Journal of Hydrology 652 (2025) 132624

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Downscaling soil moisture to sub-km resolutions with simple machine learning ensembles

Jeran Poehls [a,*], Lazaro Alonso [a], Sujan Koirala [a], Markus Reichstein [a,b], Nuno Carvalhais [a,c,d]

[a] Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany
[b] German Centre for Integrative Biodiversity Research (iDiv) Halle–Jena–Leipzig, Leipzig, Germany
[c] ELLIS Unit Jena, Jena, Germany
[d] CENSE, Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Caparica, Portugal

ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief, with the assistance of Yiwen Mei, Associate Editor.

Keywords:
Ensemble
Soil moisture
Remote sensing
Downscaling
SMAP

ABSTRACT

Soil moisture is a key factor that influences the productivity and energy balance of ecosystems and biomes. Global soil moisture measurements have coarse native resolutions of 36km and infrequent revisits of around three days. However, these limitations are not present for many variables connected to soil moisture such as land surface temperature and evapotranspiration. For this reason many previous studies have aimed to discern the relationships between these higher resolution variables and soil moisture to produce downscaled soil moisture products.

In this study, we test four ensemble machine learning models for this downscaling task. These models use a dataset of over 1,000 sites across the US to predict soil moisture at sub-km scales. We find that all models, particularly one with a very simple structure, can outperform Soil Moisture Active Passive (SMAP) measurements on a cross-fold analysis of the 1,000+ sites. This model has an average ubRMSE of **0.058** vs SMAPs **0.065** and an average R of **0.638** vs SMAPs **0.562**. Not all ensembles are beneficial, with some architectures performing better with different training weights than with ensemble averaging. However, some ensembles capture more of the land surface characteristics than ensemble members. Lastly, although general improvements over SMAP are observed, there appears to be difficulty in consistently doing so in cropland regions with high clay and low sand content.

1. Introduction

Soil water content (SWC) has a strong coupling with ecosystem stress and production (Liu et al., 2020; Fu et al., 2022; Stocker et al., 2019). SWC is most commonly measured in-situ by changes in electric current passing through the soil. Although accurate, these measurements require an investment of resources, must be calibrated for the soil being measured, and are impractical for observing SWC across regional areas (Bittelli, 2011). For larger scale SWC measurements, one can estimate SWC by observing changes in radiation intensities from absorption by water molecules in the soils surface. Field scale measurements can be made via drones using ground penetrating radar (Wu et al., 2019). But for truly global scale soil moisture mapping we need to look for the aid of satellites.

The Soil Moisture Active Passive (SMAP) mission launched by NASA in 2015 served to be the solution to global SWC measurements. This satellite combines higher resolution active radar measurements with lower resolution passive radiometer measurements (Entekhabi, 2014). The combination of these two would yield native SWC measurements at 9 km per pixel and interpolated 1–3 km products for finer resolution. However, after only three months in orbit, the power supply for the active radar component failed leaving just the low resolution radiometer sensor. The native resolution of the current radiometer sensor is 36 km per pixel. This resolution can be increased using the Backus–Gilbert optimal interpolation algorithm to 9 km per pixel with acceptable accuracy (ONeill et al., 2019). This lack of resolution has lead to multiple efforts to attempt a downscaling of the SMAP products to provide SWC predictions on scales ranging from 100 m–3 km. A higher resolution product is important as even at 1 km resolution, up to 80% of SWC variability is lost (Vergopolan et al., 2022). At native satellite resolutions, there is a complete loss of SWC variability (Vergopolan et al., 2022).

Spatial variability of SWC influences a multitude of factors including evapotranspiration, surface temperature, cloud formation, and convective rainfall to name a few of many. This loss in high resolution variability and information makes remotely sensed SWC products limiting as inputs for regional physical models. For this reason, an

* Correspondence to: Hans-Knöll-Straße 10, 07745 Jena, Germany
E-mail address: jpoehls@bgc-jena.mpg.de (J. Poehls).

https://doi.org/10.1016/j.jhydrol.2024.132624
Received 26 February 2024; Received in revised form 29 October 2024; Accepted 15 December 2024
Available online 9 January 2025
0022-1694/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

increase in understanding for SWC variability and a higher resolution SWC data product would have a wide range of applications and benefits in Earth science modeling (Naz et al., 2019; Koné et al., 2022a,b). Efforts to increase resolution or "downscale" soil moisture measurements, generally, are either empirically based or derived from machine learning.

The most common empirical method is the DISaggregation based on a Physical and Theoretical Scale Change (DisPATCH) algorithm. This algorithm is a theoretical conversion of soil temperature fields into soil moisture fields. SWC is predicted through the use of a semi-empirical soil evaporative efficiency (SEE) model and the soils average moisture content. DisPATCH performs well on bare soils, but struggles when the soils are occluded either by vegetation or clouds. It also demonstrates inconsistencies in more humid regions (Colliander et al., 2017; Ojha et al., 2021; Zheng et al., 2021). A strong advantage however, is that DisPATCH's resolution is only limited by temperature field resolution. This provides an opportunity to use higher resolution derived LST products for even higher resolution SWC predictions (Sánchez et al., 2020; Ojha et al., 2019). But higher resolution LST data would not improve the models performance against dense vegetation and is still limited by cloud cover.

The machine learning field has also seen a large number of approaches for this downscaling task (Abbaszadeh et al., 2019; Xu et al., 2022; Zhao et al., 2022; Montzka et al., 2018). However, a common occurrence are complex model architectures over particularly limited study areas (Abowarda et al., 2021; Xu et al., 2021; Cai et al., 2022). Complex architectures and workflows serve to further reveal the scope and capabilities of machine learning methods in this task. But their complexities also decrease their reproducibility as they require an increased effort to incorporate. Additionally, many of these complex architectures have only been validated on smaller more homogeneous regions. Therefore, an ideal scenario is an easy to reproduce architecture with a wider region of validation. The works of Abbaszadeh et al. 2018 and more recently Xu et al. 2022 serve as great inspirations to this concept. They employed relatively simple models over larger regions of interest. Abbaszadeh's approach demonstrated the advantage of an ensemble of random forest predictions whereas Xu's approach demonstrated the capabilities of a simple neural network architecture.

Using the work of Abbaszadeh and Xu as inspiration, this study will explore the performance of four different model architectures for downscaling coarse spatial resolution soil moisture data to sub-km resolutions. The four models include: two probabilistic estimators consisting of simple neural networks, a wide-deep learning (WDL) architecture modeled after the work of Xu et al. 2022, and a random forest (RF) model. These models will be trained on a large dataset comprised of in-situ soil moisture measurements and ancillary remote sensing predictors across the continental US with sub-km resolutions. The models will then be used to make spatial and temporal predictions of soil moisture. Additionally, analysis will be conducted to conclude the robustness of these models and generalizability. Lastly, we will look at the viability of using ensembles. This will assess if the models derive any benefit from ensemble averaging, or if single ensemble members can predict adequately on their own. The overarching goal is to demonstrate the feasibility of using ensembles of simple machine learning architectures to downscale spatially coarse soil moisture products to sub-km resolutions across a heterogeneous landscape.

2. Data

Machine learning models like decision trees and non-linear regression can predict outcomes given certain input parameters. However, they require large amounts of data to identify meaningful trends and patterns. To ensure our models can make soil moisture predictions across a large spatial area (Fig. 1), we first need to accumulate a sizable dataset with relevant input variables for analysis. The first step is feature selection.

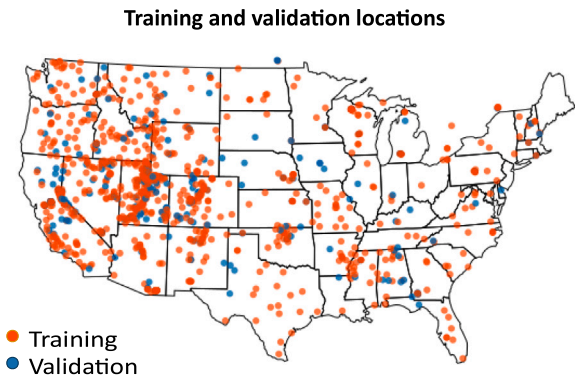


Fig. 1. Spatial distribution of CONUS dataset.

2.0.1. Feature selection

A large dataset of variables used in previous studies and with known or potential correlations with SWC was compiled. Variables were compared to in-situ SWC measurements. Those with Pearson, Kendall or Spearman correlations below a certain threshold were marked as potential candidates for exclusion from the final dataset. Additionally, a single large RF model with 54 million nodes was fit to this dataset to predict SWC. Inputs to this large RF model with a variable importance less than 25% of the most important variable (SMAP) were marked for potential exclusion from the final dataset. Variables who received two or more marks from the four tests were removed from the dataset.

2.0.2. Feature importance

The final dataset consisted of the following variables: *SMAP, NDVI, LST, Precipitation, Sand* and *Clay content, pH, Evapotranspiration,* and *Topography/Elevation*.

This dataset was then iteratively trained over while excluding one variable. The magnitude of drop in performance for each session was then used to assign a rank of importance for that variable. These variables ranked by importance are as follows:

$$SMAP > LST > Sand > ET > Precip > Topography > Clay > NDVI > pH$$

2.1. Datasets

Acquired data extends over the entire continental United States (CONUS) within a temporal period from **January 1st, 2017** through **December 31st, 2021**. This period ensured that soil moisture readings would include seasonal and, potentially, yearly variability. A map of coverage can be seen in Fig. 1.

Locations are categorized by soil texture class. For each class, 80% of sites and all samples are allocated to the training set, while the remaining 20% of sites and samples are designated for the validation set. This ensures that models are learning generalizable trends that translate to new locations. In-situ measurement's are aggregated to daily readings and are paired with daily aggregates of covariate inputs. The final dataset consists of 657,935 samples from 1054 stations, with 206 stations included in the validation set.

For additional validation, two datasets from a network of soil moisture stations used to calibrate SMAP will be utilized to evaluate performance. Details on these datasets are provided in the supplementary document.

