATION AND DECLARATIONS**

**Funding**

This work was supported by the Rangeland Tracking and Management project funded by Conscience Bay Research, LLC and the Woodwell Fund for Climate Solutions and Conscience. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Grant Disclosures**

The following grant information was disclosed by the authors:
Conscience Bay Research, LLC. and Woodwell Fund for Climate Solutions.
Rangeland Tracking and Management.

**Competing Interests**

The authors declare that they have no competing interests.

**Author Contributions**

- Yushu Xia conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jennifer D. Watts conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Megan B. Machmuller conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Jonathan Sanderman conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Xia et al. (2022), *PeerJ*, DOI 10.7717/peerj.14275
Data Availability
The following information was supplied regarding data availability:

The codes used to generate the models in this study are available in the Supplemental File and at GitHub (https://github.com/xiayushu/RCTM-soil-moisture).

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.14275#supplemental-information.

REFERENCES
Abowarda AS, Bai L, Zhang C, Long D, Li X, Huang Q, Sun Z. 2021. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. Remote Sensing of Environment 255:112301 DOI 10.1016/j.rse.2021.112301.

Adab H, Morbidelli R, Saltalippi C, Moradian M, Ghalhari GAF. 2020. Machine learning to estimate surface soil moisture from remote sensing data. Water 12(11):3223 DOI 10.3390/w12113223.

Ahmad S, Kalra A, Stephen H. 2010. Estimating soil moisture using remote sensing data: a machine learning approach. Advances in Water Resources 33(1):69–80 DOI 10.1016/j.advwatres.2009.10.008.

Akuraju VR, Ryu D, George B. 2021. Estimation of root-zone soil moisture using crop water stress index (CWSI) in agricultural fields. GIScience & Remote Sensing 58(3):340–353 DOI 10.1080/15481603.2021.1877009.

Amatulli G, McInerney D, Sethi T, Strobl P, Domisch S. 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Scientific Data 7(1):162 DOI 10.1038/s41597-020-0479-6.

Babaeian E, Sadeghi M, Jones SB, Montzka C, Vereecken H, Tuller M. 2019. Ground, proximal, and satellite remote sensing of soil moisture. Reviews of Geophysics 57(2):530–616 DOI 10.1029/2018RG000618.

Baldocchi D, Falge E, Gu L, Olson R, Hollinger D, Running S, Anthoni P, Bernhofer C, Davis K, Evans R, Fuentes J, Goldstein A, Katul G, Law B, Lee X, Malhi Y, Meyers T, Munger W, Oechel W, Paw UKT, Pilegaard K, Schmid HP, Valentini R, Verma S, Vesala T, Wilson K, Wofsy S. 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the American Meteorological Society 82(11):2415–2434 DOI 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Banimahd SA, Zand-Parsa S. 2013. Simulation of evaporation, coupled liquid water, water vapor and heat transport through the soil medium. Agricultural Water Management 130:168–177 DOI 10.1016/j.agwat.2013.08.022.

Belgu M, Drăguţ L. 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114(Part A):24–31 DOI 10.1016/j.isprsjprs.2016.01.011.

Bormann H, Klaassen K. 2008. Seasonal and land use dependent variability of soil hydraulic and soil hydrological properties of two Northern German soils. Geoderma 145(3–4):295–302 DOI 10.1016/j.geoderma.2008.03.017.

Breiman L. 2001. Random forests. Machine Learning 45(1):5–32 DOI 10.1023/A:1010933404324.
Briske DD, Joyce LA, Polley HW, Brown JR, Wolter K, Morgan JA, McCarl BA, Bailey DW. 2015. Climate-change adaptation on rangelands: linking regional exposure with diverse adaptive capacity. *Frontiers in Ecology and the Environment* 13(5):249–256 DOI 10.1890/140266.

Cammalleri C, Vogt J. 2015. On the role of land surface temperature as proxy of soil moisture status for drought monitoring in Europe. *Remote Sensing* 7(12):16849–16864 DOI 10.3390/rs71215857.

Carranza C, Nolet C, Pezij M, van der Ploeg M. 2021. Root zone soil moisture estimation with Random Forest. *Journal of Hydrology* 593:125840 DOI 10.1016/j.jhydrol.2020.125840.

Chen S, Arrouays D, Letitia Mulder V, Poggio L, Minasny B, Roudier P, Libohova Z, Lagacherie P, Shi Z, Hannam J, Meersmans J, Richer-de-Forges AC, Walter C. 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: a review. *Geoderma* 409(3):115567 DOI 10.1016/j.geoderma.2021.115567.

