Downscaling Soil Moisture to Sub-km Resolutions with Simple Machine Learning Ensembles

Jeran Poehls[1], Lazaro Alonso[1], Sujan Koirala[1], Markus Reichstein[1,2], and Nuno Carvalhais[1,3,4]

1Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany
2 German Centre for Integrative Biodiversity Research (iDiv) Halle–Jena–Leipzig, Leipzig, Germany
3ELLIS Unit Jena, Jena, Germany
4CENSE, Departamento de Ciˆencias e Engenharia do Ambiente, Faculdade de Ciˆencias e Tecnologia, Universidade NOVA de Lisboa, Caparica, Portugal

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 ensembles use a dataset of over 1,000 sites across the US to predict soil moisture at sub-km scales. We find that all ensembles, particularly one with a very simple structure, can outperform SMAP on a cross-fold analysis of the 1,000+ sites. This ensemble has an average ubRMSE of **0.058** vs SMAPs **0.065** and an average R of **0.639** vs SMAPs **0.562**. Not all ensembles are beneficial, with some architectures performing better with different training weights than with ensemble averaging. However, some ensemble architectures 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.

Keywords

Ensemble, Soil Moisture, Remote Sensing, Downscaling, SMAP

*coorespondence : jpoehls@bgc-jena.mpg.de ; Hans-Kn¨oll-Straße 10, 07745 Jena, DE

Contents

1 Introduction

The water in the soil or soil water content (SWC) has a strong coupling with ecosystem stress and production[1][2][3]. 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[4]. 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[5]. But for truly global scale soil moisture mapping we need to look for the aid of satellites.

The Soil Moisture Active Passive (SMAP) radar 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[6]. The combination of these two would yield native SWC measurements at 9km per pixel and interpolated 1-3km 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 36km per pixel. This resolution can be increased using the Backus-Gilbert optimal interpolation algorithm to 9km per pixel with acceptable accuracy[7]. 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 100m-3km. Since, even at 1km resolution, up to 80% of SWC variability is lost[8]. At native satellite resolutions, there is a complete loss of SWC variability[8]. The 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 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 modelling[9][10][11]. 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[12][13][14]. 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[15][16]. But higher resolution LST data wouldn't 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[17][18][19][20].

However, a common occurrence are complex model architectures over particularly limited study areas[21][22][23].

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 ensemble architectures for downscaling coarse spatial resolution soil moisture data to sub-km resolutions. The four ensembles include: two probabilistic estimators consisting of simple neural networks, a wide-deep learning (WDL) architecture modelled after the work of Xu et al. 2022, and a random forest (RF) model. These ensembles 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 methods 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 coarse resolution 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 that allow accurate and generalizable predictions. Therefore, 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 deciding which variables to include in the dataset. After a process of feature selection that is covered in the supplemental document, a dataset comprised of the following variables was assembled: *SMAP, NDVI, LST, Precipitation, Sand* and *Clay content, pH, Evapotranspiration,* and *Topography/Elevation*.

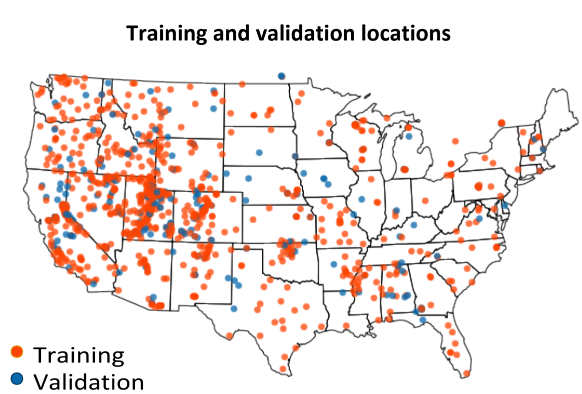


Figure 1: For this study, data within a temporal period extending from **January 1st, 2017** through **December 31st, 2021** was selected. This period ensured that soil moisture readings would include seasonal and, potentially, yearly variability.

This dataset was then iteratively trained over while excluding one of these variables. 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 folows:

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

Next we will discuss the sources used for this data.

2.1 Soil Moisture Active Passive (SMAP) Satellite Readings

The remotely sensed soil moisture readings are provided by NASAs SMAP satellite mission. The SMAP satellite provides passive radiometer measurements which allows for inference of the soil moisture content in the top 5cm of soil. Satellite readings have global coverage with a return period between 2-3 days for each pass[6]. SMAP data is offered at varying levels of post-processing. The two levels of interest are L3 and L4. L3 data consists of preprocessed measurements that are gridded and mapped spatiotemporally across the globe. L4 data is a further processed gapfilled product derived from L3. In principle, the L4 product offers much greater spatio-temporal coverage and would offer greater data availability. However, training on the L3 product yielded better results and so the L3 product was used throughout. The L3 product records two daily passes of AM (morning) and PM (evening) as it orbits. This does not mean the L3 product has an AM and PM reading for every location on Earth for every day. But, if there exists a reading for a location on that day, it will be either an AM or PM reading. In order to increase SMAP L3 temporal coverage, a simple gap filling method was employed. This involved ignoring the AM and PM designation and using these passes as a single daily reading. Any areas that experienced both AM and PM passes were averaged. This was done because in-situ data will be aggregated into daily readings and as such are less sensitive to the specific time of SMAP measurement. Therefore, SWC measurements with greater than daily resolution precision are not considered.

2.2 Moderate Resolution Imaging Spectroradiometer (MODIS)

The Moderate Resolution Imaging Spectroradiometer (MODIS) mission provides daily temporal resolution remote sensing data from sun-synchronous orbits. MODIS offers a wide variety of spectral reflectances across multiple wavelengths to characterize and infer the Earth surface and its properties.

The three MODIS inferred properties we use are Land Surface Temperature (LST), Evapotranspiration (ET), and the Normalized Difference Vegetation Index (NDVI). In this study, the 500m NDVI (MOD13A1) product is used for training and temporal predictions. The finer 250m NDVI product (MOD13Q1) is used for spatial predictions. The 8-day LST (MOD11A2) product was used during training and prediction to avoid cloud coverage. The daily LST product (MOD21A1) was used for spatial prediction. The 8-day ET product (MOD16A1) based on a modified Penman-Montieth equation is used for ET estimation. This product has a spatial resolution of 500m.

For land cover type classification, the MCD12Q1 product is used with a temporal resolution of 1-year and a spatial resolution of 500m.

2.3 CHIRPS 2.0 Precipitation

Precipitation data was retrieved from the Climate Hazards Center at Santa Barbara[24]. 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 (~5.5km).

2.4 Soil Texture and Soilgrids

The International Soil Reference and Information Centre (ISRIC) has produced a global harmonised soil properties database called SoilGrids[25]. 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 1km 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.

Topography

The Multi-Error-Removed Improved-Terrain (MERIT) Digital Elevation Model (DEM) topography product was used for this study[26]. This product has a spatial resolution of ~90m.

2.5 In-Situ soil moisture measurements

Ground truth data for training the models were obtained from in-situ SWC measurements at sites distributed from two networks 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[27]. Ameriflux is a network of flux towers spread across North America recording various atmospheric and meteorological data and fluxes[28]. 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 downloaded and used in this study. 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 quality, the Ameriflux data was pruned to only contain readings with similar properties to ISMN readings with a 'G' quality flag. This means Ameriflux samples were dropped if either the LST reading was below 3 $^{\circ}$C or the SWC reading was above 0.7 $\mathbf{m^3/m^3}$. Additionally, sites in wetland and chronically inundated regions were excluded from the dataset.

SWC measurements are then aggregated to daily averages.

2.6 Datasets

The primary dataset is comprised of all available data from ISMN and Ameriflux soil moisture measurements within the temporal and spatial boundaries. Each location is classified by soil texture class. For each soil texture class, 80% of sites and all of the samples belonging to them are moved to a training set and the remaining 20% of sites and their samples are sent to the validation set. This split makes certain that not only are the validation data samples unseen by training, but they are also locations not seen by the model. This ensured that we can generalize the results to the greater CONUS area. Each daily aggregate of in-situ measurements is accompanied by daily aggregate measurements for the covariate inputs. The final dataset is comprised of 657,935 samples and 1054 stations. 206 of which were moved into the validation dataset. For further validation, two more datasets comprising a small network of soil moisture stations, originally used to calibrate SMAP, will be used to assess performance. Further discussion of their contents can be found in the supplementary document.

Next, we will look at how the information within the datasets is utilized to train the ensembles.

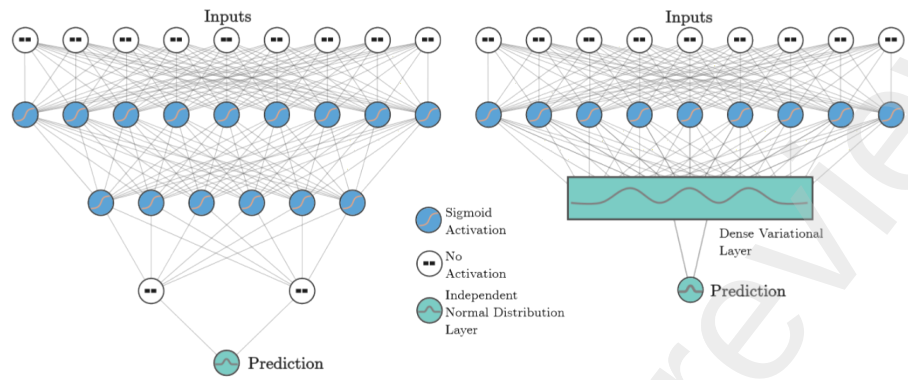
3 Models and Methods

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. This study looks at the performance of four different ensemble architectures. Two of the ensembles are replications of the architectures used by Abazsddeh (RF) and Xu (WDL). The remaining two models are simple distance based models. The first being a feed-forward network (Dense) and the other using a probabilistic layer (Prob). Both of their architectures were chosen so as to have almost the same number of hidden parameters. The architectures of the two smaller networks and WDL architectures can be seen in Figures 2 and 3 respectively. More detailed descriptions of their architectures can be found in the supplement.

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

Table 1: All of the categorical land characteristic subclasses.



(a) Dense Architecture (b) Prob Architecture

Figure 2: Probabilistic model architectures

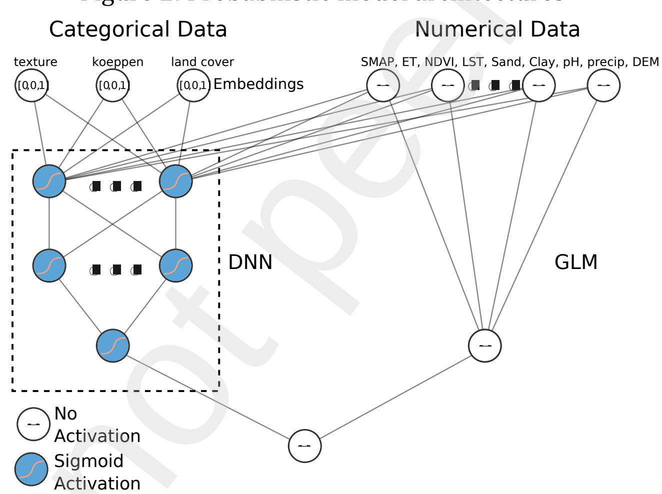


Figure 3: WDL Architecture

3.1 Training

In this study, we assume that static variables as seen in Table 1 either aide or hinder the models ability to discern SWC. Since these variables are not balanced in the dataset, the model may focus on the most abundant subclass types while neglecting to learn how to predict on other underrepresented subclasses. To account for these imbalances, instead of additional data manipulation, a simple approach is undertaken in the form of ensembles. Each ensemble member is trained with sample weights accounting for imbalances within a static characteristic. For example, an ensemble member trains on data weighted to the different soil texture class abundances giving extra weight/importance to correctly predicting the less abundant texture types. For the Dense, Probabilistic, and WDL ensembles, those static characteristics are **texture**, **clay** and **sand content**, **Köppen climate class**, **land cover class**, and an **unweighted** category that does not use any balancing. Therefore, there are 7 members per ensemble (one per characteristic) as seen in Fig. 4.

The weighting scheme for each static class 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 doesn’t use sample weights. Instead, balance is accounted for by training a unique model for each soil texture domain as done by Abbaszadeh et al.[17]. 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 to learn from due to less gaps from cloud cover. This means 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. Although this is not ideal, the rationale is that SMAP would account for the temporal variation in SWC while the other variables would 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. This loss of temporal information seems to be offset by the increase in samples to learn from and is discussed further in the supplement document.