2.2. Data sources

2.2.1. Soil moisture active passive (SMAP) satellite product

Remotely sensed soil moisture readings are provided by NASAs SMAP satellite mission. The SMAP satellite provides passive radiometer measurements which permits inference of moisture content in the top 5 cm of soil. Satellite readings have global coverage with a return period between 2–3 days (Entekhabi, 2014). SMAP data is offered at varying levels of post-processing. Two levels of interest are L3 and L4. L3 consists of preprocessed measurements that are gridded and mapped spatiotemporally across the globe. L4 data is a gapfilled product derived from L3. The L4 product offers much greater spatio-temporal coverage and would offer greater data availability. However, training on the L3 product yielded better metrics and so the L3 product was selected.

Observations from L3 passes occur at either AM or PM timestamps. To improve SMAP L3's temporal coverage, both AM and PM readings are treated as one daily reading and both are averaged if a location receives both in the same day. Because in-situ data will be aggregated into daily readings, SWC measurements with finer than daily resolution are not considered.

2.2.2. Moderate resolution imaging spectroradiometer (MODIS) data

The MODIS mission provides daily data with sun-synchronous orbits, capturing spectral reflectance across various wavelengths to analyze Earth's surface. Land Surface Temperature (LST), Evapotranspiration (ET), and the Normalized Difference Vegetation Index (NDVI) are surface properties used in this study. For NDVI, the 500 m product (MOD13A1) is used for training and temporal predictions, while the 250 m product (MOD13Q1) is used for spatial predictions. For LST, the 8-day product (MOD11A2) is utilized during training to minimize cloud interference, and the daily LST product (MOD21A1) is used for spatial prediction. ET inputs come from the 8-day ET product (MOD16A1) derived from a modified Penman–Montieth equation. This has a spatial resolution of 500 m. For land cover classification, the MCD12Q1 product is employed with a 1-year temporal and 500 m spatial resolution.

2.2.3. CHIRPS 2.0 Precipitation

Precipitation data was retrieved from the Climate Hazards Center at Santa Barbara (Funk et al., 2015). Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) is a combination between models of terrain-induced precipitation enhancement with interpolated station data and satellite based precipitation estimates. This data provides daily global precipitation coverage estimates at 0.05° spatial resolution (~5.5 km).

2.2.4. Soil texture and soilgrids

The International Soil Reference and Information Centre (ISRIC) has produced a global harmonized soil properties database called SoilGrids (Hengl et al., 2017). Although higher fidelity datasets are available for specific regions of interest from local entities, the globally consistent nature of the SoilGrids data implies wider implementation of methods using it. A 1 km resolution version of SoilGrids was used as the coarser resolution will be less sensitive to interpolation artifacts. The Sand, Clay, pH, and USDA soil classification data products were used for this study.

2.2.5. Topography

The Multi-Error-Removed Improved-Terrain (MERIT) Digital Elevation Model (DEM) topography product was used for this study (Yamazaki et al., 2017). This product has a spatial resolution of ~90 m.

2.2.6. In-situ soil moisture measurements

Ground truth data for training were obtained from in-situ SWC measurements from two networks at sites distributed throughout CONUS. The International Soil Moisture Network (ISMN) is an international cooperation to provide and maintain a global database of in-situ soil moisture measurements (Dorigo et al., 2011). Ameriflux is a network of flux towers spread across North America recording various atmospheric and meteorological data and fluxes (Boden et al., 2013). Some sites are equipped with SWC sensors. Data for sites from both networks located within the study area and active during the study period were used. ISMN data comes with a quality flag, thus, only data with a 'G' [good] quality flag were accepted.

Ameriflux data does not have quality flags for all measurements. In order to maintain consistency with ISMN measurements, the Ameriflux data was pruned to only contain readings with similar properties to ISMN readings with a 'G' quality flag (Dorigo et al., 2013). Ameriflux samples were dropped if either the LST reading was below 3 °C or the SWC reading was above 0.7 $\mathbf{m^3/m^3}$. This removed sites with potentially frozen ground and saturated soils. Additionally, sites in wetland and chronically inundated regions were excluded from the dataset.

SWC measurements are then aggregated to daily averages.

3. Models and methods

In order to increase SWC remote sensing resolution, a multivariate dataset comprising variables with a known correlation to SWC was assembled. These covariates are *SMAP, LST, sand* and *clay content, pH, NDVI, ET, Topography,* and *Precipitation.* These variables are spatially confined to locations with in-situ soil moisture measurements that are used as a target for the training of model architectures. The study evaluates four model architectures, referred to as **RF**, **WDL**, **Dense**, and **Prob**. The **RF** and **WDL** models replicate the architectures used by Abbaszadeh et al. (2019) and Xu et al. (2022), respectively. The **Dense** and **Prob** models are simpler, distance-based models: **Dense** is a feed-forward network, while **Prob** includes a probabilistic layer. Both Dense and Prob models were designed to have a similar number of hidden parameters. The architectures for all models are illustrated in Fig. 2.

3.1. Architectures

Random forest

The random forest (RF) ensemble borrows heavily from the work of Abbaszadeh et al. (2019). This ensemble is comprised of ten unique random forest models composed of 100 trees (Fig. 2(a)). Each of the ten forests are trained on specific data belonging to one of the texture classes present in the training data as seen in the texture column of Table 1. The different soil texture classes are not equally represented in the training dataset. This imbalanced distribution of texture types introduces a bias in the data that each model sees during training. This means each member on it's own would produce inconsistent predictions on samples outside of it's texture class. However, averaging all of the predictions of the forests for each sample has been shown to produce robust predictions (Abbaszadeh et al., 2019).

Wide-deep model

The Wide-Deep Learning (WDL) architecture (Fig. 2(b)) is based on the work of Cheng et al. (2016) which was applied to soil moisture downscaling (Xu et al., 2022). The WDL structure consists of one simple Dense Neural Network (DNN) with two hidden layers and a Generalized Linear Model (GLM) which consists of no hidden layers. This architecture aims to capture fine details with the DNN while using the GLM to guide the output and prevent overfitting. The WDL model in the referenced paper consists of two hidden layers with the first containing 128 units and the second containing 64. The WDL architecture in this study also uses the same number of units. The WDL model is designed to incorporate categorical data into the DNN portion of the model. This means this method will have additional information to learn from in the form of categorical data as seen in Table 1.

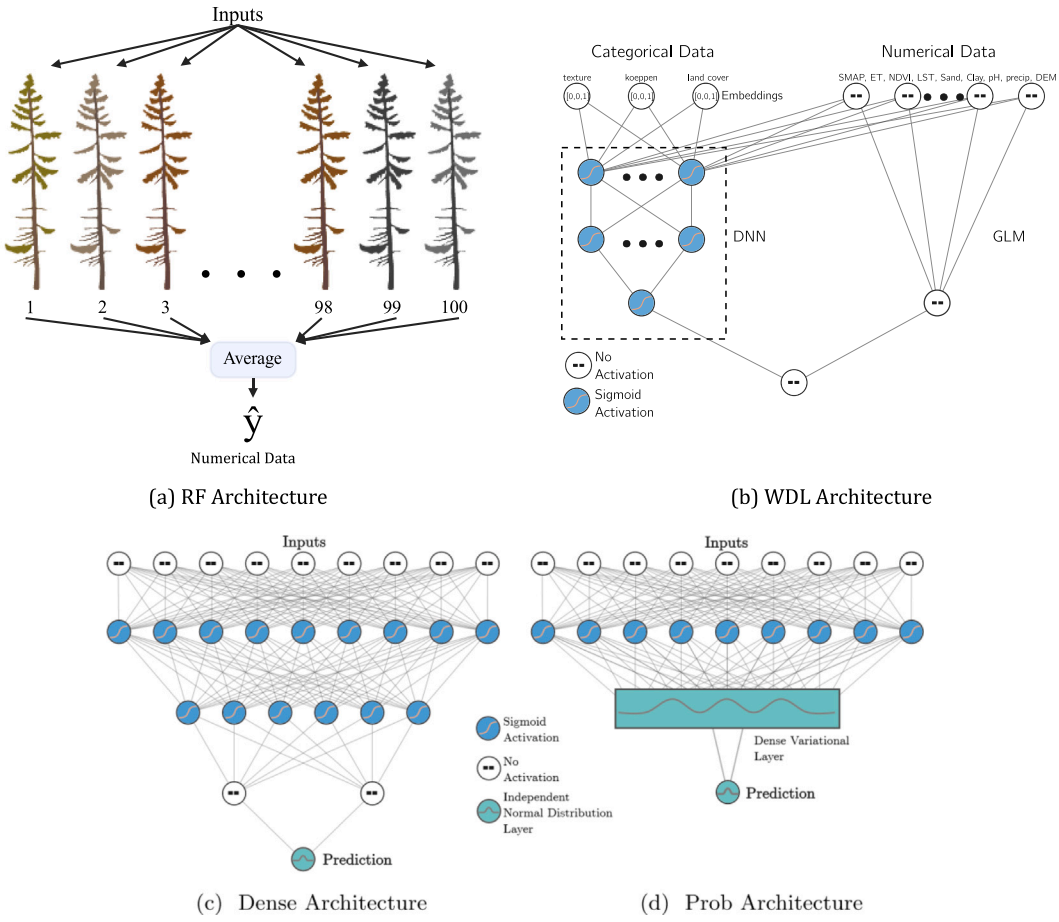


Fig. 2. Model architectures.

Table 1
Static categorical land variables and their classes.

Texture	Land Cover	Koeppen Climate Class
Loam	Grasslands	Dfb
Sandy loam	Savannahs	Cfa
Silt loam	Woody savannahs	BSk
Clay loam	Croplands	Dfc
Sandy clay loam	Deciduous broad-leaf forests	Csb
Silty clay loam	Open Shrublands	Dsb
Loamy sand	Evergreen needle-leaf forests	Csa
Sand	Mixed forests	Dfa
Clay	Barren	ET
N/A	Cropland/Vegetation Mosaic	Dsc
	Urban and Built-up	Bwk
	Evergreen Broad-leaf forests	Cfb
	Closed shrublands	Bwh
		Bsh
		Cfc
		Am
		Aw

Dense and prob

The Dense (Fig. 2(c)) and Prob (Fig. 2(d)) models follow a feed forward (or dense) architecture. They consist of just two or three hidden layers and a total parameter count of 164 and 150 parameters for Dense and Prob respectively. The output layer for both networks is an Independent Normal distribution layer. This layer outputs a single distribution which allows us to leverage the Negative Log Likelihood loss function. The Prob model utilizes a Bayesian layer (Dense Variational Layer) to learn a posterior distribution over the weights. This layer is discussed further in the supplement.