Chen M, Willgoose GR, Saco PM. 2015. Investigating the impact of leaf area index temporal variability on soil moisture predictions using remote sensing vegetation data. *Journal of Hydrology* 522(1–2):274–284 DOI 10.1016/j.jhydrol.2014.12.027.

Cichota R, Vogeler I, Snow VO, Webb TH. 2013. Ensemble pedotransfer functions to derive hydraulic properties for New Zealand soils. *Soil Research* 51(2):94 DOI 10.1071/SR12338.

Cosh MH, Caldwell TG, Baker CB, Bolten JD, Edwards N, Goble P, Hofman H, Ochsner TE, Quiring S, Schalk C, Skumanich M, Svoboda M, Woloszyn ME. 2021. Developing a strategy for the national coordinated soil moisture monitoring network. *Vadose Zone Journal* 20(4):398 DOI 10.1002/vzj2.20139.

Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP. 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology* 28(15):2031–2064 DOI 10.1002/joc.1688.

Dari J, Morbidelli R, Saltalippi C, Massari C, Brocca L. 2019. Spatial-temporal variability of soil moisture: Addressing the monitoring at the catchment scale. *Journal of Hydrology* 570(3–4):436–444 DOI 10.1016/j.jhydrol.2019.01.014.

Das NN, Mohanty BP. 2006. Root zone soil moisture assessment using remote sensing and vadose zone modeling. *Vadose Zone Journal* 5(1):296–307 DOI 10.2136/vzj2005.0033.

Deng Y, Wang S, Bai X, Wu L, Cao Y, Li H, Wang M, Li C, Yang Y, Hu Z, Tian S, Lu Q. 2020. Comparison of soil moisture products from microwave remote sensing, land model, and reanalysis using global ground observations. *Hydrological Processes* 34(3):836–851 DOI 10.1002/hyp.13636.

Derner J, Briske D, Reeves M, Brown-Brandl T, Meehan M, Blumenthal D, Travis W, Augustine D, Wilmer H, Scasta D, Hendrickson J, Volesky J, Edwards L, Peck D. 2018. Vulnerability of grazing and confined livestock in the Northern Great Plains to projected mid- and late-twenty-first century climate. *Climatic Change* 146(1–2):19–32 DOI 10.1007/s10584-017-2029-6.

Diamond HJ, Karl TR, Palecki MA, Baker CB, Bell JE, Leeper RD, Easterling DR, Lawrimore JH, Meyers TP, Helfert MR, Goodge G, Thorne PW. 2013. U.S. Climate reference network after one decade of operations: status and assessment. *Bulletin of the American Meteorological Society* 94(4):485–498 DOI 10.1175/BAMS-D-12-00170.1.

Dominati E, Patterson M, Mackay A. 2010. A framework for classifying and quantifying the natural capital and ecosystem services of soils. *Ecological Economics* 69(9):1858–1868 DOI 10.1016/j.ecolecon.2010.05.002.
Dorigo W, Wagner W, Albergel C, Albrecht F, Balsamo G, Brocca L, Chung D, Ertl M, Forkel M, Gruber A, Haas E, Hamer PD, Hirschi M, Ikonen J, de Jeu R, Kidd R, Lahoz W, Liu YY, Miralles D, Mistelbauer T, Nicolai-Shaw N, Parinussa R, Pratola C, Reimer C, van der Schalie R, Seneviratne SI, Smolander T, Lecomte P. 2017. ESA CCI soil moisture for improved earth system understanding: state-of-the art and future directions. Remote Sensing of Environment 203(Part B):185–215 DOI 10.1016/j.rse.2017.07.001.

Dorigo WA, Wagner W, Hohensinn R, Hahn S, Paulik C, Xaver A, Gruber A, Drusch M, Mecklenburg S, van Oevelen P, Robock A, Jackson T. 2011. The international soil moisture network: a data hosting facility for global in situ soil moisture measurements. Hydrology and Earth System Sciences 15(5):1675–1698 DOI 10.5194/hess-15-1675-2011.

Ek MB, Mitchell KE, Lin Y, Rogers E, Grunmann P, Koren V, Gayno G, Tarpley JD. 2003. Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. Journal of Geophysical Research: Atmospheres 108(D22):8851 DOI 10.1029/2002JD003296.

Entekhabi D, Njoku EG, O’Neill PE, Kellogg KH, Crow WT, Edelstein WN, Entin JK, Goodman SD, Jackson TJ, Johnson J, Kimball J, Piepmeier JR, Koster RD, Martin N, McDonald KC, Moghaddam M, Moran S, Reiche R, Shi JC, Spencer MW, Thurman SW, Tsang L, Van Zyl J. 2010. The soil moisture active passive (SMAP) mission. Proceedings of the IEEE 98(5):704–716 DOI 10.1109/JPROC.2010.2043918.