3.2 Predictions

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

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

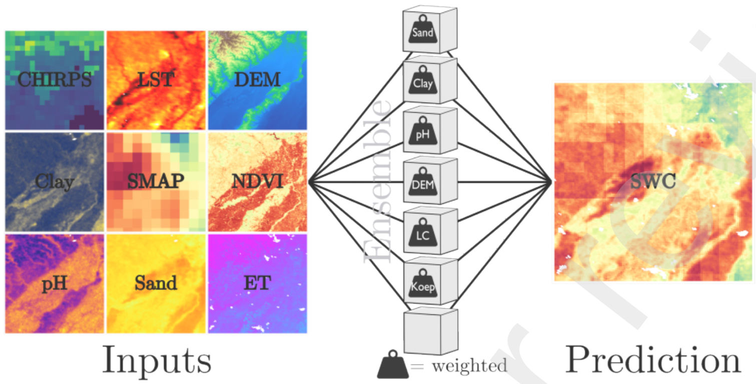


Figure 4: Prediction regime for the Dense, Prob, and WDL ensembles. Each ensemble member (cube) is trained while weighted against imbalances in a specific characteristic. These predictions are then averaged to provide an ensemble prediction.

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).

When making spatial predictions, spatial data are resampled to the highest resolution (90m) using nearest neighbor interpolation. This prevents interpolation error, but introduces some pixelation at higher levels of zoom.

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 [6]. 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 dictates 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 a large dataset comprising the validation data set aside during training. The second and third comprise smaller networks of soil moisture stations located in Oklahoma. Predictions will be compared against in-situ measurements as well as the predictions made by SMAP at that location.

4.1 CONUS Dataset

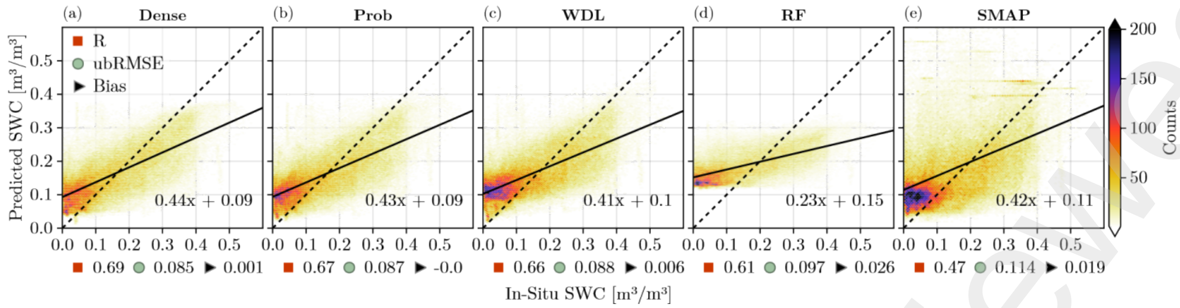


Figure 5: Heatmaps and metrics for algorithm predictions on the validation dataset as a whole.

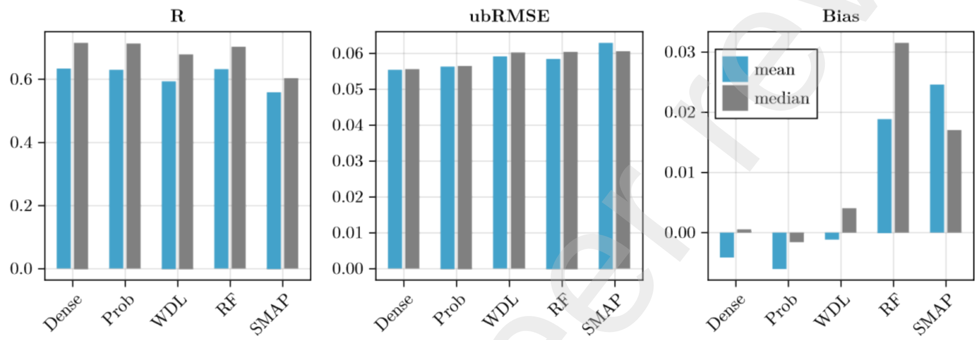


Figure 6: The average metric score for every site in the validation dataset. (a) numerically (b) visually

Because downscaling is an attempt at spatial prediction and reasoning, it's 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.

As shown in Fig. 5, every method 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 0.18m^3/m^3. Because of this, the lowest SWC prediction by the RF ensemble on the entire dataset is 0.10m^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 ensembles again outperform SMAP on every metric with exception to RFs bias. This is displayed in Figure 6. In the same figure we also see that timeseries are less consistent from site to site as the mean is notably lower than the median, but the ubRMSE shows a strong agreement between mean and median values demonstrating general consistency for prediction accuracy. Overall, this suggests all methods and their predictions should be as reliable or moreso than SMAP.

4.1.1 Spatial Predictions

To compare the spatial predictions of each method, a 1°x 1°box is cut out around a specific in-situ location on a summer day with the least cloud cover. Of the resulting predictions, six examples that exhibit unique characteristics are presented, two of which are highlighted in Figure 7. Overall, the ensembles tend to exhibit similar spatial patterns. In some cases, as exhibited in the predictions around *PBO: H2O LITTLELOST*, the 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. The other four examples can be seen and are discussed in the supplement.

Next we will look at the ensembles predictions over time.

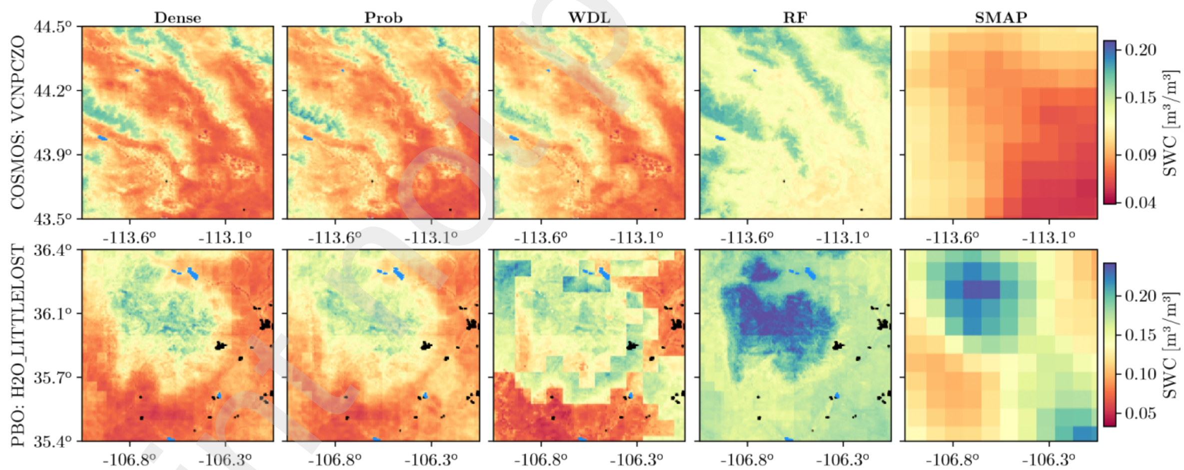


Figure 7: 1°x 1°spatial SWC predictions of ensembles vs SMAP. Black pixels represent pixels masked as ’urban’ and blue pixels are water surfaces.

4.1.2 Temporal Predictions

Although the R metric is calculated for each site in the validation set, it’s also important to view the time-series plotted against each other. For this analysis, the ten sites with the most data were selected and the time-series from 2018 is plotted. One of which is seen in Figure 8. The same figure also shows the R scores for the validation dataset on each station. Here we can see that the two top performing models in this metric (Dense and RF) both have drastically tightened distributions for R values compared to SMAP. Despite RF having similar performance to Dense, it's clear in the additional timeseries found in the supplement that RF possesses a strong bias and is often distinct from the SMAP, Dense, and in-situ markers. In general, the timeseries predictions of all models are as good or better than those of SMAP.

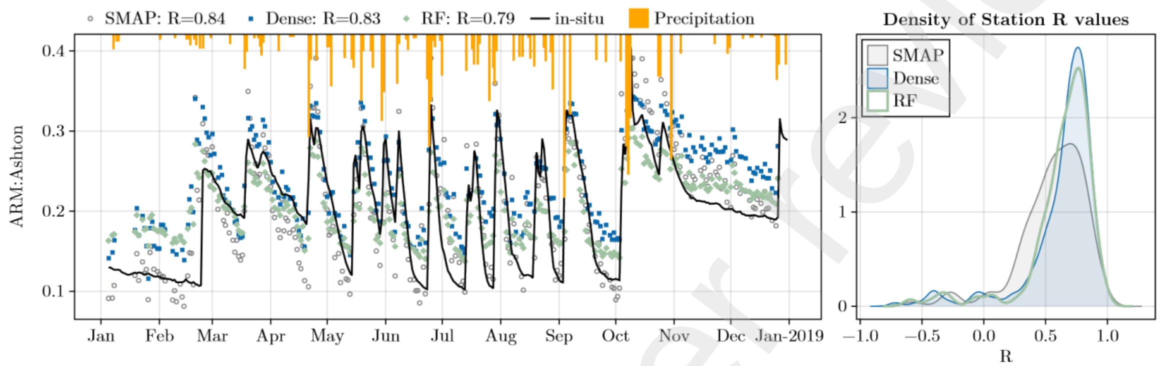


Figure 8: (Left) Temporal predictions on a station in the validation dataset. (Right) Density plot of the R values for each station in the validation dataset.

In the next subsection we will look at the performance of the ensembles on two additional test datasets.

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[6] 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.

Washita

The first dataset is the Washita River basin network.

	Dense	Prob	WDL	RF	SMAP

In this region, all methods struggle on the Washita dataset as a whole as seen in Fig 9. All methods 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.

R	**0.752**	0.661	0.681	0.700	0.745
ubRMSE	**0.041**	0.062	0.046	0.044	0.046
Bias	0.053	0.246	0.076	**0.006**	0.011

Table 2: Average site metric scores on the Washita dataset

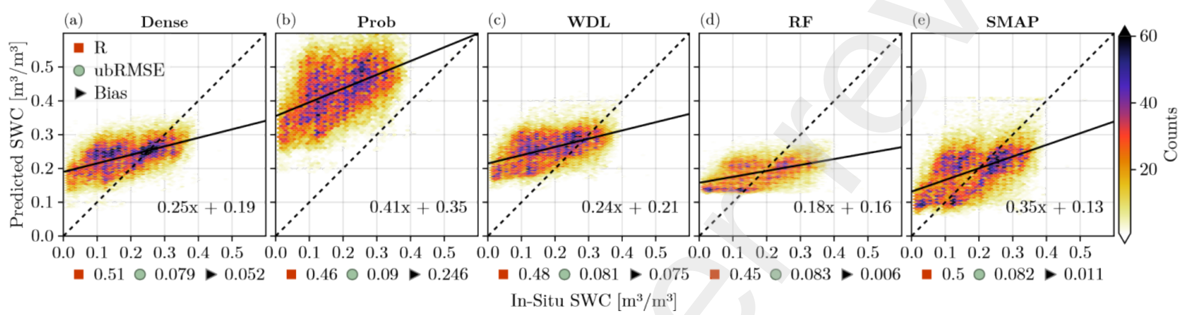


Figure 9: Heatmaps and metrics for algorithm predictions on the Washita dataset as a whole.

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. SMAP also exhibits good performance as expected. The other methods 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 is composed of measurements from the Fort Cobb network. Due to it's close proximity to Washita, its no suprise that we see similar trends. All methods demonstrate poor fitting to the dataset as a whole and the models show a strong positive bias at low SWC measurements. The RF model yields the best bias metric, although likely due to values being squashed towards a mean value.

	Dense	Prob	WDL	RF	SMAP
R	0.748	0.708	0.673	0.704	**0.752**
ubRMSE	**0.042**	0.049	0.043	0.043	0.046
Bias	**0.060**	0.116	0.079	0.062	0.062

Table 3: Average site metric scores on Fort Cobb dataset

Again, the model performance metrics increase on a site level (Table 3). The dense model is the closest method to the 0.04 m^3/m^3 ubRMSE threshold established by the SMAP mission. RF also scores within the realms of acceptability for this metric. The Prob and WDL models are unable to outperform SMAP on any metric with SMAP having the best R score.