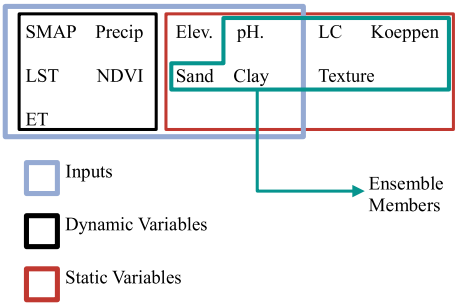


Fig. 3. Inputs comprise data seen by the models. The three variables not included as inputs (LC, Koeppen, Texture) are only used as references to apply weights to samples during training. This is an exception for the WDL model which also uses these variables as inputs (Fig. 2(b)).

3.2. Training

In this study, static variables and landscape characteristics are considered to either assist or hinder the models' ability to predict SWC. These static variables are not equally represented during training. For example, there are a lot more *field* land cover types than *forest*. Models might prioritize the most common class for a static variable, potentially overlooking less common ones. To counter this, each ensemble member is trained on samples weighted to emphasize imbalances in a specific static variable. An overview of all variables can be seen in Fig. 3. For the Dense, Prob, and WDL models, samples are weighted by the following variables: **texture**, **clay** and **sand content**, **pH**, **Köppen climate class**, **land cover class**, and **no weights**. This results in the

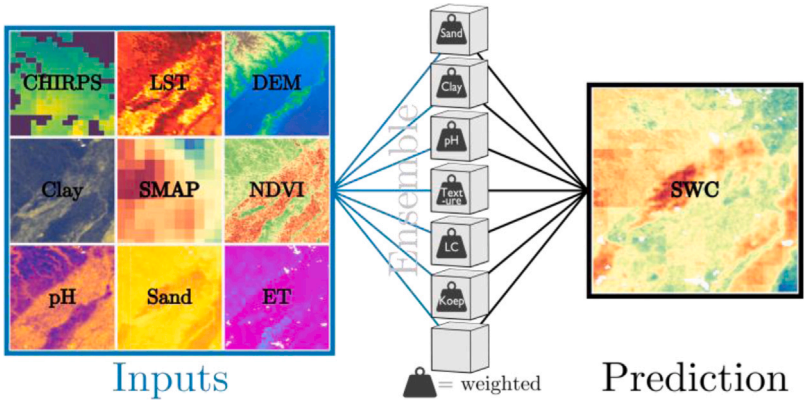


Fig. 4. Prediction regime for the Dense, Prob, and WDL ensembles. Each ensemble member (cube) is trained on samples weighted against imbalances in a static variable. These predictions are then averaged to provide an ensemble prediction.

training of seven ensemble members one for each of the previously mentioned variables. Each ensemble member receives all nine input variables (*Inputs* in Fig. 3) with samples weighted according to imbalances in static variables (Green outline in Fig. 3). After training, the predictions from each ensemble member are averaged to produce the final prediction for the entire ensemble. See Fig. 4

The weighting scheme for each static variable follows a "balanced" procedure, namely:

$$w_i = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_{\text{i}}}, \tag{1}$$

where w_i is the weight for class_{i}, n_{samples} is the total number of samples, n_{classes} is the total number of classes and n_{i} is the number of samples for class_{i}.

The RF model does not use sample weights. Instead, balance is accounted for by training a unique model for each soil texture domain as done by Abbaszadeh et al. (2019). The characteristics learned for each texture then contribute equally to the final prediction regardless of that textures representation in the dataset. This RF approach does not account for imbalances in other domains.

Temporal resolution

The models were trained on the 8-day composite LST product as this permitted more samples due to less gaps from cloud cover. Each sample uses padded or the last recorded LST composite temperature as it's daily value. This value could be, in the worst case scenario, out of date by 7 days. This is also the case for the ET product which also has an 8-day resolution. Although not ideal, it is assumed that SMAP accounts for the temporal variation in SWC while the other variables account for the spatial variation. Thus, these temporally coarse datasets are acceptable as long as their "description" of the spatial variability is consistent for that period. When comparing the daily vs. 8-day average, this loss of temporal information seems to be offset by the increase in samples to learn from. As seen in Fig. 5, training on the temporally coarse LST product yielded better metrics on both datasets. This is discussed further in the supplement document.

3.3. Predictions

For all models/ensembles, a prediction constitutes the average over all ensemble members. This can be represented by the following equation:

$$p(SM_d|C) = \frac{1}{M}\sum_{t=1}^{M} p_t(SM_d|C), \tag{2}$$

where $p(SM_d|C)$ is the downscaled ensemble posterior. This is derived from the average of the posterior predictions of M ensemble member models over covariate vector C (A stacked vector of input variables).

For spatial predictions, spatial data are resampled to the highest resolution (90 m) using nearest neighbor interpolation. This prevents interpolation error, but introduces some pixelation at higher levels of zoom. Pixelation can be eliminated if one wishes to spatially interpolate the input data to higher resolutions. This study does not use interpolated data so as to best display the underlying data structure.

Metrics

In order to assess the performance of the downscaling results, predictions will be evaluated on new spatial domains outside of the training dataset. The metrics used to assess the performance are *ubRMSE, R,* and *bias*.

$$Bias = E[(\theta_p - \theta_m)], \tag{3}$$

$$RMSE = \sqrt{E[(\theta_p - \theta_m)^2]}, \tag{4}$$

$$ubRMSE = \sqrt{RMSE^2 - bias^2}, \tag{5}$$

$$R = \frac{\sum_i^n (\theta_p - \bar{\theta_p})(\theta_m - \bar{\theta_m})}{\sqrt{\sum_i^n (\theta_p - \bar{\theta_p})^2 (\theta_m - \bar{\theta_m})^2}}, \tag{6}$$

where θ_p is the predicted value, θ_m is the measured or in-situ SWC value, and E represents the cumulative average.

Unbiased Root Mean Squared Error (*ubRMSE*) is the standard metric to evaluate SWC products employed by NASA. The SMAP mission considers an ubRMSE of less than 0.04 m^3/m^3 acceptable for a SWC product (Entekhabi, 2014). An ideal value for ubRMSE is 0. The Pearsons correlation coefficient, $R \in [-1, 1]$, shows linearity between changes in data points and is especially useful for time series analysis. For this study, an ideal value for R is 1. Lastly, bias detects whether a model overestimates (positive) or underestimates (negative) values compared to ground truth. An ideal value for bias is 0.

4. Results

Predictions were made on three datasets. The first is the validation data set aside during training. The second and third comprise smaller networks of soil moisture stations used to calibrate the SMAP measurements.

4.1. CONUS dataset

Because downscaling is an attempt at spatial prediction and reasoning, it is important that evaluations are done on new spatial areas. For this reason, all data in the validation dataset represents spatial domains previously unseen during training. This comprised ~20% of the sites available for each texture class.

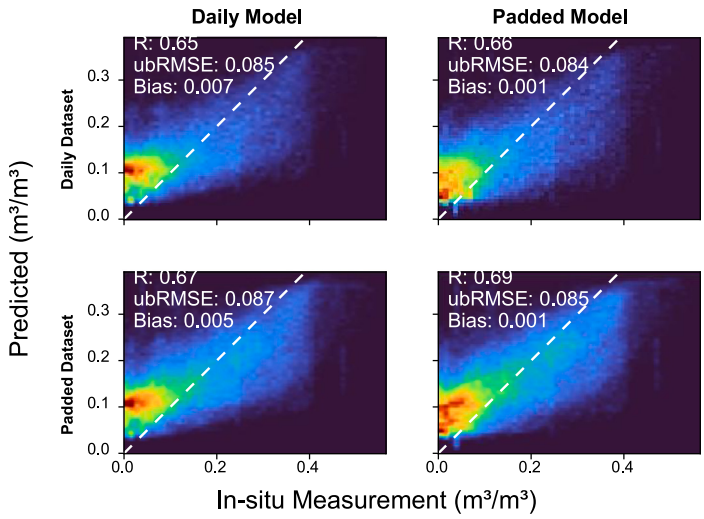


Fig. 5. Predictions for a model trained on a time-padded dataset which contains much more samples (658,000) to learn from and a model trained on a temporally accurate dataset (372,000). Both models predict on the validation sets for each dataset.

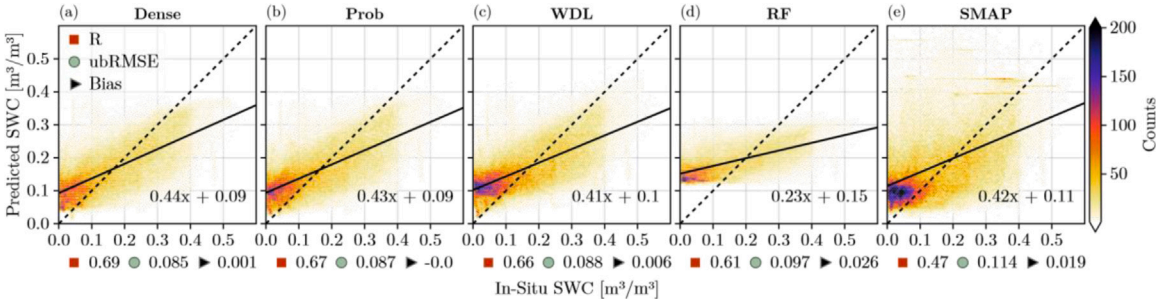


Fig. 6. Heatmaps and metrics for model predictions on the validation dataset as a whole.