Ford TW, Quiring SM. 2019. Comparison of contemporary in situ, model, and satellite remote sensing soil moisture with a focus on drought monitoring. Water Resources Research 55(2):1565–1582 DOI 10.1029/2018WR024039.

Fu Z, Ciais P, Bastos A, Stoy PC, Yang H, Green JK, Wang B, Yu K, Huang Y, Knohl A, Šigut L, Gharun M, Cuntz M, Arriga N, Roland M, Peichl M, Migliavacca M, Cremonese E, Varlagin A, Brümmer C, Gourlez de la Motte I, Fares S, Buchmann N, El-Madany TS, Pitacco A, Vendrame N, Li Z, Vincke C, Magliulo E, Koebsch F. 2020. Sensitivity of gross primary productivity to climatic drivers during the summer drought of 2018 in Europe. Philosophical Transactions of the Royal Society B: Biological Sciences 375(1810):20190747 DOI 10.1098/rstb.2019.0747.

Fuentes I, Padarian J, Vervoort RW. 2022. Towards near real-time national-scale soil water content monitoring using data fusion as a downscaling alternative. Journal of Hydrology 609(1):127705 DOI 10.1016/j.jhydrol.2022.127705.

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202(3):18–27 DOI 10.1016/j.rse.2017.06.031.

Grassini P, van Bussel LGJ, Van Wart J, Wolf J, Claessens L, Yang H, Boogaard H, de Groot H, van Ittersum MK, Cassman KG. 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Field Crops Research 177:49–63 DOI 10.1016/j.fcr.2015.03.004.

Greifeneder F, Notarnicola C, Wagner W. 2021. A machine learning-based approach for surface soil moisture estimations with Google Earth Engine. Remote Sensing 13(11):2099 DOI 10.3390/rs13112099.

Gyamerah SA, Ngare P, Ikpe D. 2020. Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov Kernel function. Agricultural and Forest Meteorology 280:e107808 DOI 10.1016/j.agrformet.2019.107808.

Hartemink AE, Hempel J, Lagacherie P, McBratney A, McKenzie N, MacMillan RA, Minasny B, Montanarella L, de Mendonça Santos ML, Sanchez P, Walsh M, Zhang G-L. 2010.
GlobalSoilMap.net—a new digital soil map of the world. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, eds. Digital Soil Mapping. Dordrecht: Springer Netherlands, 423–428.

Havstad KM, Peters DPC, Skaggs R, Brown J, Bestelmeyer B, Fredrickson E, Herrick J, Wright J. 2007. Ecological services to and from rangelands of the United States. Ecological Economics 64(2):261–268 DOI 10.1016/j.ecolecon.2007.08.005.

He M, Kimball JS, Running S, Ballantyne A, Guan K, Huemmrich F. 2016. Satellite detection of soil moisture related water stress impacts on ecosystem productivity using the MODIS-based photochemical reflectance index. Remote Sensing of Environment 186:173–183 DOI 10.1016/j.rse.2016.08.019.

Hengl T, Gupta S. 2019. Soil water content (volumetric %) for 33kPa and 1500kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (Version v01). Zenodo. Available at https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_WATERCONTENT-33KPA_USDA-4B1C_M_v01 (accessed 21 March 2022).

Hengl T, Nussbaum M, Wright MN, Heuvelink GB, Gräler B. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6(8):e5518 DOI 10.7717/peerj.5518.

Homer C, Huang C, Yang L, Wylie B, Coan M. 2004. Development of a 2001 national land-cover database for the United States. Photogrammetric Engineering & Remote Sensing 70(7):829–840 DOI 10.14358/PERS.70.7.829.

Huang J, Desai AR, Zhu J, Hartemink AE, Stoy PC, Loheide SP II, Bogena HR, Zhang Y, Zhang Z, Arriaga F. 2020. Retrieving heterogeneous surface soil moisture at 100 m across the globe via fusion of remote sensing and land surface parameters. Frontiers in Water 2:324 DOI 10.3389/frwa.2020.578367.

Huete A, Didan K, Leeuwen WV. 1999. MODIS vegetation index (MOD 13): algorithm theoretical basis document. Tucson, AZ, USA. Available at https://modis.gsfc.nasa.gov/data/atbd/atbd_mod13.pdf.

Hunt ER Jr, Everitt JH, Ritchie JC, Moran MS, Booth DT, Anderson GL, Clark PE, Seyfried MS. 2003. Applications and research using remote sensing for rangeland management. Photogrammetric Engineering & Remote Sensing 69(6):675–693 DOI 10.14358/PERS.69.6.675.