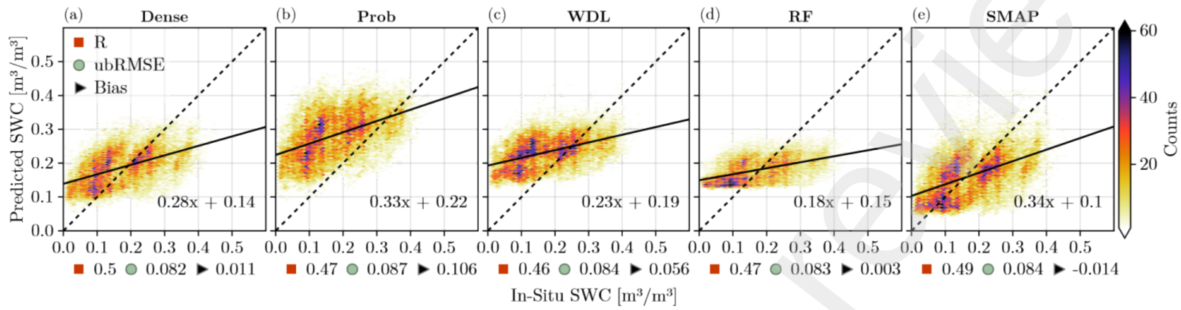


Figure 10: Heatmaps and metric scores for algorithm predictions on the Fort Cobb dataset as a whole.

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 method can reliably match or outperform SMAP here, it would suggest confidence in it's ability to perform elsewhere. The Dense architecture is the only method to reliably match or exceed SMAP on key metrics on these datasets.

Timeseries

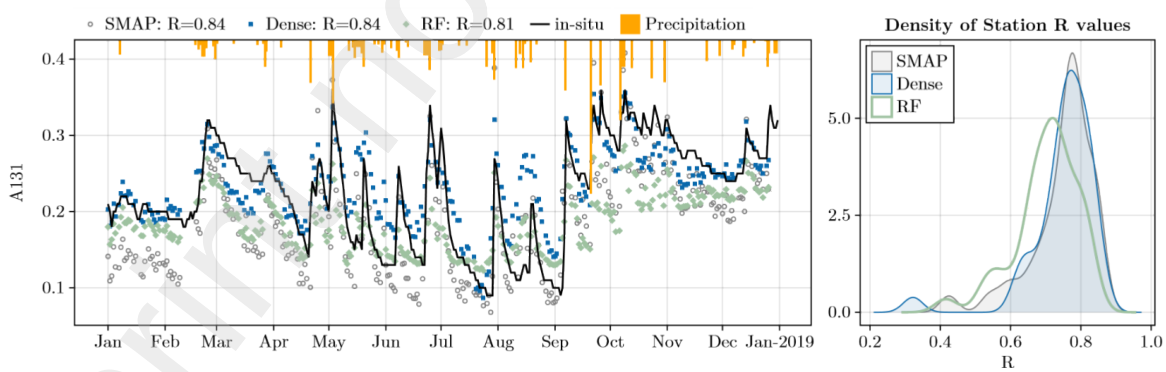


Figure 11: (Left) Temporal predictions on a station in the validation dataset. (Right) Density plot of the R values for each station in both OK datasets.

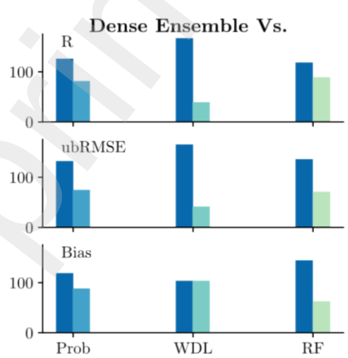
Similar to the timeseries predictions for the validation set. Timeseries predictions from the Oklahoma dataset help assure us that models are maintaining consistency through time. SMAP has a home field advantage at these sites and only the Dense architecture is able to demonstrate parity and match

SMAPs strong temporal accuracy. A timeseries of a station in the Washita dataset is plotted in Figure 11 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 and the Dense peak is a bit below that of SMAP.

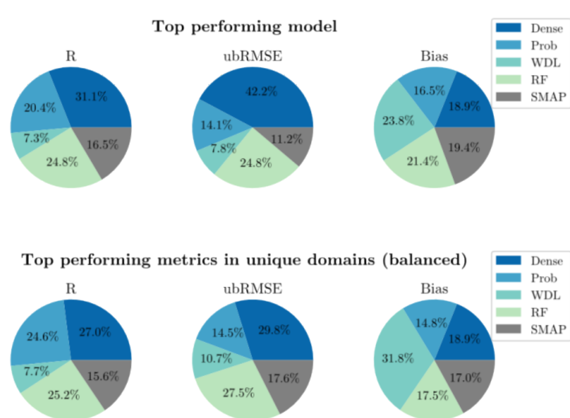
In the next section we will analyze the robustness of the results and look for potential limitations.

4.3 Top performer

We can evaluate performance based on three criteria: dataset, sites, and domains. We saw in the previous sections that the Dense model was consistently a top performer on datasets, but what about site and domain? For site level, we compare the Dense predictions on each site against the other architectures in the validation dataset. In this context, the Dense architecture outperforms every other model in every other metric as seen in Fig. 12a with the exception of the bias against WDL. In a head-to-head competition of all methods, Dense is the clear winner in ubRMSE and notable winner in R. WDL maintains the best method for bias. To see if Dense is still the top performer by domain, we look at each models performance on stations belonging to the subclasses of each categorical land surface attribute as seen in Table 1. Performance is then normalized so over/underrepresented classeas have equal impact on performance. This normalizing method is discussed further in future sections. When normalizing for class type and abundance, we can see (Fig. 12b) the Dense model is still the most consistent performer for R and ubRMSE. However, this is only slightly more dominant than the RF ensemble. WDL is again the clear top performer for bias.



(a)



(b)

Figure 12: (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 category in Table 1. Each time a model outperforms every other method for that metric it gets a point. All points for that category are normalized so that the top performer receives one point for that category. All points are summed together for all categories. This produces an unbiased assessment of model performance regardless of imbalances in representation of classes.

Having a distance based model outperform the RF has additional advantages. For starters the evaluation speed for distance based models is two orders of magnitude faster (0.16s vs 17.7s on 130k samples). Therefore, it's more feasible to predict over large domains. Additionally, the file size of the RF ensemble is three orders of magnitude larger (2.3GB vs 1.03MB) which makes transferring it less convenient than the simple distance based ensembles. For these reasons, it doesn't seem reasonable to continue using a RF architecture for this task at this resolution.

Next we will look to see how generalizable the performance of the models are for different land surface characteristics.

4.4 Domain Preference

To further explore areas of strengths and weakness', metrics are calculated across each of the three categorical static characteristics: **texture**, **climate class**, and **land cover**. These static characteristics are further broken down into the subclasses previously shown in Table 1. A significant drop in metric performance in one of these subclasses may indicate an inability for a model to fully generalize SWC from the input variables. To search for these preferences/weaknesses we compute the average metric score for a method on each station in the 40 subclasses from Table 1. We then divide this by the average performance for all models on that subclass. This final value gives us the relative performance of a model compared to all others. If any models performance is at least 10% better or worse than the mean score for all models on that subclass, then that model is deemed to have a bias for that subclass. These instances are seen in Table 4. The Bias metric was excluded as the RF model consistently exhibited poor bias. The only instance where a model demonstrates a negative or positive performance on both ubRMSE and R was on Sand. Here, the Dense R value is 40% the mean R value and the ubRMSE is 124% the mean ubRMSE value. This category constitutes only one stations worth of data and so no conclusions can be made about the models performance on sand overall.

Although there doesn't appear to be any strong or negative biases for any single static characteristics, what if there exists a combination of inputs that exhibit difficulties? The next section will explore for just such an instance.

Characteristic	Dense	Prob	WDL	RF	No.ofStations
			R		
SiClLo	1.07	1.05	**0.83**	1.05	3
MxdFrsts	1.08	0.98	**0.89**	1.04	3
Bsh	1.04	1.05	**0.88**	1.02	2
Sa	**0.44**	1.21	1.17	1.18	1
			ubRMSE		
Csa	0.92	0.99	**1.10**	0.98	24
Opn Shrblnds	0.94	1.01	**1.14**	0.91	6
SaClLo	1.03	1.04	1.04	**0.89**	3
Bsh	0.95	**1.14**	0.94	0.91	2
ET	1.00	**1.14**	0.94	0.92	2
BWh	0.99	1.13	0.99	**0.90**	1
Sa	**1.24**	**0.71**	1.05	1.00	1
Cl	**0.85**	1.03	1.09	1.03	1

Table 4: Static classes where one model displays a bias (an average metric score on that class which deviates 10% or more from the mean of all models) for that specific class. For R, values greater than 1.0 outperform the mean, for ubRMSE values below 1.0 outperform the mean. No. of stations represents number of locations possessing that characteristic

4.5 Areas of Underperformance

To find combinations of characteristics that exhibit underperformance, the static characteristics for each site in the CONUS dataset were compiled into a dataset with six dimensions (sand, clay, pH, topography, climate class, land cover type) whose values were normalized for each dimension. 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 Figure 13. This area in the PCA represents Cropland land cover type with high clay content and low sand content as seen in Table 5. These values are scaled by the standard deviation of the dataset for each static characteristic. 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 standard deviations below the mean. Most of these sites are croplands.

This brief analysis shows that the best performing model (Dense) does not have consistent performance on croplands of high clay and low sand content values. Therefore, this method 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.

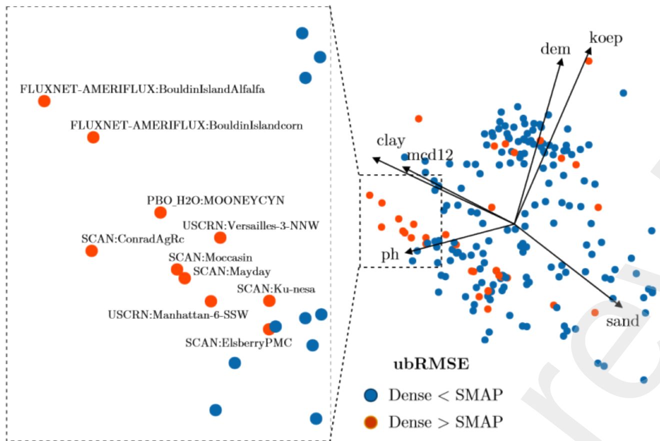


Figure 13: Reprojection of test data static characteristics into PCA space. Peach dots represent sites where the Dense ensemble's ubRMSE score was worse than SMAP

site	*Sand*	*Clay*	*pH*	*Dem*	*Koep*	*LC*
SCAN:Ku-nesa	-2.02	1.52	-0.00	-1.08	Cfa	Svnnas
USCRN:Manhattan-6-SSW	-1.88	1.52	0.58	-1.05	Cfa	Grsslnds
FLUXNET-AMERIFLUX:BouldinIslandAlfalfa	-1.60	3.63	-0.12	-1.38	Csa	Crplnds
FLUXNET-AMERIFLUX:BouldinIslandcorn	-1.52	3.14	-0.12	-1.39	Csa	Crplnds
PBO H2O:MOONEYCYN	-0.82	2.01	1.40	-0.98	Csb	Crplnds
SCAN:ConradAgRc	-1.10	2.33	1.17	-0.31	BSk	Crplnds
SCAN:ElsberryPMC	-2.09	0.39	0.11	-1.24	Cfa	Crplnds
SCAN:Mayday	-1.38	2.17	-0.35	-1.35	Cfa	Crplnds
SCAN:Moccasin	-0.82	1.84	0.93	-0.14	BSk	Crplnds
USCRN:Versailles-3-NNW	-2.37	0.39	-0.24	-1.12	Cfa	Crplnd/Natr msaic
Mean	**-1.56**	**1.89**	**0.34**	**-1.00**	–	–

Table 5: The deviations from mean values for static characteristics at the site level

4.6 Cross-fold Analysis

In order to assess whether our methodology is generalizable. A 10-fold cross validation was conducted. This involved splitting the original dataset into 10 separate datasets containing 10% of the total stations and their respective data. For each of these 10 datasets, the ensembles are trained on the other 90% and then predict the in-situ values for those left out. These datasets are produced randomly and so their proportions of different static characteristics is not curated. This randomness may have a negative impact on the RF ensemble as it has no weighting scheme to account for the imbalances it will learn from.