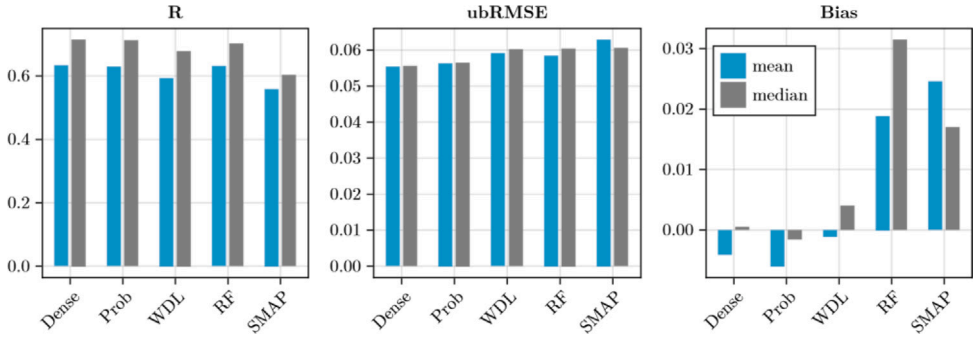


Fig. 7. The average metric score for individual sites in the validation dataset.

As shown in Fig. 6, every model was able to generalize over the entire dataset better than the raw SMAP values. The RF predictions are strongly biased with SWC measurements being squashed towards 0.18 m^3/m^3. Because of this, the lowest SWC prediction by the RF model on the entire dataset is 0.10 m^3/m^3. Although the RF output demonstrates a failure to capture the true variance of the dataset, this is not an unacceptable result as ubRMSE and R metrics are both invariant to bias. Thus, we can still observe spatial and temporal trends even with extreme biases. This does however diminish the value of RF predictions.

On a site to site level, all models surpass SMAP on every metric with exception to RFs median bias. This is displayed in Fig. 7. The figure also indicates that time series are less consistent across sites, with the mean R value being notably lower than the median. The ubRMSE shows a strong agreement between mean and median values demonstrating general consistency for prediction accuracy. Overall, this suggests all models and their predictions should be as reliable or more so than SMAP.

4.1.1. Spatial predictions

To compare the spatial predictions of each method, a 1° × 1° box is cut out around in-situ locations on a summer day with the least cloud cover. Of the resulting predictions, six examples that exhibit unique characteristics are highlighted, two of which are displayed in Fig. 8. Overall, the models tend to exhibit similar spatial patterns. In some cases, as exhibited in the predictions around *PBO: H2O_LITTLELOST*, the extra categorical inputs of the WDL model produce strong pixelation which create unpleasant and impractical outputs. Additionally the RF predictions show strong bias and little variability. Four additional

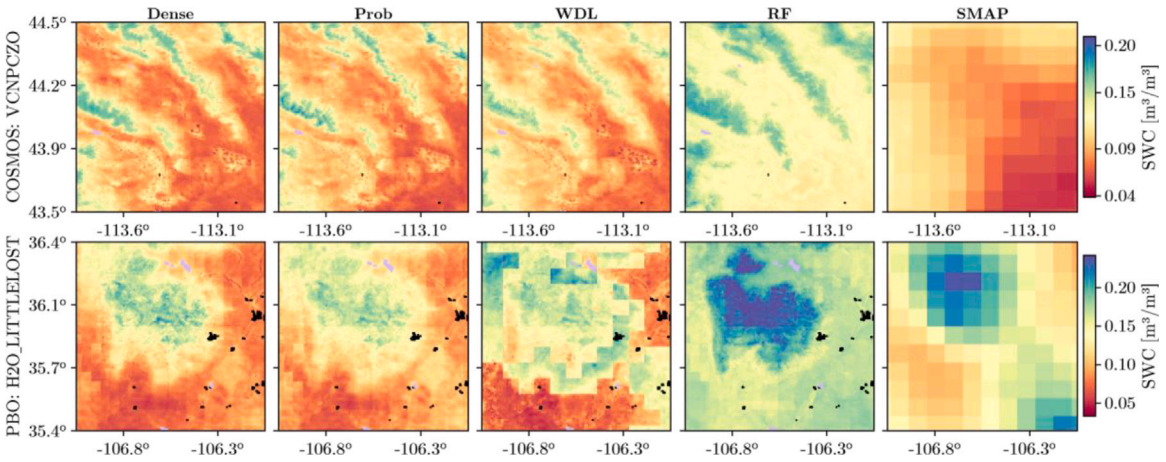


Fig. 8. 1° × 1° spatial SWC predictions of models vs. SMAP. Black pixels represent pixels masked as 'urban' and purple pixels are water surfaces.

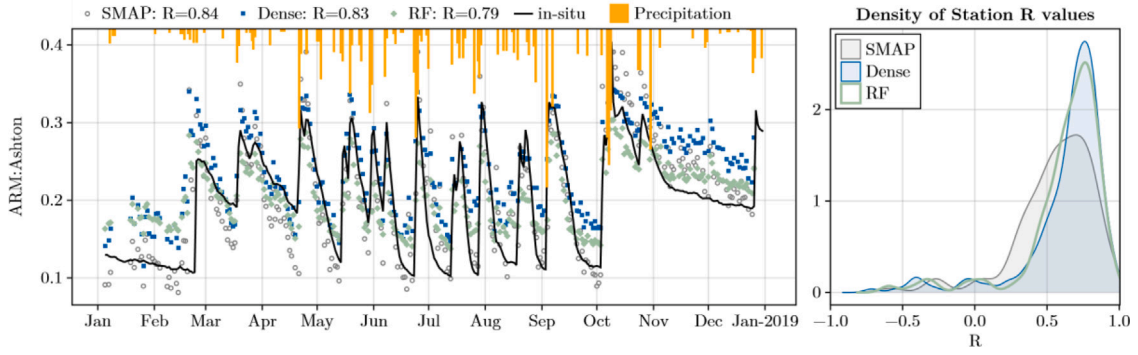


Fig. 9. (Left) Temporal predictions on a station in the validation dataset. (Right) Density plot of the R values for each station in the validation dataset.

examples can be found in the supplement document.

4.1.2. Temporal predictions

Temporal accuracy is measured with Pearson's R. This value is calculated for the timeseries of each site in the validation set is an acceptable guage of temporal accuracy. However, because R is invariant to bias, it is good to plot timeseries to observe prediction behavior. For this purpose, the ten sites with the most data were selected and a time-series for 2018 is plotted. One of which is seen in Fig. 9. The figure also displays the R scores for each station in the validation dataset. The two models with the strongest average R are plotted (Dense and RF). Both distributions have shifted to higher R values with mean, median, and every quantile being higher than their SMAP counterpart. Although RF's performance is comparable to Dense's, it exhibits a pronounced bias on sites with low SWC and often deviates from in-situ markers as shown in additional timeseries in the supplement. Overall, the time-series predictions from all models are on par with or exceed those of SMAP.

4.2. Oklahoma basin datasets

The Oklahoma Basin has two well-known neighboring regions of densely covered soil moisture networks. Not only were these networks used to calibrate SMAP (Entekhabi, 2014) but they are often used to assess downscaling efforts over a more localized region. The two regions, Fort Cobb and Washita River Basin, are comprised of 17 and 20 sites of retrievable data for the study period, respectively. All of these sites are located on loam soil texture according to soil grids data. The majority are classified as grasslands with a few cropland sites in Fort Cobb.

Table 2
The average metric score for individual sites in the Washita dataset.

	Dense	Prob	WDL	RF	SMAP
R	**0.752**	0.661	0.681	0.704	0.745
ubRMSE	**0.041**	0.062	0.046	0.044	0.046
Bias	0.053	0.248	0.076	**0.006**	0.011

Washita

The first dataset is the Washita River basin network. In this region, all models struggle on the Washita dataset as a whole as seen in Fig. 10. All models have a significant positive bias on the lower SWC readings with the Prob model having severely shifted predictions. The Prob model also is the only model that fails to outperform SMAP's ubRMSE score. Only the Dense model outperforms SMAP on 2/3 metrics.

Performance metrics improve significantly on individual sites as seen in Table 2. The Dense network performs well here with the best R score and the only ubRMSE to reach the 0.04 m^3/m^3 realm of acceptable values. The other models are unable to outperform SMAP measurements on a site to site level which can be seen further in tables of station data in the supplement document.

Fort Cobb

The second dataset consists of measurements from the Fort Cobb network, which, due to its proximity to Washita, exhibits similar trends. All models show poor overall fit to the dataset (Fig. 11) and a strong positive bias at low SWC measurements. The RF model has the best bias metric, likely because values are compressed towards a mean value.

On a site-by-site basis, model performance metrics improve (Table 3). The Dense model approaches the 0.04 m^3/m^3 ubRMSE threshold set by the SMAP mission with RF and WDL on the periphery

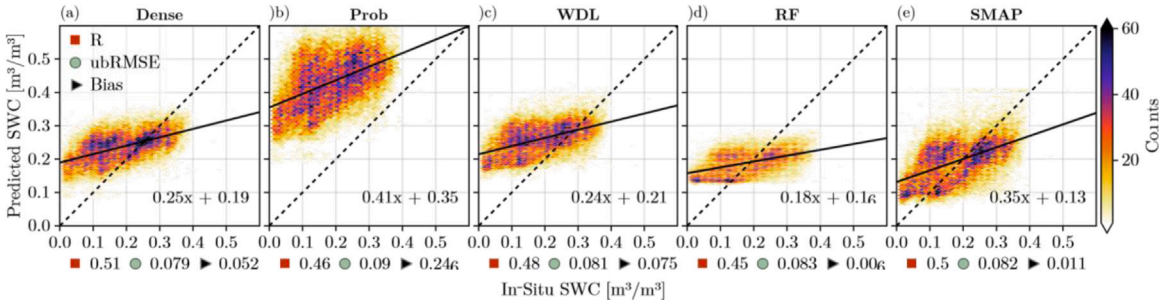


Fig. 10. Heatmaps and metrics for model predictions on the Washita dataset as a whole.

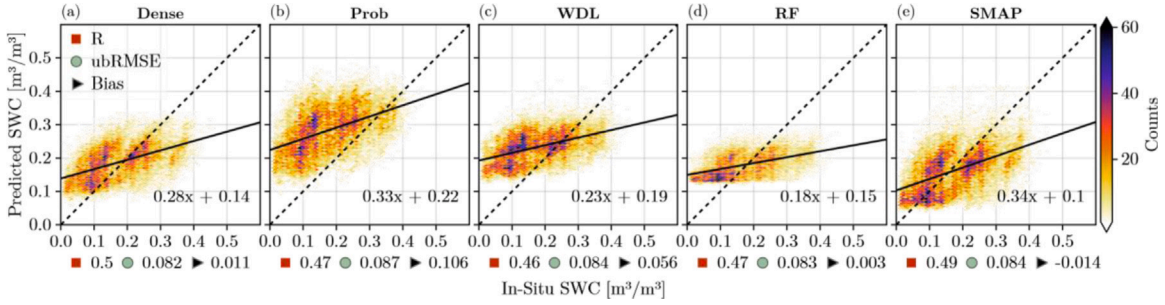


Fig. 11. Heatmaps and metric scores for model predictions on the Fort Cobb dataset as a whole.