Izaurralde RC, Williams JR, McGill WB, Rosenberg NJ, Jakas MCQ. 2006. Simulating soil C dynamics with EPIC: Model description and testing against long-term data. Ecological Modelling 192(3–4):362–384 DOI 10.1016/j.ecolmodel.2005.07.010.

James SE, Pärtel M, Wilson SD, Peltzer DA. 2003. Temporal heterogeneity of soil moisture in grassland and forest. Journal of Ecology 91(2):234–239 DOI 10.1046/j.1365-2745.2003.00758.x.

Jebellie SJ, Prasher SO, Clemente RS. 1996. Role of soil moisture content in reducing environmental pollution from pesticides. Canadian Water Resources Journal 21(4):393–402 DOI 10.4296/cwrj2104393.

Jiang B, Su H, Liu K, Chen S. 2020. Assessment of remotely sensed and modelled soil moisture data products in the U.S. Southern Great Plains. Remote Sensing 12(12):2030 DOI 10.3390/rs12122030.

Jiang G, Wang N, Zhang Y, Wang Z, Zhang Y, Yu J, Zhang Y, Wei Z, Xu Y, Geisen S, Friman V-P, Shen Q. 2021. The relative importance of soil moisture in predicting bacterial wilt disease occurrence. Soil Ecology Letters 3(4):356–366 DOI 10.1007/s42832-021-00086-2.
Jing W, Zhang P, Zhao X. 2018. Reconstructing monthly ECV global soil moisture with an improved spatial resolution. *Water Resources Management* 32(7):2523–2537 DOI 10.1007/s11269-018-1944-2.

Jones MO, Allred BW, Naugle DE, Maestas JD, Donnelly P, Metz LJ, Karl J, Smith R, Bestelmeyer B, Boyd C, Kerby JD, McIver JD. 2018. Innovation in rangeland monitoring: annual, 30 m, plant functional type percent cover maps for U.S. rangelands, 1984–2017. *Ecosphere* 9(9) DOI 10.1002/ecs2.2430.

Kahimba FC, Ranjan RS, Mann DD. 2009. Modeling soil temperature, frost depth, and soil moisture redistribution in seasonally frozen agricultural soils. *Applied Engineering in Agriculture* 25(6):871–882 DOI 10.13031/2013.29237.

Karandish F, Šimůnek J. 2016. A comparison of numerical and machine-learning modeling of soil water content with limited input data. *Journal of Hydrology* 543(3):892–909 DOI 10.1016/j.jhydrol.2016.11.007.

Karthikeyan L, Mishra AK. 2021. Multi-layer high-resolution soil moisture estimation using machine learning over the United States. *Remote Sensing of Environment* 266:112706 DOI 10.1016/j.rse.2021.112706.

Kathuria D, Mohanty BP, Katzfuss M. 2019. A nonstationary geostatistical framework for soil moisture prediction in the presence of surface heterogeneity. *Water Resources Research* 55(1):729–753 DOI 10.1029/2018WR023505.

Khaledian Y, Miller BA. 2020. Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling* 81(1–2):401–418 DOI 10.1016/j.apm.2019.12.016.

Koenker R, Bassett GJ. 1978. Regression quantiles. *Econometrica* 46(1):33–50 DOI 10.2307/1913643.

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

Kuang W, Gao X, Tenuta M, Gui D, Zeng F. 2019. Relationship between soil profile accumulation and surface emission of N₂O: effects of soil moisture and fertilizer nitrogen. *Biocybernetics and Biomedical Engineering* 39(2):127–134 DOI 10.2478/bbee-2019-0009.

Kuhn M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28(5):1–26 DOI 10.18637/jss.v028.i05.

Lamichhane S, Kumar I, Wilson B. 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review. *Geoderma* 352(1):395–413 DOI 10.1016/j.geoderma.2019.05.031.

Law B. 2005. Carbon dynamics in response to climate and disturbance: recent progress from multi-scale measurements and modeling in AmeriFlux. In: *Plant Responses to Air Pollution and Global Change*. Tokyo: Springer Japan, 205–213.

Lee E, Kim S. 2019. Wavelet analysis of soil moisture measurements for hillslope hydrological processes. *Journal of Hydrology* 575(9):82–93 DOI 10.1016/j.jhydrol.2019.05.023.

Lehner B, Verdin K, Jarvis A. 2008. New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union* 89(10):93 DOI 10.1029/2008EO100001.