In general, the metrics from the cross validation are similar to those achieved in the validation set. The exception being the RF ensemble. This is likely due to the RF method relying on needing some information from each texture class. But not every cross validation subset has every texture to learn from. The density curves for the R values for each station in the cross validation dataset are plotted in Figure 14. Compared to SMAP, the Dense and Prob methods (the two strongest performers) have

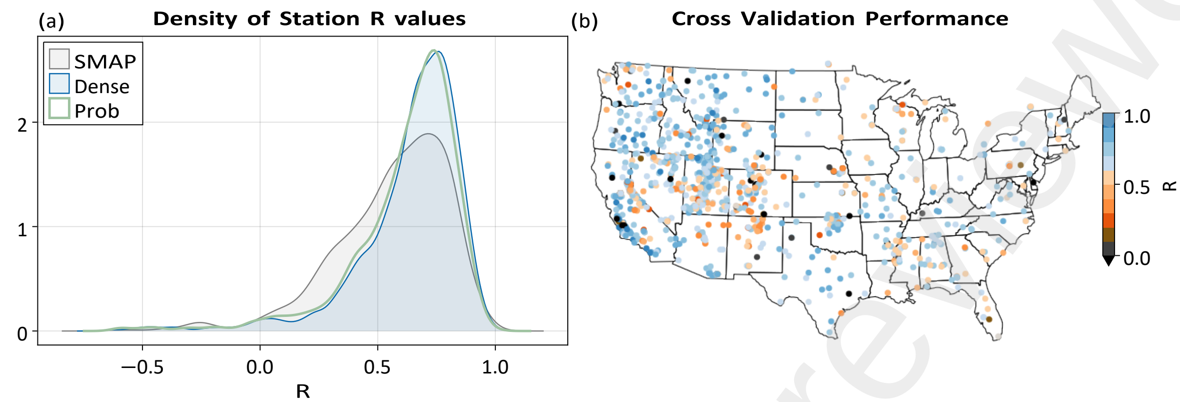


Figure 14: (a) Density plots of the Dense and Prob R values for each station in the cross validation dataset. (b) Spatial distribution of R values on each station as predicted by Dense

their distributions tightened over higher R values. This was also the case for the WDL and RF (seen in supplement), but the RF distribution is notably less impressive as expected. Density plots for ubRMSE show improvement from SMAP in all methods except with RF and can be found in the supplement. For the weighted methods (Dense, PRob, WDL), the cross validation appears to confirm that the weighting scheme limits biases in the training data.

Model	*Dataset*	R	*ubRMSE*	*Bias*
Dense	Val	0.632	**0.055**	-0.004 **-**
	Cross Val	**0.639**	0.058	**0.000**
Prob	Val	**0.628**	**0.056**	**-0.007**
	Cross Val	0.621	0.060	-0.008
WDL	Val	0.594	**0.059**	**-0.001**
	Cross Val	**0.611**	0.060	-0.003
RF	Val	**0.630**	**0.058**	0.019
	Cross Val	0.572	0.065	**0.004**
SMAP	Val	0.559	**0.063**	0.025
	Cross Val	**0.562**	0.065	**0.023**

Table 6: The mean metric score for each method on each station on the validation set vs the cross validation dataset

5 Discussion

The primary focus for this section is to evaluate the the robustness and generalizability of the methods. Additionally, we want to look at the ensemble framework in context of this work and identify whether or not there is any advantage from an ensemble prediction, or if we can achieve equally satisfactory results with just a single ensemble member.

5.1 Generalizability

Large domain predictions only yield value if we can trust that those predictions are generalizeable, or consistently accurate, across the hetereogeniety of the domain. To test whether these ensemble predictions can extrapolate beyond their training dataset, we ensured that validation data belonged to locations previously unseen and foreign to the models. After analysis yielded no concerning biases or shortcomings, we then conducted a crossfold analysis across all sites in the training and validation set. Again, we see consistent/similar performance on each site when it was previously unseen during training. The last form of analysis involved monitoring spatial predictions and their associated SHAP values. This analysis is discussed further in the supplement. We find that the SHAP values generally adhere to expectations found in literature, however strangely all methods seem to have an inverse relationship for NDVI from what is expected. Further analysis was not conducted to discern why this was the case.

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

5.2 Ensemble Advantage

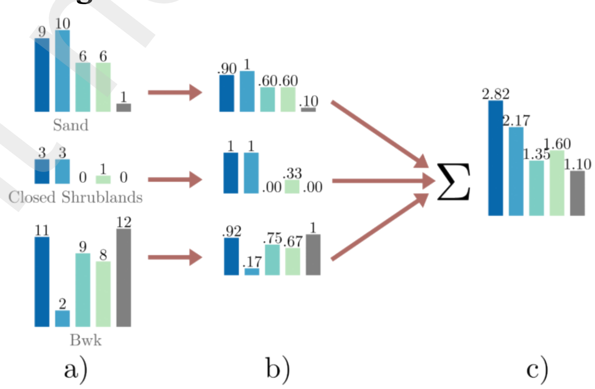


Figure 15: Weighting schema for unbiased top performers. a) All models predict on all sites belonging to a specific category. Each time a model outperforms every other model it gets a point. b) Points are then normalized. This ensures under-represented categories have equal importance in assessing model performance. c) The normalized points are summed providing a final assessment of model performance on all categories.

This study serves to assess the feasibility and advantage of using an ensemble of models to predict SWC at higher resolutions. In the case of the two probabilistic ensembles (Dense and Prob), they

Model	Metric	*Ens.*	Sand	Clay	Koep	MCD12	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

Table 7: Average station performance for each ensemble member and the ensemble as a whole on the validation dataset.

represent exceedingly simple models. The purpose of these ensembles is to permit equal representation for all unique land characteristics in the training process as to prevent overfitting to a dominant characteristic. However, perhaps the weighting scheme for one land characteristic may be a sufficient representation of the data and an ensemble is redundant.

First we compare the average performance of each ensemble member against the ensemble in the validation dataset. This is seen in Table 7. Here, we can see that for the Dense ensemble, the ensemble is only marginally better than its unweighted member. Whereas for the Prob and WDL ensembles, the Sand and Clay weighted members outperformed their respective ensembles. In all instances the ensembles average performance is not significantly improved upon when compared to the unweighted member.

To ensure that there isn't a dominant subclass that is easy to predict for both ensemble and members, we compare the ensembles performance on static domains against every ensemble member. In other words, for each texture/land cover/Koeppen class listed in Table 1, we compare the prediction performance of individual ensemble members versus the full ensemble on that subset of data. For each site a model outperforms the other, their score for that class increases. The two scores for that class are normalized so that the model that outperforms on the most sites receives a value of 1. This process is illustrated in Fig. 15. This is done for each metric (R, ubRMSE, Bias). These final scores are summed and these final sums represent the total normalized performance ratio for that ensemble vs ensemble member pairing. These final normalized performance ratios for each ensemble-member pairing are visualized in Fig. 17.

When looking at these unbiased performances across subclasses, 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 architecture is likely too simple to overfit a characteristic, and the GLM of the WDL seems to be adept at guiding predictions and preventing overfitting. From a purely numerical context, there does not exist a clear ensemble advantage.

Lastly, we compare the spatial predictions of the ensemble vs the unweighted ensemble member. Here there exists a much starker difference in behaviour. Namely, the Dense ensemble predictions seem to capture more of the land surface characteristics than the single ensemble member. This is seen in Figure 16. Although not directly quantifiable, it is clear that the Ensemble is able to incorporate more of the land surface characteristics into it's prediction than the unweighted ensemble member. This however, is not the case for the Prob architecture. The single ensemble member for Prob seemed do distinguish the same land characteristic fidelity as the ensemble. For the WDL architecture, ensemble member prediction is noisier than the ensemble. Further analysis will need to be conducted to asses whether these behaviours constitutes a substantial improvement of one over the other.

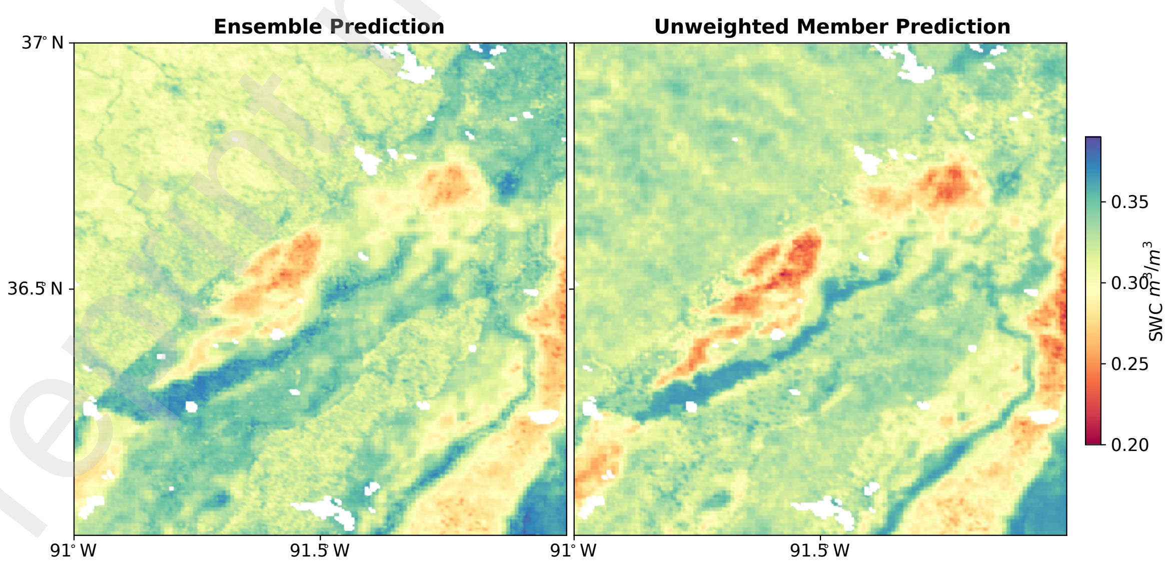


Figure 16: Spatial Predictions comparing the Dense ensemble vs the unweighted (Free) ensemble member

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

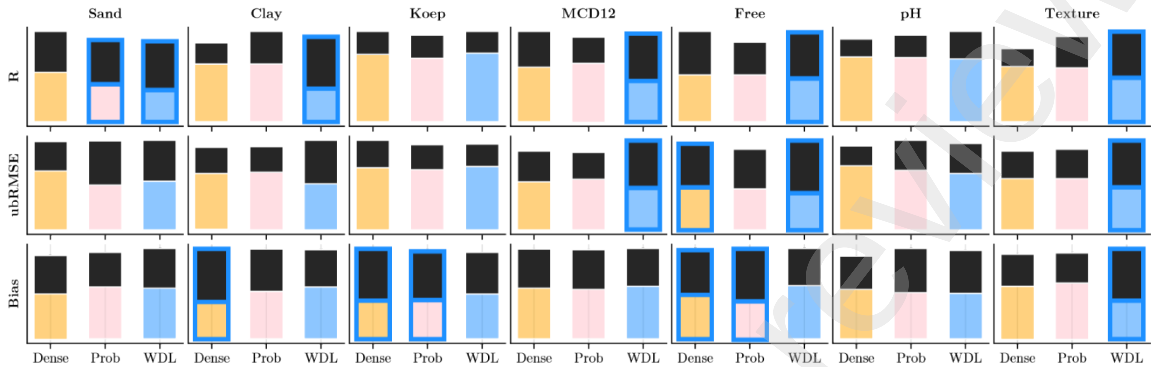


Figure 17: Head to head comparison of Ensembles (Bottom label) vs their member constituents (Top label) with normalized performances. Bars highlighted in blue indicate an instance where an ensemble member outperformed the ensemble on that metric (Left label). An explanation of this head to head competition is seen in Figure 15

6 Conclusion

The work conducted in this paper served to demonstrate that an ensemble of simple ML architecture can yield acceptable SWC downscaling results. Analysis revealed that these ensembles can reliably do this with strong generalizability. However, certain ensemble members can outperform or achieve parity with the full ensemble on the validation dataset. This suggests there is no/little benefit one would achieve from an ensemble that one would not also achieve with a rigorous sample weighting scheme. Despite this, Comparison of the spatial predictions between Ensembles vs these seemingly similarly performing members showed that ensembles appear to capture more of the land surface characteristics. More analysis is needed to assess whether or not this is advantageous and by how much. Multi-variable analysis of ensemble 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 ensemble which needs curated dated to learn from and so struggles on the random crossfold data.