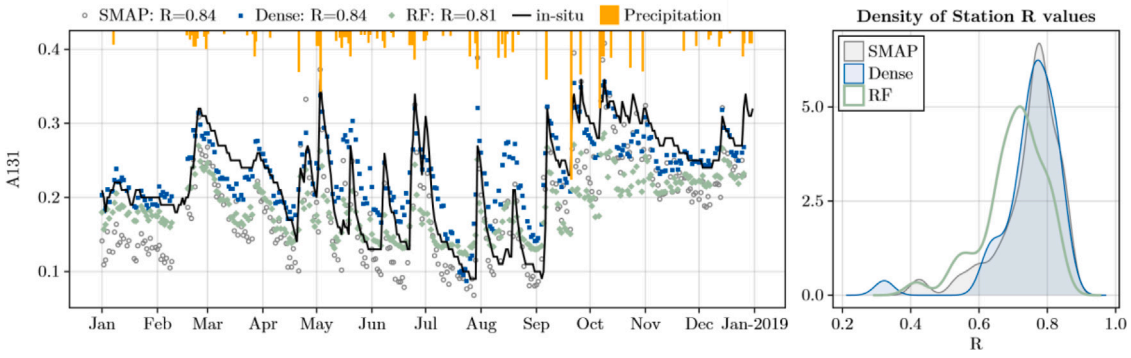


Fig. 12. (Left) Temporal predictions on a station in the Washita dataset. (Right) Density plot of the R values for each station in both OK datasets.

Table 3
The average metric score for individual sites in Fort Cobb dataset.

	Dense	Prob	WDL	RF	SMAP
R	0.748	0.708	0.673	0.709	**0.752**
ubRMSE	**0.042**	0.049	0.044	0.043	0.046
Bias	0.017	0.112	0.062	0.010	**−0.008**

for this metric. The Prob model does not surpass SMAP on any metric, with SMAP achieving the highest R score and best bias.

Because the Oklahoma Basin networks were used to calibrate the SMAP mission, we expect SMAP to exhibit one of it's strongest performances here. If a model can reliably match or outperform SMAP here, it would suggest confidence in it's ability to perform elsewhere. The Dense architecture is the only model to reliably match or exceed SMAP on key metrics on these datasets.

Timeseries

Only the Dense model is able to demonstrate parity and match SMAPs strong temporal accuracy. A timeseries of a station in the Washita dataset is plotted in Fig. 12 along with the density plot of the R values of all of the stations in both Oklahoma datasets. Here we can see that RF has a distribution shifted slightly to the left as it fails to match SMAPs performance. The Dense distributions peak (Q2) is lower than SMAPs (0.766 vs. 0.774) but the dense distribution has a higher mean (0.750 vs. 0.748) and both Q1 and Q3 quartiles have slightly higher values compared to SMAP.

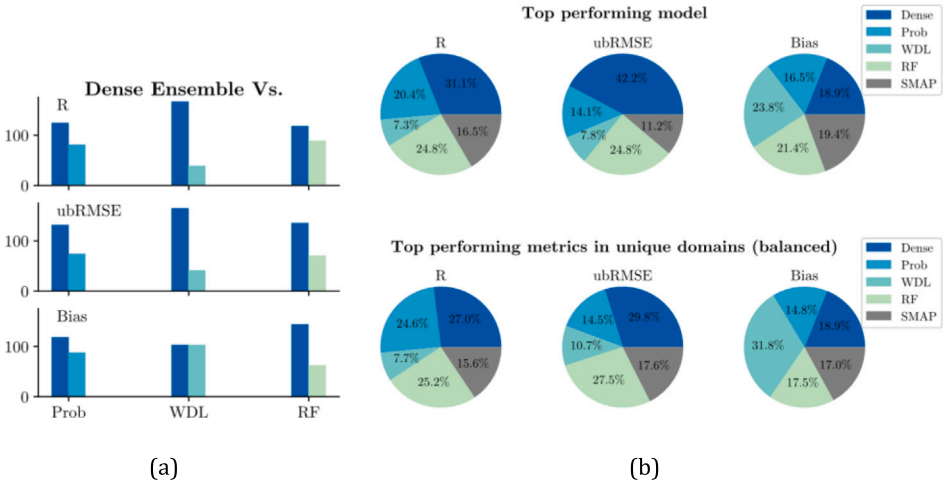
4.3. Top performer

Performance evaluation is based on three criteria: dataset, sites, and domains. The Dense model is the clear a top performer across datasets (Figs. 6, 10, 11). For site-level comparison, in a head-to-head competition, Dense outperforms all other models in every metric except for bias against WDL, as shown in Fig. 13(a).

To determine if Dense remains the top performer by domain, we assess each model's performance on sites categorized by land surface variables, as detailed in Table 1. Performance is normalized to ensure that over- or underrepresented classes equally influence the results. This normalization process is elaborated upon in subsequent sections. After normalization for class type and abundance, Fig. 13(b) reveals that Dense continues to be the most consistent performer for

Table 4
The deviations from mean values for static variables at the site level.

site	Sand	Clay	pH	Elev	Koep	LC
SCAN:Ku-nesa	−2.02	1.52	−0.00	−1.08	Cfa	Savannahs
USCRN:Manhattan-6-SSW	−1.88	1.52	0.58	−1.05	Cfa	Grasslands
FLUXNET-AMERIFLUX:BouldinIslandAlfalfa	−1.60	3.63	−0.12	−1.38	Csa	Croplands
FLUXNET-AMERIFLUX:BouldinIslandcorn	−1.52	3.14	−0.12	−1.39	Csa	Croplands
PBO_H2O:MOONEYCYN	−0.82	2.01	1.40	−0.98	Csb	Croplands
SCAN:ConradAgRc	−1.10	2.33	1.17	−0.31	BSk	Croplands
SCAN:ElsberryPMC	−2.09	0.39	0.11	−1.24	Cfa	Croplands
SCAN:Mayday	−1.38	2.17	−0.35	−1.35	Cfa	Croplands
SCAN:Moccasin	−0.82	1.84	0.93	−0.14	BSk	Croplands
USCRN:Versailles-3-NNW	−2.37	0.39	−0.24	−1.12	Cfa	Cropland/Vegetation Mosaic
Mean	**−1.56**	**1.89**	**0.34**	**−1.00**	–	–



(a) (b)

Fig. 13. (a) The Dense model against every other model. For each site one model outperforms the other, the value increases. (b) (**Top**) Percentage of stations where a model was the top performer for a given metric (**Bottom**) Each model predicts on all sites belonging to a specific class in Table 1. Each time a model outperforms every other model for a metric it gets a point. All points for that class are normalized so that the top performer receives one point for that class. All points are summed together for all classes. This produces an unbiased assessment of model performance regardless of imbalances in representation of classes.

R and ubRMSE, albeit slightly less dominant than the RF model. WDL maintains its position as the best model for bias.

Having a distance based model outperform RF has additional advantages. For starters the evaluation speed for distance based models is two orders of magnitude faster (0.16 s vs. 17.7 s on 130k samples). Therefore, it is more feasible to predict over large domains. Additionally, the file size of the RF model is three orders of magnitude larger (2.3 GB vs. 1.03 MB) which makes transferring it less convenient than the simple distance based models. For these reasons, it is unreasonable to continue using a RF architecture for this task at this resolution.

4.4. Areas of underperformance

As the Dense model was the strongest performer. It is important to find circumstances where it struggles. To do so, the static variables for each site in the CONUS dataset were compiled into a dataset with six normalized dimensions (sand, clay, pH, topography, climate class, land cover type). This dataset was then projected into 2D space using Principle Component Analysis (PCA). This reduction allows one to visualize the high-dimensional six static variables as a 2D image. The sites from the validation set are then plotted and colored if the Dense model failed to outperform SMAP's ubRMSE score at that site. The 2D projection shows a clear grouping in the box in Fig. 14. This area in the PCA represents Cropland land cover type with high clay content and low sand content as seen in Table 4. These values are scaled by the standard deviation of the dataset for each static variable. A value of −2.0, means two standard deviations below the mean. Some sites have very high clay content and others, like *USCRN:Versailles-3-NNW* and *SCAN:ElsberryPMC*, have very low sand content. More than two

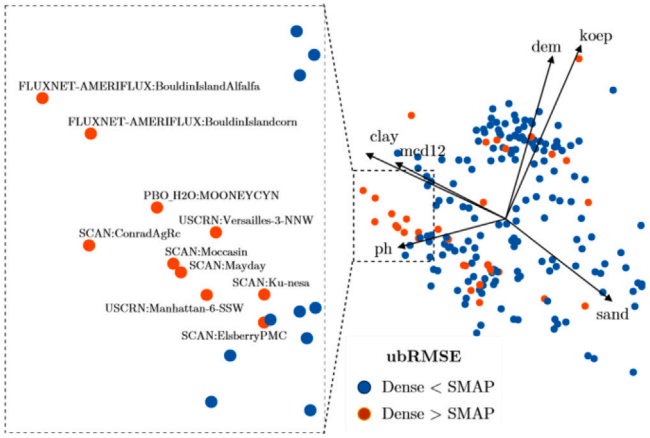


Fig. 14. Reprojection of test data static variables into PCA space. Orange dots represent sites where the Dense model's ubRMSE score was worse than SMAP.

standard deviations below the mean. Most of these sites are croplands.

This brief analysis shows the Dense model does not have consistent performance on croplands of high clay and low sand content values. Therefore, this model would not be an ideal representation of soil moisture in these conditions and should not be relied upon if a given use case should arise.