Li X, Shao M, Zhao C, Liu T, Jia X, Ma C, Naeth MA. 2019. Regional spatial variability of root-zone soil moisture in arid regions and the driving factors—a case study of Xinjiang, China (M.A. Naeth, editor). *Canadian Journal of Soil Science* 99(3):277–291 DOI 10.1139/cjss-2019-0006.
Liao W, Wang D, Wang G, Xia Y, Liu X. 2019. Quality control and evaluation of the observed daily data in the North American Soil Moisture Database. *Journal of Meteorological Research* 33(3):501–518 DOI 10.1007/s13351-019-8121-2.

Liu Y, Xia X, Yao L, Jing W, Zhou C, Huang W, Li Y, Yang J. 2020. Downscaling satellite retrieved soil moisture using regression tree-based machine learning algorithms over southwest France. *Earth and Space Science* 7(10):358 DOI 10.1029/2020EA001267.

Lowery B, Hickey WJ, Arshad MAC, Lal R. 1997. Soil water parameters and soil quality. *Methods for Assessing Soil Quality* 49:143–155 DOI 10.2136/sssaspecpub49.

Martínez-Murillo JF, Hueso-González P, Ruiz-Sinoga JD. 2017. Topsoil moisture mapping using geostatistical techniques under different Mediterranean climatic conditions. *Science of the Total Environment* 595(9):400–412 DOI 10.1016/j.scitotenv.2017.03.291.

McBratney AB, Mendonça Santos ML, Minasny B. 2003. On digital soil mapping. *Geoderma* 117(1–2):3–52 DOI 10.1016/S0016-7061(03)00223-4.

Meinshausen N. 2006. Quantile regression forests. *Journal of Machine Learning* 7(35):983–999.

Metzger S, Ayres E, Durden D, Florian C, Lee R, Lunch C, Luo H, Pingintha-Durden N, Roberti JA, SanClements M, Sturtevant C, Xu K, Zulueta RC. 2019. From NEON field sites to data portal: a community resource for surface-atmosphere research comes online. *Bulletin of the American Meteorological Society* 100(11):2305–2325 DOI 10.1175/BAMS-D-17-0307.1.

Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software* 101(Part A):1–9 DOI 10.1016/j.envsoft.2017.12.001.

Miller BA, Koszinski S, Hierold W, Rogasik H, Schröder B, Van Oost K, Wehrhan M, Sommer M. 2016. Towards mapping soil carbon landscapes: issues of sampling scale and transferability. *Soil and Tillage Research* 156:194–208 DOI 10.1016/j.still.2015.07.004.

Moran MS, Hymer DC, Qi J, Sano EE. 2000. Soil moisture evaluation using multi-temporal synthetic aperture radar (SAR) in semiarid rangeland. *Agricultural and Forest Meteorology* 105(1–3):69–80 DOI 10.1016/S0168-1923(00)00189-1.

Nawar S, Mouazen AM. 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *CATENA* 151(4):118–129 DOI 10.1016/j.catena.2016.12.014.

Niu CY, Musa A, Liu Y. 2015. Analysis of soil moisture condition under different land uses in the arid region of Horqin sandy land, northern China. *Solid Earth* 6(4):1157–1167 DOI 10.5194/se-6-1157-2015.

Parton WJ, Rasmussen PE. 1994. Long-term effects of crop management in wheat-fallow: II. CENTURY model simulations. *Soil Science Society of America Journal* 58(2):530–536 DOI 10.2136/sssaj1994.0361599500580020040x.

Patel NR, Anapashsha R, Kumar S, Saha SK, Dadhwal VK. 2009. Assessing potential of MODIS derived temperature/vegetation condition index (TVDI) to infer soil moisture status. *International Journal of Remote Sensing* 30(1):23–39 DOI 10.1080/01431160802108497.

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

Platnick S, King MD, Ackerman SA, Menzel WP, Baum BA, Riedi JC, Frey RA. 2003. The MODIS cloud products: algorithms and examples from terra. *IEEE Transactions on Geoscience and Remote Sensing* 41(2):459–473 DOI 10.1109/TGRS.2002.808301.
Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, Lyapustin A, Gourlet-Fleury S, PéliSSier R. 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* 11(1):4540 DOI 10.1038/s41467-020-18321-y.

Quiring SM, Ford TW, Wang JK, Khong A, Harris E, Lindgren T, Goldberg DW, Li Z. 2016. The North American Soil Moisture Database: development and applications. *Bulletin of the American Meteorological Society* 97(8):1441–1459 DOI 10.1175/BAMS-D-13-00263.1.

R Core Team. 2021. R: a language and environment for statistical computing. Available at https://www.R-project.org/ (accessed 10 March 2022).