Final summary:

- Ensembles of simple ML architectures can downscale SWC predictions to sub 1km resolutions

- Simpler architectures can outperform or match the performance of these ensembles on datasets. However, the spatial predictions of the ensembles can capture more of the land characteristics than the ensemble member and reduce noise.

- Training the models on temporally padded data provides more benefits than drawbacks in terms of overall performance.

- The top performing ensemble is unreliable on croplands with higher than average clay and lower than average sand content.

6.1 Acknowledgements

The research was generously support by funding from the Max Planck Institute of Biogeochemistry with support from the International Max Planck Research School for Global Biogeochemical Cycles. 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.

Competing Interests

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

References

[1] Laibao Liu, Lukas Gudmundsson, Mathias Hauser, Dahe Qin, Shuangcheng Li, and Sonia I.

Seneviratne. Soil moisture dominates dryness stress on ecosystem production globally. *Nature Communications*, 11(1):4892, December 2020.

[2] Zheng Fu, Philippe Ciais, I. Colin Prentice, Pierre Gentine, David Makowski, Ana Bastos, Xiangzhong Luo, Julia K. Green, Paul C. Stoy, Hui Yang, and Tomohiro Hajima. Atmospheric dryness reduces photosynthesis along a large range of soil water deficits. *Nature Communications*, 13(1):989, December 2022.

[3] Benjamin D. Stocker, Jakob Zscheischler, Trevor F. Keenan, I. Colin Prentice, Sonia I. Seneviratne, and Josep Pen˜uelas. Drought impacts on terrestrial primary production underestimated by satellite monitoring. *Nature Geoscience*, 12(4):264–270, April 2019.

[4] Marco Bittelli. Measuring Soil Water Content: A Review. *HortTechnology*, 21(3):293–300, June 2011.

[5] Kaijun Wu, Gabriela Arambulo Rodriguez, Marjana Zajc, Elodie Jacquemin, Michiels Cl´ement, Alb´eric De Coster, and S´ebastien Lambot. A new drone-borne GPR for soil moisture mapping. *Remote Sensing of Environment*, 235:111456, December 2019.

[6] Dara Entekhabi. *SMAP Handbook Soil Moisture Active Passive*. JPL Publication JPL, 2014.

[7] Peggy E. ONeill, Steven Chan, Eni G. Njoku, Tom Jackson, and Rajat Bindlish. SMAP Enhanced L3 Radiometer Global Daily 9 km EASE-Grid Soil Moisture, Version 3, 2019.

[8] Noemi Vergopolan, Justin Sheffield, Nathaniel W. Chaney, Ming Pan, Hylke E. Beck, Craig R. Ferguson, Laura Torres-Rojas, Felix Eigenbrod, Wade Crow, and Eric F. Wood. High-Resolution Soil Moisture Data Reveal Complex Multi-Scale Spatial Variability Across the United States. *Geophysical Research Letters*, 49(15):e2022GL098586, August 2022.

[9] Bibi S. Naz, Wolfgang Kurtz, Carsten Montzka, Wendy Sharples, Klaus Goergen, Jessica Keune, Huilin Gao, Anne Springer, Harrie-Jan Hendricks Franssen, and Stefan Kollet. Improving soil moisture and runoff simulations at 3 km over Europe using land surface data assimilation. *Hydrology and Earth System Sciences*, 23(1):277–301, January 2019.

[10] Brahima Kon´e, Arona Diedhiou, Adama Diawara, Sandrine Anquetin, N'datchoh Evelyne Tour´e, Adama Bamba, and Arsene Toka Kobea. Influence of initial soil moisture in a regional climate model study over West Africa – Part 1: Impact on the climate mean. *Hydrology and Earth System Sciences*, 26(3):711–730, February 2022.

[11] Brahima Kon´e, Arona Diedhiou, Adama Diawara, Sandrine Anquetin, N'datchoh Evelyne Tour´e, Adama Bamba, and Arsene Toka Kobea. Influence of initial soil moisture in a regional climate model study over West Africa – Part 2: Impact on the climate extremes. *Hydrology and Earth System Sciences*, 26(3):731–754, February 2022.

[12] Andreas Colliander, Joshua B. Fisher, Gregory Halverson, Olivier Merlin, Sidharth Misra, Rajat Bindlish, Thomas J. Jackson, and Simon Yueh. Spatial Downscaling of SMAP Soil Moisture Using MODIS Land Surface Temperature and NDVI During SMAPVEX15. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2107–2111, November 2017.

[13] Nitu Ojha, Olivier Merlin, Christophe Suere, and Maria Jos´e Escorihuela. Extending the Spatio-Temporal Applicability of DISPATCH Soil Moisture Downscaling Algorithm: A Study Case Using SMAP, MODIS and Sentinel-3 Data. *Frontiers in Environmental Science*, 9:555216, March 2021.

[14] Jingyao Zheng, Haishen Lu¨, Wade T. Crow, Tianjie Zhao, Olivier Merlin, Nemesio Rodriguez-Fernandez, Jiancheng Shi, Yonghua Zhu, Jianbin Su, Chuen Siang Kang, Xiaoyi Wang, and Qiqi Gou. Soil moisture downscaling using multiple modes of the DISPATCH algorithm in a semi-humid/humid region. *International Journal of Applied Earth Observation and Geoinformation*, 104:102530, December 2021.

[15] Juan M. S´anchez, Joan M. Galve, Jos´e Gonz´alez-Piqueras, Ram´on L´opez-Urrea, Raquel Nicl`os, and Alfonso Calera. Monitoring 10-m LST from the Combination MODIS/Sentinel-2, Validation in a High Contrast Semi-Arid Agroecosystem. *Remote Sensing*, 12(9):1453, May 2020.

[16] Nitu Ojha, Olivier Merlin, Beatriz Molero, Christophe Suere, Luis Olivera-Guerra, Bouchra Ait Hssaine, Abdelhakim Amazirh, Ahmad Al Bitar, Maria Escorihuela, and Salah Er-Raki.

Stepwise Disaggregation of SMAP Soil Moisture at 100 m Resolution Using Landsat-7/8 Data and a Varying Intermediate Resolution. *Remote Sensing*, 11(16):1863, August 2019.

[17] Peyman Abbaszadeh, Hamid Moradkhani, and Xiwu Zhan. Downscaling SMAP Radiometer Soil Moisture Over the CONUS Using an Ensemble Learning Method. *Water Resources Research*, 55(1):324–344, January 2019.

[18] Mengyuan Xu, Ning Yao, Haoxuan Yang, Jia Xu, Annan Hu, Luis Gustavo Goncalves de Goncalves, and Gang Liu. Downscaling SMAP soil moisture using a wide & deep learning method over the Continental United States. *Journal of Hydrology*, 609:127784, June 2022.

[19] Hongfei Zhao, Jie Li, Qiangqiang Yuan, Liupeng Lin, Linwei Yue, and Hongzhang Xu. Downscaling of soil moisture products using deep learning: Comparison and analysis on Tibetan Plateau. *Journal of Hydrology*, 607:127570, April 2022.

[20] Carsten Montzka, Kathrina R¨otzer, Heye Bogena, Nilda Sanchez, and Harry Vereecken. A New Soil Moisture Downscaling Approach for SMAP, SMOS, and ASCAT by Predicting Sub-Grid Variability. *Remote Sensing*, 10(3):427, March 2018.

[21] Ahmed Samir Abowarda, Liangliang Bai, Caijin Zhang, Di Long, Xueying Li, Qi Huang, and Zhangli Sun. 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, March 2021.

[22] Wei Xu, Zhaoxu Zhang, Zehao Long, and Qiming Qin. Downscaling SMAP Soil Moisture Products With Convolutional Neural Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4051–4062, 2021.

[23] Yulin Cai, Puran Fan, Sen Lang, Mengyao Li, Yasir Muhammad, and Aixia Liu. Downscaling of SMAP Soil Moisture Data by Using a Deep Belief Network. *Remote Sensing*, 14(22):5681, November 2022.

[24] Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen.

The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1):150066, December 2015.

[25] Tomislav Hengl, Jorge Mendes De Jesus, Gerard B. M. Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagoti´c, Wei Shangguan, Marvin N. Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, Mario Antonio Guevara, Rodrigo Vargas, Robert A. MacMillan, Niels H. Batjes, Johan G. B. Leenaars, Eloi Ribeiro, Ichsani Wheeler, Stephan Mantel, and Bas Kempen. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2):e0169748, February 2017.

[26] Dai Yamazaki, Daiki Ikeshima, Ryunosuke Tawatari, Tomohiro Yamaguchi, Fiachra O'Loughlin, Jeffery C. Neal, Christopher C. Sampson, Shinjiro Kanae, and Paul D. Bates. A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853, June 2017.

[27] W. A. Dorigo, W. Wagner, R. Hohensinn, S. Hahn, C. Paulik, A. Xaver, A. Gruber, M. Drusch, S. Mecklenburg, P. Van Oevelen, A. Robock, and T. Jackson. 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, May 2011.

[28] T. A. Boden, M. Krassovski, and B. Yang. The AmeriFlux data activity and data system: an evolving collection of data management techniques, tools, products and services. *Geoscientific Instrumentation, Methods and Data Systems*, 2(1):165–176, June 2013.

Downscaling Soil Moisture to Sub-km Resolutions with Simple Machine Learning Ensembles

Jeran Poehls[1], Lazaro Alonso[1], Sujan Koirala[1], Markus Reichstein[1,2], and Nuno Carvalhais[1,3,4]

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

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 ensembles use a dataset of over 1,000 sites across the US to predict soil moisture at sub-km scales. We find that all ensembles, particularly one with a very simple structure, can outperform SMAP on a cross-fold analysis of the 1,000+ sites. This ensemble has an average ubRMSE of **0.058** vs SMAPs **0.065** and an average R of **0.639** vs SMAPs **0.562**. Not all ensembles are beneficial, with some architectures performing better with different training weights than with ensemble averaging. However, some ensemble architectures 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.

Keywords

Ensemble, Soil Moisture, Remote Sensing, Downscaling, SMAP

*coorespondence : jpoehls@bgc-jena.mpg.de ; Hans-Knöll-Straße 10, 07745 Jena, DE

Contents

1 Introduction

The water in the soil or soil water content (SWC) has a strong coupling with ecosystem stress and production[1][2][3]. 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[4]. 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[5]. But for truly global scale soil moisture mapping we need to look for the aid of satellites.

The Soil Moisture Active Passive (SMAP) radar 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[6]. The combination of these two would yield native SWC measurements at 9km per pixel and interpolated 1-3km 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 36km per pixel. This resolution can be increased using the Backus-Gilbert optimal interpolation algorithm to 9km per pixel with acceptable accuracy[7]. 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 100m-3km. Since, even at 1km resolution, up to 80% of SWC variability is lost[8]. At native satellite resolutions, there is a complete loss of SWC variability[8]. The 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 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 modelling[9][10][11]. 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[12][13][14]. 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[15][16]. But higher resolution LST data wouldn't 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[17][18][19][20]. However, a common occurrence are complex model architectures over particularly limited study areas[21][22][23]. 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 ensemble architectures for downscaling coarse spatial resolution soil moisture data to sub-km resolutions. The four ensembles include: two probabilistic estimators consisting of simple neural networks, a wide-deep learning (WDL) architecture modelled after the work of Xu et al. 2022, and a random forest (RF) model. These ensembles 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 methods 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 coarse resolution 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 that allow accurate and generalizable predictions. Therefore, 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 deciding which variables to include in the dataset. After a process of feature selection that is covered in the supplemental document, a dataset comprised of the following variables was assembled: *SMAP, NDVI, LST, Precipitation, Sand* and *Clay content, pH, Evapotranspiration,* and *Topography/Elevation.*

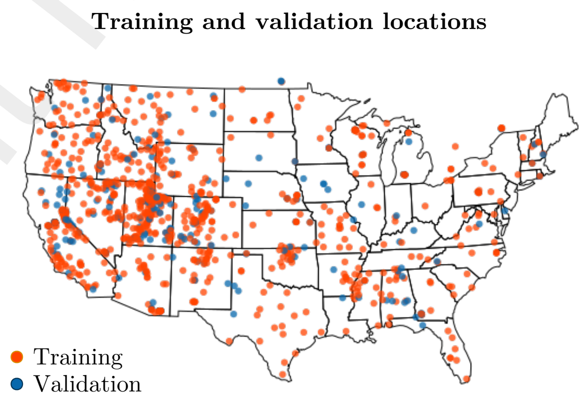


Figure 1: For this study, data within a temporal period extending from **January 1st, 2017** through **December 31st, 2021** was selected. This period ensured that soil moisture readings would include seasonal and, potentially, yearly variability.