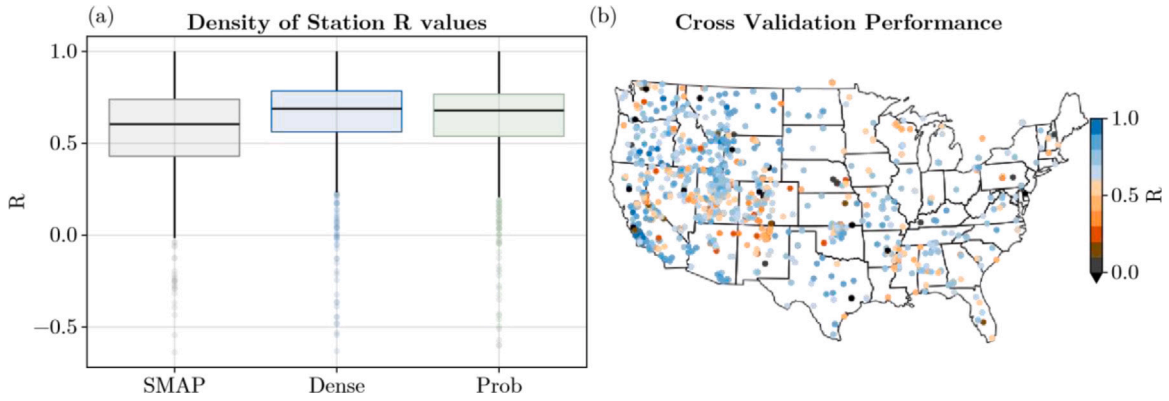


Fig. 15. (a) Box plots of the SMAP, Dense, and Prob R values for each station in the cross validation dataset. (b) Spatial distribution of R values on each station as predicted by Dense.

4.5. Cross-fold analysis

In order to assess whether our methodology is generalizable, a 10-fold cross validation was conducted. Using stratified random sampling, the original dataset is grouped into soil texture classes where the sites belonging to that texture class are divided into 10 groups of equivalent count. One of these groups from each texture class is moved into a validation dataset. If a texture class has less than 10 sites of data, (**Sand**(2), **Clay**(2), and **Undefined**(4)) half of the stations are randomly sampled. Models are then trained on the data not in the validation dataset.

The cross-validation metrics generally align with those from the validation set, with one exception: the Random Forest (RF) model. When comparing temporal consistency (R) with SMAP, the Dense and Probabilistic models — our top performers — show improved results. Each quantile of their distribution is higher than that of SMAP (as shown in Fig. 15). Neither were able to sufficiently correct outliers and as a result each have a higher kurtosis than the SMAP distribution. The Wide and Deep Learning (WDL) model exhibited a similar trend, which can be found in the supplementary materials.

The Random Forest (RF) model's performance unexpectedly declined in the cross-validation compared to the validation dataset. Its average temporal consistency (R) showed only marginal improvement over SMAP. This drop is surprising because each cross-validation training set included all texture classes, similar to the validation dataset. The cause of this significant performance decrease remains unclear. However, we observed that nearly all forests in the ensemble (except the forest trained on Sandy Clay Loam data) showed less variability (lower standard deviation) in their predictions compared to the forests trained on the original dataset.

One possible explanation could be related to the amount of training data used. The cross-validation forests were trained on approximately 90% of the data, whereas the original forests used about 80%. This additional data might have led to some form of overfitting. However, we did not conduct further analysis to confirm this hypothesis or identify other potential causes.

The cross validation and additional plots found in the supplement appear to confirm that the weighting scheme for distance-based models limits biases in the training data and that model predictions remain robust on unseen locations when compared to SMAP (see Table 5).

Table 5
The mean metric score for each model on each station on the validation set vs. the cross validation dataset.

Model	Dataset	R	ubRMSE	Bias
Dense	Val	0.632	**0.055**	−0.004
	Cross Val	**0.638**	0.058	**−0.002**
Prob	Val	**0.628**	**0.056**	−0.007
	Cross Val	0.620	0.060	**−0.003**
WDL	Val	0.594	**0.059**	**−0.001**
	Cross Val	**0.608**	0.061	−0.004
RF	Val	**0.630**	**0.058**	0.019
	Cross Val	0.564	0.066	0.019
SMAP	Val	0.559	**0.063**	0.025
	Cross Val	**0.562**	0.065	**0.023**

5. Discussion

The primary focus for this section is to evaluate the robustness and generalizability of model performance. Additionally, we want to assess the ensemble framework and identify whether or not there is any advantage from an ensemble prediction.

It is worth noting the existence of SMAP-HydroBlocks (Vergopolan et al., 2021), a high-resolution (30 m) surface soil moisture product for CONUS. While SMAP-HydroBlocks employs a sophisticated physical model incorporating detailed hydrological processes and high-resolution regional datasets, our approach differs in its use of simple machine learning ensembles and globally available datasets. This distinction makes our method potentially more adaptable for global applications, particularly in regions where high-resolution local data may be limited. However, for studies focused within the United States, SMAP-HydroBlocks likely offers superior fidelity due to its incorporation of detailed local information.

5.1. Generalizability

Large-scale domain predictions are valuable only if they are consistently accurate across the domain's heterogeneity. To test our model predictions' generalizability, we:

1. Validated using data from previously unseen locations.
2. Conducted cross-fold analysis across all sites in the training and validation sets.
3. Monitored spatial predictions and their associated SHAP values.

Results showed consistent performance on unseen sites during training. SHAP values generally aligned with literature expectations, except for an unexpected inverse relationship with NDVI across all models. This anomaly warrants further investigation. Detailed analysis of spatial predictions and SHAP values is available in the supplement.

Results from these analyses demonstrate the generalizability of using ensembles of simple ML architectures for downscaling SWC at sub-km resolutions.

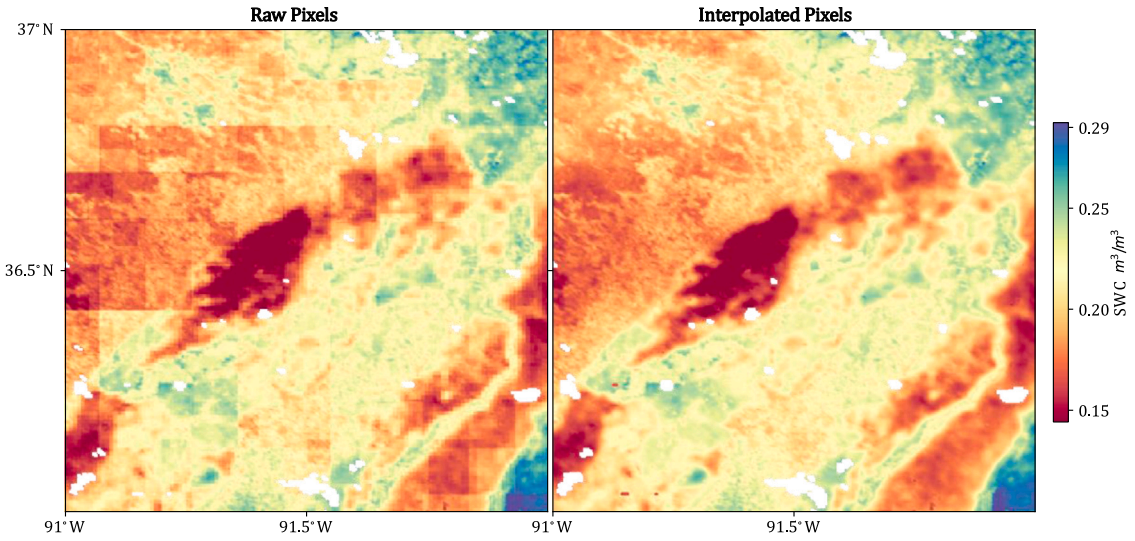


Fig. 16. Spatial predictions can be smoothen if input data is interpolated. This increases fidelity and continuity but can introduce interpolation error.

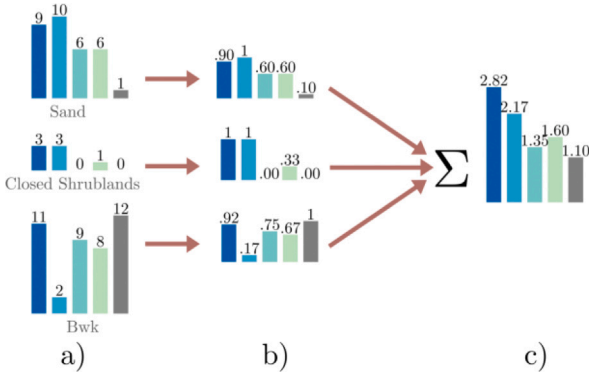


Fig. 17. Weighting schema for unbiased top performers. (a) All models predict on all sites belonging to a specific category. Each time a model outperforms every other model it gets a point. (b) Points are then normalized. This ensures under-represented categories have equal importance in assessing model performance. (c) The normalized points are summed providing a final assessment of model performance on all categories.

5.2. Interpolation

Although the finest spatial resolution is 90 m (elev), the native resolution of other variables are coarser. As a result, plotting the pixels of coarser data on this finer data creates a "true" prediction but introduces some harsh boundaries or pixelation. One can completely eliminate this pixelation if all input data is interpolated to the highest resolution. As mentioned, this was not done in this study to preserve the structure of the data. However, predictions using interpolated data present high-fidelity and continuous predictions (Fig. 16).

5.3. Ensemble advantage

This study assesses the feasibility and advantages of using model ensembles to predict Soil Water Content (SWC) at higher resolutions. The rationale behind these ensembles is to ensure equal representation of all unique land characteristics during training, potentially preventing overfitting to dominant features. However, the necessity of this approach remains in question. We begin by comparing the ensemble performance against the performance of each ensemble member in the validation dataset (Table 6):

1. The Dense ensemble shows only marginal improvement over its unweighted member.
2. For both Prob and WDL ensembles, the Sand and Clay weighted members outperformed their respective ensembles.
3. In all cases, the ensembles' average performance is not significantly better than the unweighted member.

To ensure an unbiased comparison, we evaluate performance using static variables in a head-to-head competition. Scores are normalized by class abundance to account for varying sample sizes. The process is as follows:

1. For each texture, land cover, and Koeppen class listed in Table 1, we compare individual ensemble members against the full ensemble.
2. Each time a model outperforms the other for a given site, its score for that class increases. Scores for each class are normalized, with the model performing best on most sites receiving a value of 1.
3. We sum these normalized scores to obtain a total normalized performance ratio for each ensemble vs. ensemble member pairing.