Ramcharan A, Hengl T, Nauman T, Brungard C, Waltman S, Wills S, Thompson J. 2018. Soil property and class maps of the conterminous United States at 100-meter spatial resolution. *Soil Science Society of America Journal* 82(1):186–201 DOI 10.2136/sssaj2017.04.0122.

Reeves MC, Mitchell JE. 2012. A synoptic review of U.S. rangelands: a technical document supporting the Forest Service 2010 RPA Assessment. Fort Collins, CO, USA.

Ren S, Guo B, Wang Z, Wang J, Fang Q, Wang J. 2022. Optimized spectral index models for accurately retrieving soil moisture (SM) of winter wheat under water stress. *Agricultural Water Management* 261:107333 DOI 10.1016/j.agwat.2021.107333.

Richardson AJ, Everitt JH. 1992. Using spectral vegetation indices to estimate rangeland productivity. *Geocarto International* 7(1):63–69 DOI 10.1080/10106049209354353.

Robinson NP, Allred BW, Smith WK, Jones MO, Moreno A, Erickson TA, Naugle DE, Running SW, Pettorelli N, Paruelo J. 2018. Terrestrial primary production for the conterminous United States derived from Landsat 30 m and MODIS 250 m. *Remote Sensing in Ecology and Conservation* 4(3):264–280 DOI 10.1002/rse2.74.

Robinson DA, Campbell CS, Hopmans JW, Hornbuckle BK, Jones SB, Knight R, Ogden F, Selker J, Wendroth O. 2008. Soil moisture measurement for ecological and hydrological Watershed-scale observatories: a review. *Vadose Zone Journal* 7(1):358–389 DOI 10.2136/vzj2007.0143.

Rodell M, Houser PR, Jambor U, Gottschalck J, Mitchell K, Meng C-J, Arsanault K, Cosgrove B, Radakovitch J, Bosilovich M, Entin JK, Walker JP, Lohmann D, Toll D. 2004. The global land data assimilation system. *Bulletin of the American Meteorological Society* 85(3):381–394 DOI 10.1175/BAMS-85-3-381.

Sadeghi M, Babaiean E, Tuller M, Jones SB. 2017. The optical trapezoid model: a novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat-8 observations. *Remote Sensing of Environment* 198(11):52–68 DOI 10.1016/j.rse.2017.05.041.

Safanelli J, Poppil R, Ruiz L, Bonfatti B, Mello F, Rizzo R, Dematté J. 2020. Terrain analysis in google earth engine: a method adapted for high-performance global-scale analysis. *ISPRS International Journal of Geo-Information* 9(6):400 DOI 10.3390/ijgi9060400.

Salvucci GD, Saleem JA, Kaufmann R. 2002. Investigating soil moisture feedbacks on precipitation with tests of Granger causality. *Advances in Water Resources* 25(8–12):1305–1312 DOI 10.1016/S0309-1708(02)00057-X.

Sánchez N, Martín-Calvo A, Martínez-Fernández J, González-Piquer J, González-Dugo MP, Baroncini-Turrichia G, Torres E, Calera A, Pérez-Gutiérrez C. 2012. Water balance at plot scale for soil moisture estimation using vegetation parameters. *Agricultural and Forest Meteorology* 166–167:1–9 DOI 10.1016/j.agrformet.2012.07.005.

Sandholt I, Rasmussen K, Andersen J. 2002. A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote Sensing of Environment* 79(2–3):213–224 DOI 10.1016/S0034-4257(01)00274-7.
Saxton KE, Rawls WJ. 2006. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Science Society of America Journal* 70(5):1569–1578 DOI 10.2136/sssaj2005.0117.

Schaefer GL, Paetzold RF. 2000. SNOTEL (SNOwpack TELemetry) and SCAN (soil climate analysis network). Automated weather stations for applications in agriculture and water resources management: current use and future perspectives. *Journal of Hydrometeorology* 7(2):187–194.

Schaefer GL, Cosh MH, Jackson TJ. 2007. The USDA natural resources conservation service soil climate analysis network (SCAN). *Journal of Atmospheric and Oceanic Technology* 24(12):2073–2077 DOI 10.1175/2007JTECHA930.1.

Schuman GE, Janzen HH, Herrick JE. 2002. Soil carbon dynamics and potential carbon sequestration by rangelands. *Environmental Pollution* 116(3):391–396 DOI 10.1016/S0269-7491(01)00215-9.

Schnur MT, Xie H, Wang X. 2010. Estimating root zone soil moisture at distant sites using MODIS NDVI and EVI in a semi-arid region of southwestern USA. *Ecological Informatics* 5(5):400–409 DOI 10.1016/j.ecoinf.2010.05.001.