This dataset was then iteratively trained over while excluding one of these variables. 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 folows:

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

Next we will discuss the sources used for this data.

2.1 Soil Moisture Active Passive (SMAP) Satellite Readings

The remotely sensed soil moisture readings are provided by NASAs SMAP satellite mission. The SMAP satellite provides passive radiometer measurements which allows for inference of the soil moisture content in the top 5cm of soil. Satellite readings have global coverage with a return period between 2-3 days for each pass[6]. SMAP data is offered at varying levels of post-processing. The two levels of interest are L3 and L4. L3 data consists of preprocessed measurements that are gridded and mapped spatiotemporally across the globe. L4 data is a further processed gapfilled product derived from L3. In principle, the L4 product offers much greater spatio-temporal coverage and would offer greater data availability. However, training on the L3 product yielded better results and so the L3 product was used throughout. The L3 product records two daily passes of AM (morning) and PM (evening) as it orbits. This does not mean the L3 product has an AM and PM reading for every location on Earth for every day. But, if there exists a reading for a location on that day, it will be either an AM or PM reading. In order to increase SMAP L3 temporal coverage, a simple gap filling method was employed. This involved ignoring the AM and PM designation and using these passes as a single daily reading. Any areas that experienced both AM and PM passes were averaged. This was done because in-situ data will be aggregated into daily readings and as such are less sensitive to the specific time of SMAP measurement. Therefore, SWC measurements with greater than daily resolution precision are not considered.

2.2 Moderate Resolution Imaging Spectroradiometer (MODIS)

The Moderate Resolution Imaging Spectroradiometer (MODIS) mission provides daily temporal resolution remote sensing data from sun-synchronous orbits. MODIS offers a wide variety of spectral reflectances across multiple wavelengths to characterize and infer the Earth surface and its properties. The three MODIS inferred properties we use are Land Surface Temperature (LST), Evapotranspiration (ET), and the Normalized Difference Vegetation Index (NDVI). In this study, the 500m NDVI (MOD13A1) product is used for training and temporal predictions. The finer 250m NDVI product (MOD13Q1) is used for spatial predictions. The 8-day LST (MOD11A2) product was used during training and prediction to avoid cloud coverage. The daily LST product (MOD21A1) was used for spatial prediction. The 8-day ET product (MOD16A1) based on a modified Penman-Montieth equation is used for ET estimation. This product has a spatial resolution of 500m.

For land cover type classification, the MCD12Q1 product is used with a temporal resolution of 1-year and a spatial resolution of 500m.

2.3 CHIRPS 2.0 Precipitation

Precipitation data was retrieved from the Climate Hazards Center at Santa Barbara[24]. 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 ($\sim$5.5km).

2.4 Soil Texture and Soilgrids

The International Soil Reference and Information Centre (ISRIC) has produced a global harmonised soil properties database called SoilGrids[25]. 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 1km 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.

Topography

The Multi-Error-Removed Improved-Terrain (MERIT) Digital Elevation Model (DEM) topography product was used for this study[26]. This product has a spatial resolution of ~90m.

2.5 In-Situ soil moisture measurements

Ground truth data for training the models were obtained from in-situ SWC measurements at sites distributed from two networks 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[27]. Ameriflux is a network of flux towers spread across North America recording various atmospheric and meteorological data and fluxes[28]. 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 downloaded and used in this study. 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 quality, the Ameriflux data was pruned to only contain readings with similar properties to ISMN readings with a 'G' quality flag. This means Ameriflux samples were dropped if either the LST reading was below 3°C or the SWC reading was above 0.7 $\mathbf{m^3/m^3}$. Additionally, sites in wetland and chronically inundated regions were excluded from the dataset.
SWC measurements are then aggregated to daily averages.

2.6 Datasets

The primary dataset is comprised of all available data from ISMN and Ameriflux soil moisture measurements within the temporal and spatial boundaries. Each location is classified by soil texture class. For each soil texture class, 80% of sites and all of the samples belonging to them are moved to a training set and the remaining 20% of sites and their samples are sent to the validation set. This split makes certain that not only are the validation data samples unseen by training, but they are also locations not seen by the model. This ensured that we can generalize the results to the greater CONUS area. Each daily aggregate of in-situ measurements is accompanied by daily aggregate measurements for the covariate inputs. The final dataset is comprised of 657,935 samples and 1054 stations. 206 of which were moved into the validation dataset. For further validation, two more datasets comprising a small network of soil moisture stations, originally used to calibrate SMAP, will be used to assess performance. Further discussion of their contents can be found in the supplementary document.

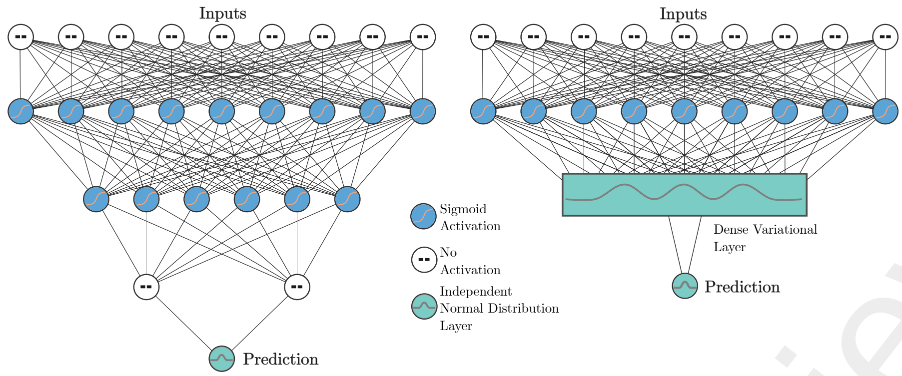
Next, we will look at how the information within the datasets is utilized to train the ensembles.

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. This study looks at the performance of four different ensemble architectures. Two of the ensembles are replications of the architectures used by Abazsddeh (RF) and Xu (WDL). The remaining two models are simple distance based models. The first being a feed-forward network (Dense) and the other using a probabilistic layer (Prob). Both of their architectures were chosen so as to have almost the same number of hidden parameters. The architectures of the two smaller networks and WDL architectures can be seen in Figures 2 and 3 respectively. More detailed descriptions of their architectures can be found in the supplement.

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

Table 1: All of the categorical land characteristic subclasses.



(a) Dense Architecture (b) Prob Architecture

Figure 2: Probabilistic model architectures

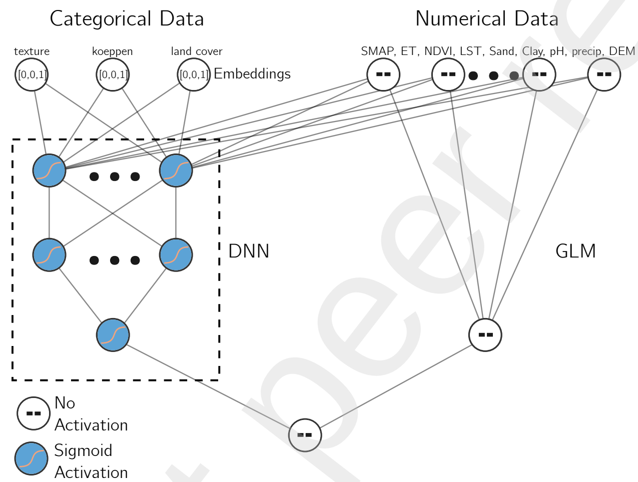


Figure 3: WDL Architecture

3.1 Training

In this study, we assume that static variables as seen in Table 1 either aide or hinder the models ability to discern SWC. Since these variables are not balanced in the dataset, the model may focus on the most abundant subclass types while neglecting to learn how to predict on other underrepresented subclasses. To account for these imbalances, instead of additional data manipulation, a simple approach is undertaken in the form of ensembles. Each ensemble member is trained with sample weights accounting for imbalances within a static characteristic. For example, an ensemble member trains on data weighted to the different soil texture class abundances giving extra weight/importance to correctly predicting the less abundant texture types. For the Dense, Probabilistic, and WDL ensembles, those static characteristics are **texture**, **clay** and **sand content**, **Köppen climate class**, **land cover class**, and an **unweighted** category that does not use any balancing. Therefore, there are 7 members per ensemble (one per characteristic) as seen in Fig. 4.

The weighting scheme for each static class 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 doesn't use sample weights. Instead, balance is accounted for by training a unique model for each soil texture domain as done by Abbaszadeh et al.[17]. 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 to learn from due to less gaps from cloud cover. This means 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. Although this is not ideal, the rationale is that SMAP would account for the temporal variation in SWC while the other variables would 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. This loss of temporal information seems to be offset by the increase in samples to learn from and is discussed further in the supplement document.

3.2 Predictions

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

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

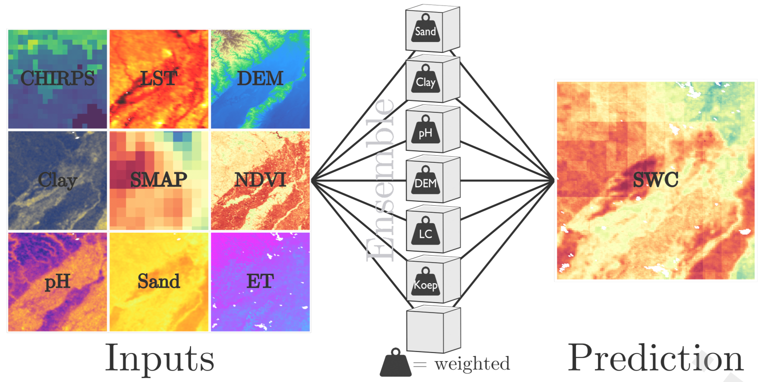


Figure 4: Prediction regime for the Dense, Prob, and WDL ensembles. Each ensemble member (cube) is trained while weighted against imbalances in a specific characteristic. These predictions are then averaged to provide an ensemble prediction.

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).

When making spatial predictions, spatial data are resampled to the highest resolution (90m) using nearest neighbor interpolation. This prevents interpolation error, but introduces some pixelation at higher levels of zoom.

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 [6]. 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 dictates 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 a large dataset comprising the validation data set aside during training. The second and third comprise smaller networks of soil moisture stations located in Oklahoma. Predictions will be compared against in-situ measurements as well as the predictions made by SMAP at that location.

4.1 CONUS Dataset

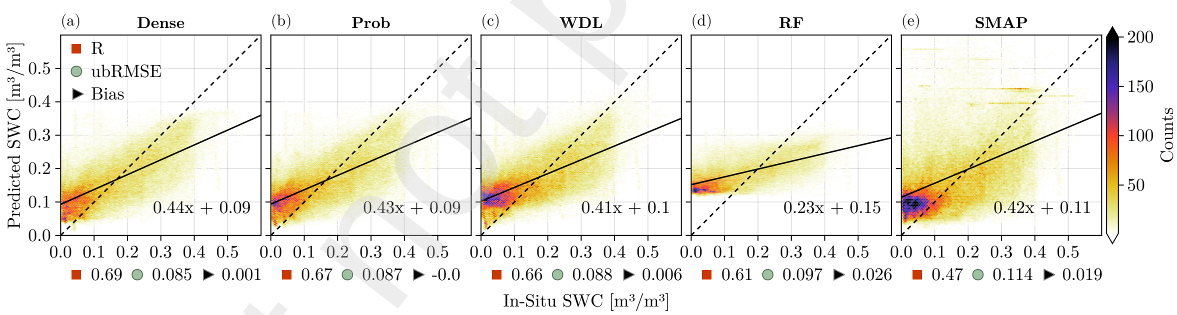


Figure 5: Heatmaps and metrics for algorithm predictions on the validation dataset as a whole.