This weighting schema is illustrated in Fig. 17, and the resulting performance ratios are visualized in Fig. 18.

Again we see the same trend with no clear ensemble advantage across all of it's members. Each ensemble achieves parity or is outperformed by an ensemble member at least once. The Dense/Prob architectures are likely too simple and the GLM of the WDL seems adept enough at guiding predictions that any overfitting of training data is negligible. From a purely numerical context, there does not exist a clear ensemble advantage. However, for the Dense/Prob architectures, there is a clear advantage for timeseries predictions (R).

The RF ensemble has a dominant ensemble advantage due to the nature of how it was trained. This is discussed further in the supplement.

Lastly, we compared the spatial predictions of the ensemble to those of the top-performing ensemble member. Our analysis revealed that the Dense ensemble predictions appear to capture more land surface characteristics than the single ensemble member, as illustrated in Fig. 19. While not directly quantifiable, it is evident that the Ensemble incorporates more land surface features into its prediction than the unweighted member.

However, this observation does not hold true for the Prob architecture. In this case, the ensemble member achieved similar land characteristic fidelity to the ensemble prediction. For the WDL architecture, the ensemble member prediction shows more noise than the

Table 6
Average performance for each ensemble member and the ensemble as a whole on all sites in the validation dataset. Bold indicates top performer for that metric.

Model	Metric	Ens.	Sand	Clay	Koep	Land Cvr	Free	pH	Texture
Dense	R	**0.632**	0.621	0.615	0.607	0.618	0.631	0.613	0.558
	ubRMSE	**0.055**	0.056	0.056	0.058	0.057	**0.055**	0.057	0.058
	Bias	−0.004	**−0.000**	−0.001	−0.001	−0.019	−0.003	−0.006	0.001
Prob	R	**0.629**	**0.629**	0.620	0.592	0.618	0.623	0.613	0.596
	ubRMSE	**0.056**	**0.056**	0.057	0.059	0.057	**0.056**	0.057	0.059
	Bias	−0.007	**−0.004**	**−0.004**	−0.011	−0.008	−0.007	−0.006	**−0.004**
WDL	R	0.594	0.594	**0.598**	0.586	0.594	0.594	0.586	0.589
	ubRMSE	**0.059**	**0.059**	**0.059**	0.060	**0.059**	**0.059**	0.060	**0.059**
	Bias	−0.001	−0.004	−0.002	0.002	−0.006	−0.002	**0.000**	0.003

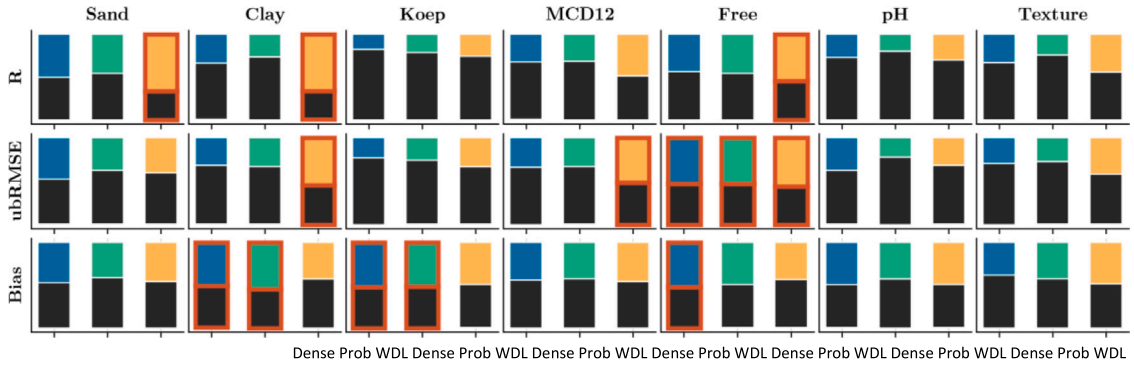


Fig. 18. Head to head comparison of Ensembles (Bottom label) vs. their member constituents (Top label) with normalized performances. Bars highlighted in red indicate an instance where an ensemble member outperformed the ensemble on that metric (Left label). An explanation of this head to head competition is seen in Fig. 17.

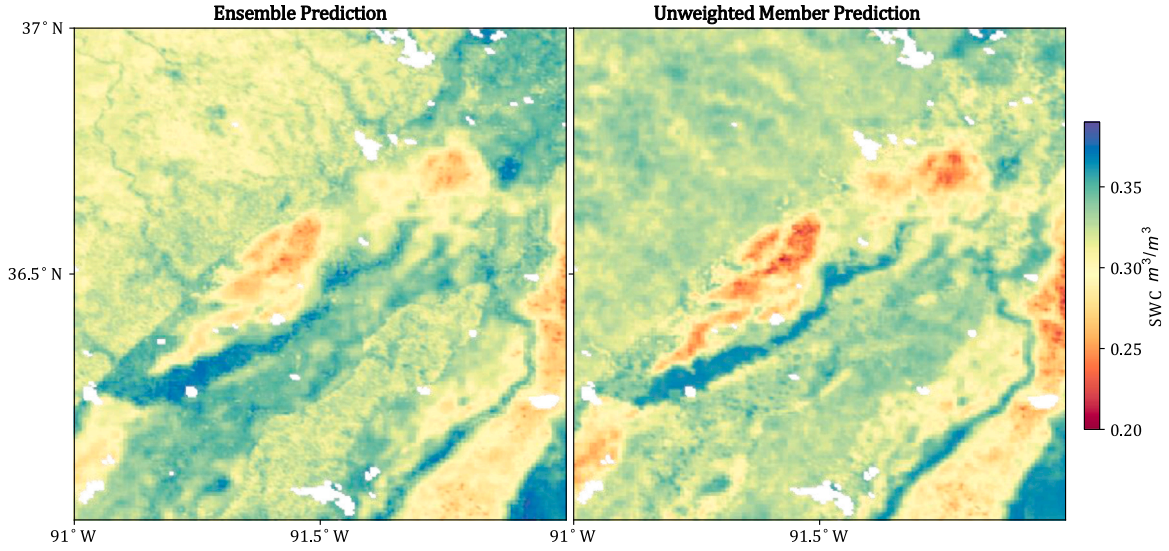


Fig. 19. Spatial Predictions comparing the Dense ensemble vs. the unweighted (Free) ensemble member.

ensemble but retains similar details. These comparisons can be found in the supplement.

Further analysis is needed to determine whether these differences constitute a substantial improvement of one approach over the other.

6. Conclusion

The work conducted in this paper served to demonstrate that an ensemble of simple ML architectures can acceptably downscale SWC. Models can reliably predict SWC with strong generalizability. However, certain ensemble members can outperform or achieve parity with the full ensemble on the validation dataset. This suggests that an ensemble is unnecessary and the same generlizability can be achieved with more rigorous weighting during training. However, for the top performing ensemble/model, the ensemble captured more of the land characteristics than its top performing single member. More analysis is needed to assess whether or not this is advantageous and by how much.

Multi-variable analysis of model predictions suggest the top performing model struggles on croplands with higher than average clay and silt content. This model cannot reliably outperform SMAP readings in these areas.

Training conducted with time-padded data benefits the performance more than the temporal inaccuracies of these readings hinder the training process. This suggests that models rely on SMAP to describe the temporal evolution of SWC, while using higher spatial resolution data to modulate SWC based on land characteristics.

Overall, all models were able to outperform SMAP on the validation and cross-fold datasets. The only exception being the RF model failing to achieve similar performance on both the validation and cross-validation datasets.

Final summary:

- Ensembles of simple ML architectures can downscale SWC predictions to sub 1 km resolutions
- Ensemble members can outperform or match the performance of these ensembles on datasets. However, the spatial predictions of some ensembles can capture more of the land characteristics than the ensemble member and reduces noise.
- Training the models on temporally padded data provides more benefits than drawbacks in terms of overall performance.
- The top performing model is unreliable on croplands with higher than average clay and lower than average sand content.

CRediT authorship contribution statement

Jeran Poehls: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lazaro Alonso:** Writing – review & editing, Visualization, Resources. **Sujan Koirala:** Writing – review & editing. **Markus Reichstein:** Supervision, Funding acquisition. **Nuno Carvalhais:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors of this paper have no conflicts of interest regarding the research conducted in this study.

Acknowledgments

Funding for this study was provided by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). The work was supported by colleagues at the Max Planck Institute for Biogeochemistry. This research could not be possible without the free and accessible data from Ameriflux, NASA's MODIS Mission, The International Soil Moisture Network, and the International Soil Reference and Information Centre.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jhydrol.2024.132624.

Data availability

Data is available in the supplement.

References

Abbaszadeh, Peyman, Moradkhani, Hamid, Zhan, Xiwu, 2019. Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. Water Resour. Res. 55 (1), 324–344, URL https://onlinelibrary.wiley.com/doi/10.1029/2018WR023354.

Abowarda, Ahmed Samir, Bai, Liangliang, Zhang, Caijin, Long, Di, Li, Xueying, Huang, Qi, Sun, Zhangli, 2021. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. Remote Sens. Environ. 255, 112301, URL https://linkinghub.elsevier.com/retrieve/pii/S0034425721000195.

Bittelli, Marco, 2011. Measuring soil water content: A review. HortTechnology 21 (3), 293–300, URL https://journals.ashs.org/view/journals/horttech/21/3/article-p293.xml.

Boden, T.A., Krassovski, M., Yang, B., 2013. The AmeriFlux data activity and data system: an evolving collection of data management techniques, tools, products and services. Geosci. Instrum. Methods Data Syst. 2 (1), 165–176, URL https://gi.copernicus.org/articles/2/165/2013/.

Cai, Yulin, Fan, Puran, Lang, Sen, Li, Mengyao, Muhammad, Yasir, Liu, Aixia, 2022. Downscaling of SMAP soil moisture data by using a deep belief network. Remote Sens. 14 (22), 5681, URL https://www.mdpi.com/2072-4292/14/22/5681.