Schoener G, Stone MC. 2020. Monitoring soil moisture at the catchment scale—a novel approach combining antecedent precipitation index and radar-derived rainfall data. *Journal of Hydrology* 589(10):125155 DOI 10.1016/j.jhydrol.2020.125155.

Sehler R, Li J, Reager J, Ye H. 2019. Investigating relationship between soil moisture and precipitation globally using remote sensing observations. *Journal of Contemporary Water Research & Education* 168(1):106–118 DOI 10.1111/j.1936-704X.2019.03324.x.

Senyurek V, Lei F, Boyd D, Gurbuz AC, Kurum M, Moorhead R. 2020. Evaluations of machine learning-based CYGNSS soil moisture estimates against SMAP observations. *Remote Sensing* 12(21):3503 DOI 10.3390/rs12213503.

Siabi N, Sanaeinejad SH, Ghahrama B. 2022. Effective method for filling gaps in time series of environmental remote sensing data: an example on evapotranspiration and land surface temperature images. *Computers and Electronics in Agriculture* 193(5):e106619 DOI 10.1016/j.compag.2021.106619.

Sohrabi MM, Ryu JH, Abatzoglou J, Tracy J. 2015. Development of soil moisture drought index to characterize droughts. *Journal of Hydrologic Engineering* 20(11):04015025 DOI 10.1061/(ASCE)HE.1943-5584.0001213.

Spangler K, Burchfield EK, Schumacher B. 2020. Past and current dynamics of U.S. agricultural land use and policy. *Frontiers in Sustainable Food Systems* 4:e00098 DOI 10.3389/fsufs.2020.00098.

Srivastava PK. 2017. Satellite soil moisture: review of theory and applications in water resources. *Water Resources Management* 31(10):3161–3176 DOI 10.1007/s11269-017-1722-6.

Stafford JV. 1988. Remote, non-contact and in-situ measurement of soil moisture content: a review. *Journal of Agricultural Engineering Research* 41(3):151–172 DOI 10.1016/0021-8634(88)90175-8.

Steinert A, Morgan F. 2016. NEON site-level plot summary central plains experimental range (CPER). 1–5. Available at https://data.neonscience.org/documents/10179/2361410/CPER_Soil_SiteSummary/e40eb8b-4abd-4061-8077-05bd29d25a7c (accessed 2 May 2022).

Strudley MW, Green TR, Ascough JC. 2008. Tillage effects on soil hydraulic properties in space and time: state of the science. *Soil and Tillage Research* 99(1):4–48 DOI 10.1016/j.still.2008.01.007.
Sylvain ZA, Wall DH. 2011. Linking soil biodiversity and vegetation: Implications for a changing planet. *American Journal of Botany* 98(3):517–527 DOI 10.3732/ajb.1000305.

Taghizadehghasab A, Safadoust A, Mosaddeghi MR. 2021. Effects of salinity and sodicity of water on friability of two texturally-different soils at different matric potentials. *Soil and Tillage Research* 209:e104950 DOI 10.1016/j.still.2021.104950.

Tavakol A, Rahmani V, Quiring SM, Kumar SV. 2019. Evaluation analysis of NASA SMAP L3 and L4 and SPoRT-LIS soil moisture data in the United States. *Remote Sensing of Environment* 229(10):234–246 DOI 10.1016/j.rse.2019.05.006.

Thornton MM, Shrestha R, Wei Y, Thornton PE, Kao S, Wilson B. 2020. Daymet: daily surface weather data on a 1-km grid for North America, version 4. Oak Ridge, Tennessee, USA DOI 10.3334/ORNLDAAC/1840.

Townsend J, DiMiceli C. 2015. MOD44B MODIS/Terra vegetation continuous fields yearly L3 global 500m SIN grid. University of Maryland and MODAPS SIPS – NASA DOI 10.5067/MODIS/MOD44B.006.

Udawatta RP, Gantzer CJ. 2022. Soil and water ecosystem services of agroforestry. *Journal of Soil and Water Conservation* 77(1):5A–11A DOI 10.2489/jswc.2022.1028A.

van Zyl JJ. 2001. The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. *Acta Astronautica* 48(5–12):559–565 DOI 10.1016/S0094-5765(01)00020-0.

Vaysse K, Lagacherie P. 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* 291(325):55–64 DOI 10.1016/j.geoderma.2016.12.017.

Velpuri NM, Senay GB, Morisette JT. 2016. Evaluating new SMAP soil moisture for drought monitoring in the rangelands of the US High Plains. *Rangelands* 38(4):183–190 DOI 10.1016/j.rala.2016.06.002.