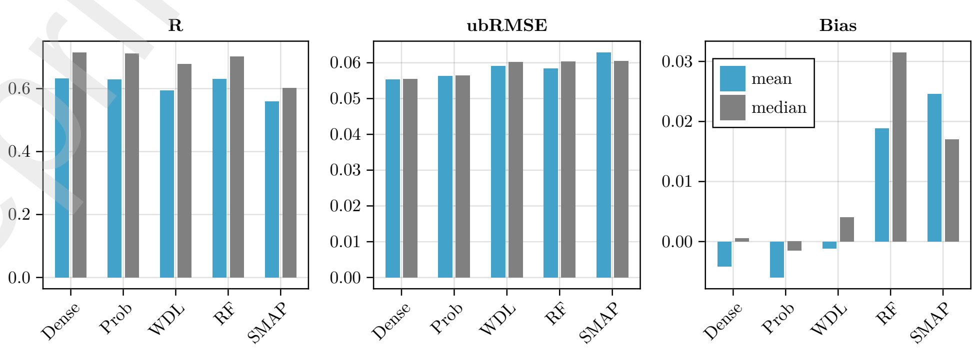


Figure 6: The average metric score for every site in the validation dataset. (a) numerically (b) visually

Because downscaling is an attempt at spatial prediction and reasoning, it's 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.

As shown in Fig. 5, every method 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 $0.18m^3/m^3$. Because of this, the lowest SWC prediction by the RF ensemble on the entire dataset is $0.10m^3/m^3$. 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 ensembles again outperform SMAP on every metric with exception to RFs bias. This is displayed in Figure 6. In the same figure we also see that timeseries are less consistent from site to site as the mean is notably lower than the median, but the ubRMSE shows a strong agreement between mean and median values demonstrating general consistency for prediction accuracy. Overall, this suggests all methods and their predictions should be as reliable or moreso than SMAP.

4.1.1 Spatial Predictions

To compare the spatial predictions of each method, a 1°x 1°box is cut out around a specific in-situ location on a summer day with the least cloud cover. Of the resulting predictions, six examples that exhibit unique characteristics are presented, two of which are highlighted in Figure 7. Overall, the ensembles tend to exhibit similar spatial patterns. In some cases, as exhibited in the predictions around *PBO: H2O_LITTLELOST*, the 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. The other four examples can be seen and are discussed in the supplement.
Next we will look at the ensembles predictions over time.

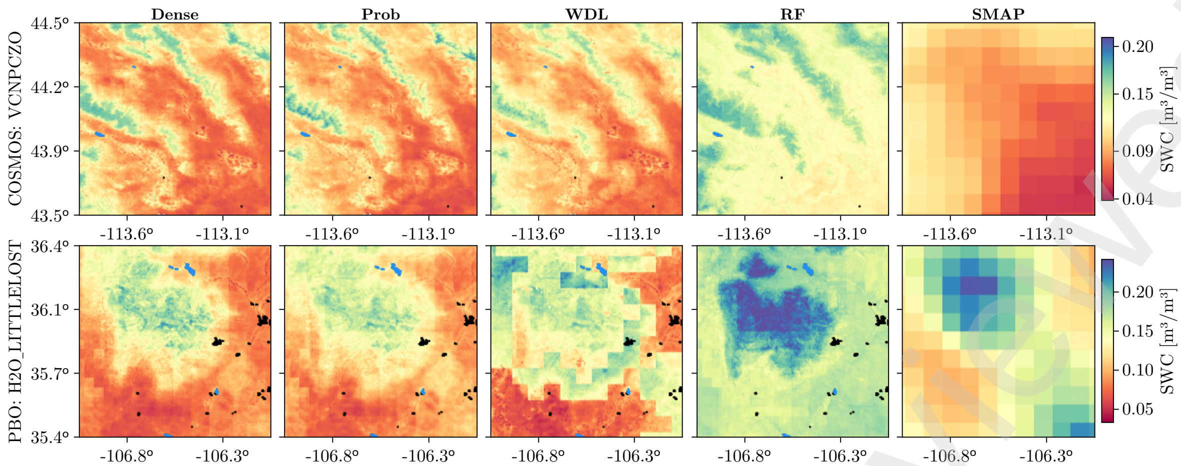


Figure 7: 1°x 1°spatial SWC predictions of ensembles vs SMAP. Black pixels represent pixels masked as 'urban' and blue pixels are water surfaces.

4.1.2 Temporal Predictions

Although the R metric is calculated for each site in the validation set, it's also important to view the time-series plotted against each other. For this analysis, the ten sites with the most data were selected and the time-series from 2018 is plotted. One of which is seen in Figure 8. The same figure also shows the R scores for the validation dataset on each station. Here we can see that the two top performing models in this metric (Dense and RF) both have drastically tightened distributions for R values compared to SMAP. Despite RF having similar performance to Dense, it's clear in the additional timeseries found in the supplement that RF possesses a strong bias and is often distinct from the SMAP, Dense, and in-situ markers. In general, the timeseries predictions of all models are as good or better than those of SMAP.

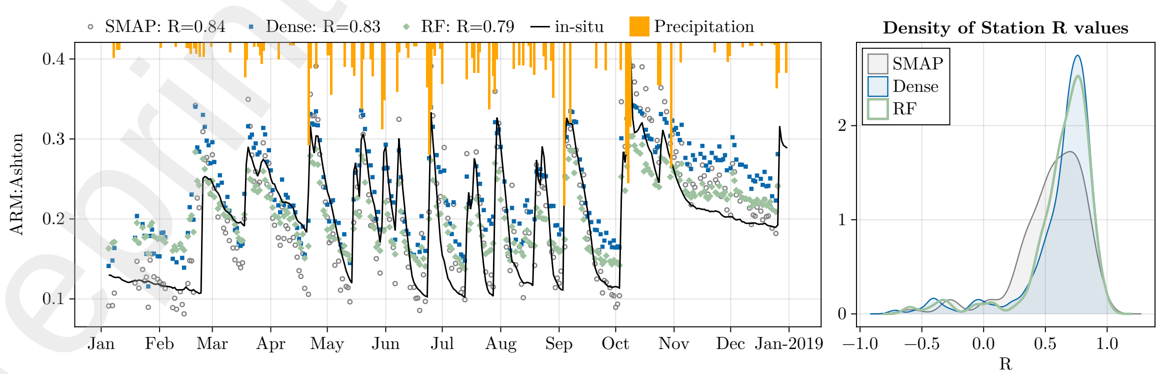


Figure 8: (Left) Temporal predictions on a station in the validation dataset. (Right) Density plot of the R values for each station in the validation dataset.

In the next subsection we will look at the performance of the ensembles on two additional test datasets.

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[6] 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.

Washita

The first dataset is the Washita River basin network. In this region, all methods struggle on the Washita dataset as a whole as seen in Fig 9. All methods 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.

	Dense	Prob	WDL	RF	SMAP
R	**0.752**	0.661	0.681	0.700	0.745
ubRMSE	**0.041**	0.062	0.046	0.044	0.046
Bias	0.053	0.246	0.076	**0.006**	0.011

Table 2: Average site metric scores on the Washita dataset

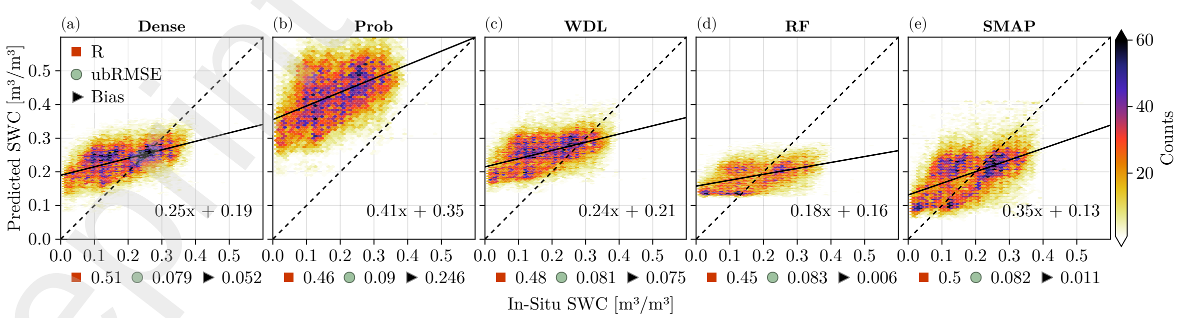


Figure 9: Heatmaps and metrics for algorithm predictions on the Washita dataset as a whole.

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. SMAP also exhibits good performance as expected. The other methods 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 is composed of measurements from the Fort Cobb network. Due to it's close proximity to Washita, its no suprise that we see similar trends. All methods demonstrate poor fitting to the dataset as a whole and the models show a strong positive bias at low SWC measurements. The RF model yields the best bias metric, although likely due to values being squashed towards a mean value.

	Dense	Prob	WDL	RF	SMAP
R	0.748	0.708	0.673	0.704	**0.752**
ubRMSE	**0.042**	0.049	0.043	0.043	0.046
Bias	**0.060**	0.116	0.079	0.062	0.062

Table 3: Average site metric scores on Fort Cobb dataset

Again, the model performance metrics increase on a site level (Table 3). The dense model is the closest method to the 0.04 m^3/m^3 ubRMSE threshold established by the SMAP mission. RF also scores within the realms of acceptability for this metric. The Prob and WDL models are unable to outperform SMAP on any metric with SMAP having the best R score.

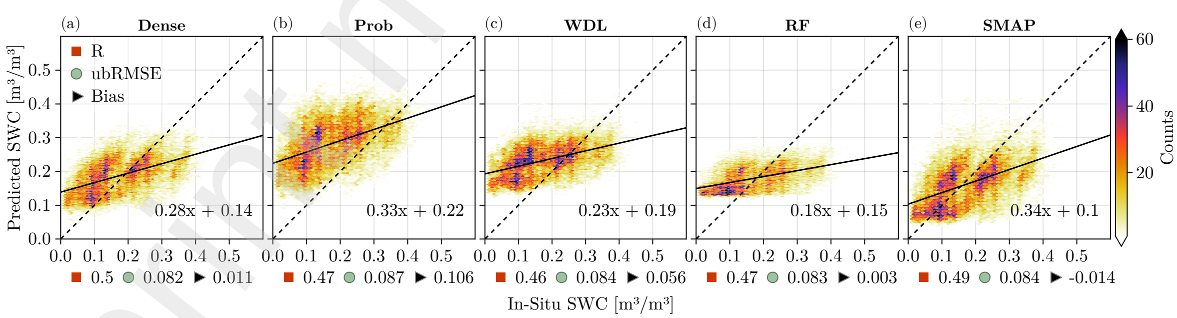


Figure 10: Heatmaps and metric scores for algorithm predictions on the Fort Cobb dataset as a whole.

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 method can reliably match or outperform SMAP here, it would suggest confidence in it's ability to perform elsewhere. The Dense architecture is the only method to reliably match or exceed SMAP on key metrics on these datasets.

Timeseries

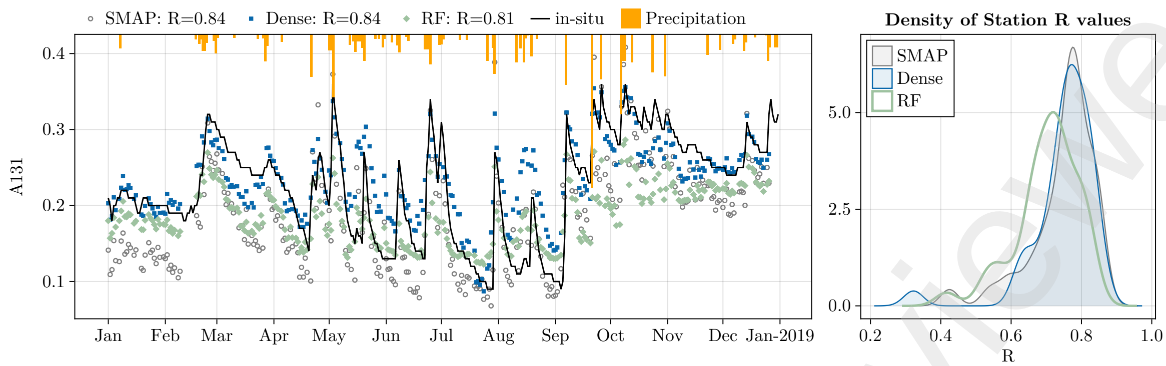


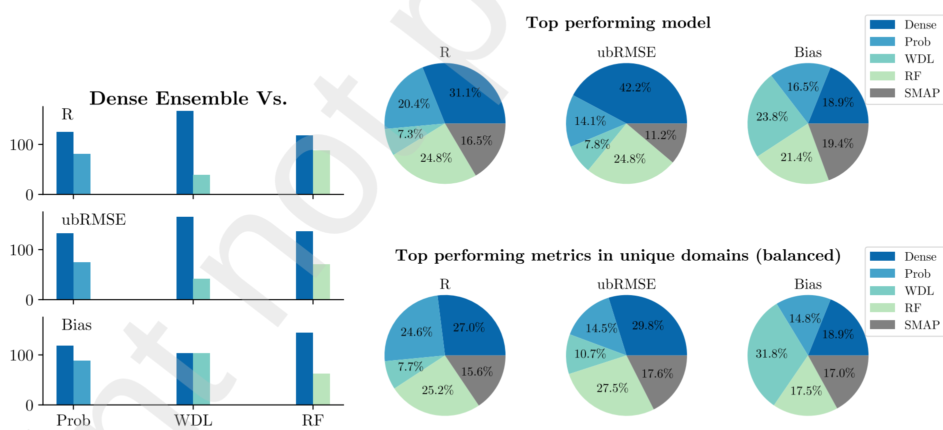
Figure 11: (Left) Temporal predictions on a station in the validation dataset. (Right) Density plot of the R values for each station in both OK datasets.