Cheng, Heng-Tze, Koc, Levent, Harmsen, Jeremiah, Shaked, Tal, Chandra, Tushar, Aradhye, Hrishi, Anderson, Glen, Corrado, Greg, Chai, Wei, Ispir, Mustafa, Anil, Rohan, Haque, Zakaria, Hong, Lichan, Jain, Vihan, Liu, Xiaobing, Shah, Hemal, 2016. Wide & deep learning for recommender systems. http://dx.doi.org/10.48550/ARXIV.1606.07792, URL https://arxiv.org/abs/1606.07792. Publisher: arXiv Version Number: 1.

Colliander, Andreas, Fisher, Joshua B., Halverson, Gregory, Merlin, Olivier, Misra, Sidharth, Bindlish, Rajat, Jackson, Thomas J., Yueh, Simon, 2017. Spatial downscaling of SMAP soil moisture using MODIS land surface temperature and NDVI during SMAPVEX15. IEEE Geosci. Remote Sens. Lett. 14 (11), 2107–2111, URL http://ieeexplore.ieee.org/document/8060985/.

Dorigo, W.A., 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. Hydrol. Earth Syst. Sci. 15 (5), 1675–1698, URL https://hess.copernicus.org/articles/15/1675/2011/.

Dorigo, W.A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, A., Sanchis-Dufau, A.D., Zamojski, D., Cordes, C., Wagner, W., Drusch, M., 2013. Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network. Vadose Zone Journal 12 (3), 1–21, https://acsess.onlinelibrary.wiley.com/doi/10.2136/vzj2012.0097.

Entekhabi, Dara, 2014. SMAP Handbook Soil Moisture Active Passive. JPL Publication JPL.

Fu, Zheng, Ciais, Philippe, Prentice, I. Colin, Gentine, Pierre, Makowski, David, Bastos, Ana, Luo, Xiangzhong, Green, Julia K., Stoy, Paul C., Yang, Hui, Hajima, Tomohiro, 2022. Atmospheric dryness reduces photosynthesis along a large range of soil water deficits. Nature Commun. 13 (1), 989, URL https://www.nature.com/articles/s41467-022-28652-7.

Funk, Chris, Peterson, Pete, Landsfeld, Martin, Pedreros, Diego, Verdin, James, Shukla, Shraddhanand, Husak, Gregory, Rowland, James, Harrison, Laura, Hoell, Andrew, Michaelsen, Joel, 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci. Data 2 (1), 150066, URL https://www.nature.com/articles/sdata201566.

Hengl, Tomislav, Mendes De Jesus, Jorge, Heuvelink, Gerard B.M., Ruiperez Gonzalez, Maria, Kilibarda, Milan, Blagotić, Aleksandar, Shangguan, Wei, Wright, Marvin N., Geng, Xiaoyuan, Bauer-Marschallinger, Bernhard, Guevara, Mario Antonio, Vargas, Rodrigo, MacMillan, Robert A., Batjes, Niels H., Leenaars, Johan G.B., Ribeiro, Eloi, Wheeler, Ichsani, Mantel, Stephan, Kempen, Bas, 2017. SoilGrids250m: Global gridded soil information based on machine learning. In: Bond-Lamberty, Ben (Ed.), PLoS ONE 12 (2), e0169748, URL https://dx.plos.org/10.1371/journal.pone.0169748.

Koné, Brahima, Diedhiou, Arona, Diawara, Adama, Anquetin, Sandrine, Touré, N'datchoh Evelyne, Bamba, Adama, Kobea, Arsene Toka, 2022a. Influence of initial soil moisture in a regional climate model study over West Africa – Part 1: Impact on the climate mean. Hydrol. Earth Syst. Sci. 26 (3), 711–730, URL https://hess.copernicus.org/articles/26/711/2022/.

Koné, Brahima, Diedhiou, Arona, Diawara, Adama, Anquetin, Sandrine, Touré, N'datchoh Evelyne, Bamba, Adama, Kobea, Arsene Toka, 2022b. Influence of initial soil moisture in a regional climate model study over West Africa – Part 2: Impact on the climate extremes. Hydrol. Earth Syst. Sci. 26 (3), 731–754, URL https://hess.copernicus.org/articles/26/731/2022/.

Liu, Laibao, Gudmundsson, Lukas, Hauser, Mathias, Qin, Dahe, Li, Shuangcheng, Seneviratne, Sonia I., 2020. Soil moisture dominates dryness stress on ecosystem production globally. Nature Commun. 11 (1), 4892, URL https://www.nature.com/articles/s41467-020-18631-1.

Montzka, Carsten, Rötzer, Kathrina, Bogena, Heye, Sanchez, Nilda, Vereecken, Harry, 2018. A new soil moisture downscaling approach for SMAP, SMOS, and ASCAT by predicting sub-grid variability. Remote Sens. 10 (3), 427, URL http://www.mdpi.com/2072-4292/10/3/427.

Naz, Bibi S., Kurtz, Wolfgang, Montzka, Carsten, Sharples, Wendy, Goergen, Klaus, Keune, Jessica, Gao, Huilin, Springer, Anne, Hendricks Franssen, Harrie-Jan, Kollet, Stefan, 2019. Improving soil moisture and runoff simulations at 3 km over Europe using land surface data assimilation. Hydrol. Earth Syst. Sci. 23 (1), 277–301, URL https://hess.copernicus.org/articles/23/277/2019/.

Ojha, Nitu, Merlin, Olivier, Molero, Beatriz, Suere, Christophe, Olivera-Guerra, Luis, Ait Hssaine, Bouchra, Amazirh, Abdelhakim, Al Bitar, Ahmad, Escorihuela, Maria, Er-Raki, Salah, 2019. Stepwise disaggregation of SMAP soil moisture at 100 m resolution using landsat-7/8 data and a varying intermediate Resolution. Remote Sens. 11 (16), 1863, URL https://www.mdpi.com/2072-4292/11/16/1863.

Ojha, Nitu, Merlin, Olivier, Suere, Christophe, Escorihuela, Maria José, 2021. Extending the spatio-temporal applicability of DISPATCH soil moisture downscaling algorithm: A study case using SMAP, MODIS and sentinel-3 data. Front. Environ. Sci. 9, 555216, URL https://www.frontiersin.org/articles/10.3389/fenvs.2021.555216/full.

ONeill, Peggy E., Chan, Steven, Njoku, Eni G., Jackson, Tom, Bindlish, Rajat, 2019. SMAP enhanced L3 radiometer global daily 9 km EASE-grid soil moisture, Version 3. http://dx.doi.org/10.5067/T90W6VRLCBHI, URL https://nsidc.org/data/SPL3SMP_E/versions/3.

Sánchez, Juan M., Galve, Joan M., González-Piqueras, José, López-Urrea, Ramón, Niclòs, Raquel, Calera, Alfonso, 2020. Monitoring 10-m LST from the Combination MODIS/Sentinel-2, validation in a high contrast semi-arid agroecosystem. Remote Sens. 12 (9), 1453, URL https://www.mdpi.com/2072-4292/12/9/1453.

Stocker, Benjamin D., Zscheischler, Jakob, Keenan, Trevor F., Prentice, I. Colin, Seneviratne, Sonia I., Peñuelas, Josep, 2019. Drought impacts on terrestrial primary production underestimated by satellite monitoring. Nat. Geosci. 12 (4), 264–270, URL http://www.nature.com/articles/s41561-019-0318-6.

Vergopolan, Noemi, Chaney, Nathaniel W., Pan, Ming, Sheffield, Justin, Beck, Hylke E., Ferguson, Craig R., Torres-Rojas, Laura, Sadri, Sara, Wood, Eric F., 2021. SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. Sci. Data 8 (1), 264, URL https://www.nature.com/articles/s41597-021-01050-2.

Vergopolan, Noemi, Sheffield, Justin, Chaney, Nathaniel W., Pan, Ming, Beck, Hylke E., Ferguson, Craig R., Torres-Rojas, Laura, Eigenbrod, Felix, Crow, Wade, Wood, Eric F., 2022. High-resolution soil moisture data reveal complex multi-scale spatial variability across the united states. Geophys. Res. Lett. 49 (15), e2022GL098586, URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022GL098586.

Wu, Kaijun, Rodriguez, Gabriela Arambulo, Zajc, Marjana, Jacquemin, Elodie, Clément, Michiels, De Coster, Albéric, Lambot, Sébastien, 2019. A new drone-borne GPR for soil moisture mapping. Remote Sens. Environ. 235, 111456, URL https://linkinghub.elsevier.com/retrieve/pii/S0034425719304754.

Xu, Mengyuan, Yao, Ning, Yang, Haoxuan, Xu, Jia, Hu, Annan, Gustavo Goncalves de Goncalves, Luis, Liu, Gang, 2022. Downscaling SMAP soil moisture using a wide & deep learning method over the continental United States. J. Hydrol. 609, 127784, URL https://linkinghub.elsevier.com/retrieve/pii/S0022169422003596.

Xu, Wei, Zhang, Zhaoxu, Long, Zehao, Qin, Qiming, 2021. Downscaling SMAP soil moisture products with convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 4051–4062, URL https://ieeexplore.ieee.org/document/9390292/.

Yamazaki, Dai, Ikeshima, Daiki, Tawatari, Ryunosuke, Yamaguchi, Tomohiro, O'Loughlin, Fiachra, Neal, Jeffery C., Sampson, Christopher C., Kanae, Shinjiro, Bates, Paul D., 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44 (11), 5844–5853, URL https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017GL072874.

Zhao, Hongfei, Li, Jie, Yuan, Qiangqiang, Lin, Liupeng, Yue, Linwei, Xu, Hongzhang, 2022. Downscaling of soil moisture products using deep learning: Comparison and analysis on Tibetan Plateau. J. Hydrol. 607, 127570, URL https://linkinghub.elsevier.com/retrieve/pii/S0022169422001457.

Zheng, Jingyao, Lü, Haishen, Crow, Wade T., Zhao, Tianjie, Merlin, Olivier, Rodriguez-Fernandez, Nemesio, Shi, Jiancheng, Zhu, Yonghua, Su, Jianbin, Kang, Chuen Siang, Wang, Xiaoyi, Gou, Qiqi, 2021. Soil moisture downscaling using multiple modes of the DISPATCH algorithm in a semi-humid/humid region. Int. J. Appl. Earth Obs. Geoinf. 104, 102530, URL https://linkinghub.elsevier.com/retrieve/pii/S0303243421002373.