Vereecken H, Huisman JA, Pachepsky Y, Montzka C, van der Kruk J, Bogena H, Weihermüller L, Herbst M, Martinez G, Vanderborght J. 2014. On the spatio-temporal dynamics of soil moisture at the field scale. *Journal of Hydrology* 516(4):76–96 DOI 10.1016/j.jhydrol.2013.11.061.

Vergopolan N, Chaney NW, Pan M, Sheffield J, Beck HE, Ferguson CR, Torres-Rojas L, Sadri S, Wood EF. 2021. SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. *Scientific Data* 8(1):264 DOI 10.1038/s41597-021-01050-2.

Verstraeten W, Veroustraete F, Feyen J. 2008. Assessment of evapotranspiration and soil moisture content across different scales of observation. *Sensors* 8(1):70–117 DOI 10.3390/s8010070.

Wagner W. 2003. Evaluation of the agreement between the first global remotely sensed soil moisture data with model and precipitation data. *Journal of Geophysical Research* 108(D19):4611 DOI 10.1029/2003JD003663.

Walker JP, Willgoose GR, Kalma JD. 2004. In situ measurement of soil moisture: a comparison of techniques. *Journal of Hydrology* 293(1–4):85–99 DOI 10.1016/j.jhydrol.2004.01.008.

Wei L, Zhang B, Wang M. 2007. Effects of antecedent soil moisture on runoff and soil erosion in alley cropping systems. *Agricultural Water Management* 94(1–3):54–62 DOI 10.1016/j.agwat.2007.08.007.

Weiss DJ, Atkinson PM, Bhatt S, Mappin B, Hay SI, Gething PW. 2014. An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS Journal of Photogrammetry and Remote Sensing* 98:106–118 DOI 10.1016/j.isprsjprs.2014.10.001.
Wu X, Walker JP, Rudiger C, Panciera R. 2015. Effect of land-cover type on the SMAP Active/Passive soil moisture downscaling algorithm performance. IEEE Geoscience and Remote Sensing Letters 12(4):846–850 DOI 10.1109/LGRS.2014.2364049.

Xia Y, Mitchell K, Ek M, Sheffield J, Cosgrove B, Wood E, Luo L, Alonge C, Wei H, Meng J, Livneh B, Lettenmaier D, Koren V, Duan Q, Mo K, Fan Y, Mocko D. 2012a. NLDAS Noah Land Surface Model L4 Hourly 0.125 x 0.125 degree V002. Goddard Earth Sciences Data and Information Services Center (GES DISC). Available at https://disc.gsfc.nasa.gov/datasets/NLDAS_NOAH0125_H_002/summary (accessed 10 March 2022).

Xia Y, Mitchell K, Ek M, Sheffield J, Cosgrove B, Wood E, Luo L, Alonge C, Wei H, Meng J, Livneh B, Lettenmaier D, Koren V, Duan Q, Mo K, Fan Y, Mocko D. 2012b. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. Journal of Geophysical Research: Atmospheres 117(D3):1–27 DOI 10.1029/2011JD016048.

Xiao L, Xue S, Liu G-B, Zhang C. 2014. Soil moisture variability under different land uses in the Zhifanggou catchment of the Loess Plateau, China. Arid Land Research and Management 28(3):274–290 DOI 10.1080/15324982.2013.860500.

Yadav BK, Mathur S, Siebel MA. 2009. Soil moisture dynamics modeling considering the root compensation mechanism for water uptake by plants. Journal of Hydrologic Engineering 14(9):913–922 DOI 10.1061/(ASCE)HE.1943-5584.0000066.

Zhang Y, Ji W, Saurette DD, Easher TH, Li H, Shi Z, Adamchuk VI, Biswas A. 2020. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma 366:114253 DOI 10.1016/j.geoderma.2020.114253.

Zhu Q, Lin H. 2011. Influences of soil, terrain, and crop growth on soil moisture variation from transect to farm scales. Geoderma 163(1–2):45–54 DOI 10.1016/j.geoderma.2011.03.015.

Zreda M, Shuttleworth WJ, Zeng X, Zweck C, Desilets D, Franz T, Rosolem R. 2012. COSMOS: the COsmic-ray soil moisture observing system. Hydrology and Earth System Sciences 16(11):4079–4099 DOI 10.5194/hess-16-4079-2012.

Zucco G, Brocca L, Moramarco T, Morbidelli R. 2014. Influence of land use on soil moisture spatial–temporal variability and monitoring. Journal of Hydrology 516(3–4):193–199 DOI 10.1016/j.jhysol.2014.01.043.