Similar to the timeseries predictions for the validation set. Timeseries predictions from the Oklahoma dataset help assure us that models are maintaining consistency through time. SMAP has a home field advantage at these sites and only the Dense architecture is able to demonstrate parity and match SMAPs strong temporal accuracy. A timeseries of a station in the Washita dataset is plotted in Figure 11 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 and the Dense peak is a bit below that of SMAP.

In the next section we will analyze the robustness of the results and look for potential limitations.

4.3 Top performer

We can evaluate performance based on three criteria: dataset, sites, and domains. We saw in the previous sections that the Dense model was consistently a top performer on datasets, but what about site and domain? For site level, we compare the Dense predictions on each site against the other architectures in the validation dataset. In this context, the Dense architecture outperforms every other model in every other metric as seen in Fig. 12a with the exception of the bias against WDL. In a head-to-head competition of all methods, Dense is the clear winner in ubRMSE and notable winner in R. WDL maintains the best method for bias. To see if Dense is still the top performer by domain, we look at each models performance on stations belonging to the subclasses of each categorical land surface attribute as seen in Table 1. Performance is then normalized so over/underrepresented classeas have equal impact on performance. This normalizing method is discussed further in future sections. When normalizing for class type and abundance, we can see (Fig. 12b) the Dense model is still the most consistent performer for R and ubRMSE. However, this is only slightly more dominant than the RF ensemble. WDL is again the clear top performer for bias.



(a) (b)

Figure 12: (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 category in Table 1. Each time a model outperforms every other method for that metric it gets a point. All points for that category are normalized so that the top performer receives one point for that category. All points are summed together for all categories. This produces an unbiased assessment of model performance regardless of imbalances in representation of classes.

Having a distance based model outperform the RF has additional advantages. For starters the evaluation speed for distance based models is two orders of magnitude faster (0.16s vs 17.7s on 130k samples). Therefore, it's more feasible to predict over large domains. Additionally, the file size of the RF ensemble is three orders of magnitude larger (2.3GB vs 1.03MB) which makes transferring it less convenient than the simple distance based ensembles. For these reasons, it doesn't seem reasonable to continue using a RF architecture for this task at this resolution.

Next we will look to see how generalizable the performance of the models are for different land surface characteristics.

4.4 Domain Preference

To further explore areas of strengths and weakness', metrics are calculated across each of the three categorical static characteristics: **texture**, **climate class**, and **land cover**. These static characteristics are further broken down into the subclasses previously shown in Table 1. A significant drop in metric performance in one of these subclasses may indicate an inability for a model to fully generalize SWC from the input variables. To search for these preferences/weaknesses we compute the average metric score for a method on each station in the 40 subclasses from Table 1. We then divide this by the average performance for all models on that subclass. This final value gives us the relative performance of a model compared to all others. If any models performance is at least 10% better or worse than the mean score for all models on that subclass, then that model is deemed to have a bias for that subclass. These instances are seen in Table 4. The Bias metric was excluded as the RF model consistently exhibited poor bias. The only instance where a model demonstrates a negative or positive performance on both ubRMSE and R was on Sand. Here, the Dense R value is 40% the mean R value and the ubRMSE is 124% the mean ubRMSE value. This category constitutes only one stations worth of data and so no conclusions can be made about the models performance on sand overall.

Although there doesn't appear to be any strong or negative biases for any single static characteristics, what if there exists a combination of inputs that exhibit difficulties? The next section will explore for just such an instance.

Characteristic	Dense	Prob	WDL	RF	No. of Stations
			R		
SiClLo	1.07	1.05	**0.83**	1.05	3
Mxd Frsts	1.08	0.98	**0.89**	1.04	3
Bsh	1.04	1.05	**0.88**	1.02	2
Sa	**0.44**	1.21	1.17	1.18	1
			ubRMSE		
Csa	0.92	0.99	**1.10**	0.98	24
Opn Shrblnds	0.94	1.01	**1.14**	0.91	6
SaClLo	1.03	1.04	1.04	**0.89**	3
Bsh	0.95	**1.14**	0.94	0.91	2
ET	1.00	**1.14**	0.94	0.92	2
BWh	0.99	1.13	0.99	**0.90**	1
Sa	**1.24**	**0.71**	1.05	1.00	1
Cl	**0.85**	1.03	1.09	1.03	1

Table 4: Static classes where one model displays a bias (an average metric score on that class which deviates 10% or more from the mean of all models) for that specific class. For R, values greater than 1.0 outperform the mean, for ubRMSE values below 1.0 outperform the mean. No. of stations represents number of locations possessing that characteristic

4.5 Areas of Underperformance

To find combinations of characteristics that exhibit underperformance, the static characteristics for each site in the CONUS dataset were compiled into a dataset with six dimensions (sand, clay, pH, topography, climate class, land cover type) whose values were normalized for each dimension. 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 Figure 13. This area in the PCA represents Cropland land cover type with high clay content and low sand content as seen in Table 5. These values are scaled by the standard deviation of the dataset for each static characteristic. 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 standard deviations below the mean. Most of these sites are croplands.

This brief analysis shows that the best performing model (Dense) does not have consistent performance on croplands of high clay and low sand content values. Therefore, this method 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.

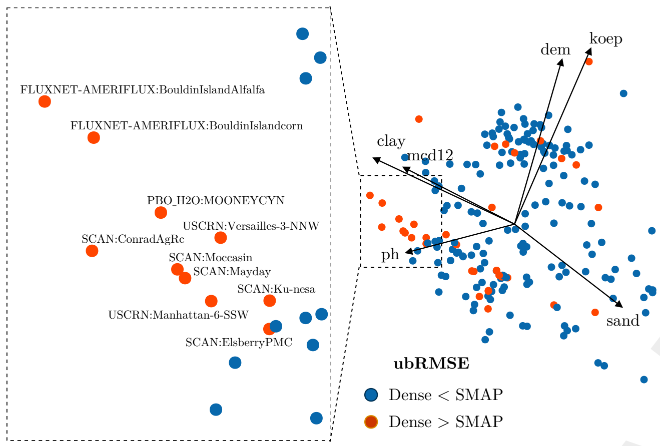


Figure 13: Reprojection of test data static characteristics into PCA space. Peach dots represent sites where the Dense ensemble's ubRMSE score was worse than SMAP

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

Table 5: The deviations from mean values for static characteristics at the site level

4.6 Cross-fold Analysis

In order to assess whether our methodology is generalizable. A 10-fold cross validation was conducted. This involved splitting the original dataset into 10 separate datasets containing 10% of the total stations and their respective data. For each of these 10 datasets, the ensembles are trained on the other 90% and then predict the in-situ values for those left out. These datasets are produced randomly and so their proportions of different static characteristics is not curated. This randomness may have a negative impact on the RF ensemble as it has no weighting scheme to account for the imbalances it will learn from.

In general, the metrics from the cross validation are similar to those achieved in the validation set. The exception being the RF ensemble. This is likely due to the RF method relying on needing some information from each texture class. But not every cross validation subset has every texture to learn from. The density curves for the R values for each station in the cross validation dataset are plotted in Figure 14. Compared to SMAP, the Dense and Prob methods (the two strongest performers) have

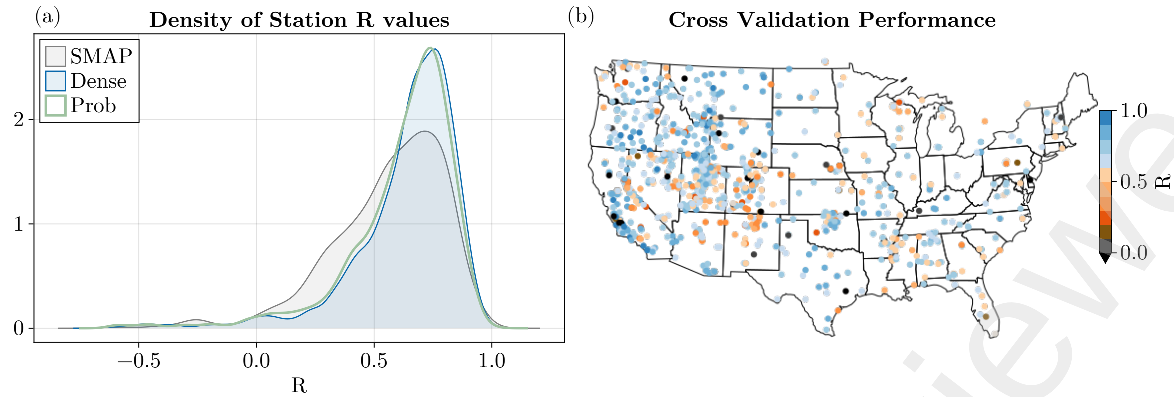


Figure 14: (a) Density plots of the Dense and Prob R values for each station in the cross validation dataset. (b) Spatial distribution of R values on each station as predicted by Dense

their distributions tightened over higher R values. This was also the case for the WDL and RF (seen in supplement), but the RF distribution is notably less impressive as expected. Density plots for ubRMSE show improvement from SMAP in all methods except with RF and can be found in the supplement. For the weighted methods (Dense, PRob, WDL), the cross validation appears to confirm that the weighting scheme limits biases in the training data.

Model	*Dataset*	R	*ubRMSE*	*Bias*
Dense	Val	0.632	**0.055**	-0.004
	Cross Val	**0.639**	0.058	**-0.000**
Prob	Val	**0.628**	**0.056**	**-0.007**
	Cross Val	0.621	0.060	-0.008
WDL	Val	0.594	**0.059**	**-0.001**
	Cross Val	**0.611**	0.060	-0.003
RF	Val	**0.630**	**0.058**	0.019
	Cross Val	0.572	0.065	**0.004**
SMAP	Val	0.559	**0.063**	0.025
	Cross Val	**0.562**	0.065	**0.023**

Table 6: The mean metric score for each method on each station on the validation set vs the cross validation dataset

5 Discussion

The primary focus for this section is to evaluate the the robustness and generalizability of the methods. Additionally, we want to look at the ensemble framework in context of this work and identify whether or not there is any advantage from an ensemble prediction, or if we can achieve equally satisfactory results with just a single ensemble member.

5.1 Generalizability

Large domain predictions only yield value if we can trust that those predictions are generalizeable, or consistently accurate, across the hetereogeniety of the domain. To test whether these ensemble predictions can extrapolate beyond their training dataset, we ensured that validation data belonged to locations previously unseen and foreign to the models. After analysis yielded no concerning biases or shortcomings, we then conducted a crossfold analysis across all sites in the training and validation set. Again, we see consistent/similar performance on each site when it was previously unseen during training. The last form of analysis involved monitoring spatial predictions and their associated SHAP values. This analysis is discussed further in the supplement. We find that the SHAP values generally adhere to expectations found in literature, however strangely all methods seem to have an inverse relationship for NDVI from what is expected. Further analysis was not conducted to discern why this was the case.

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

5.2 Ensemble Advantage

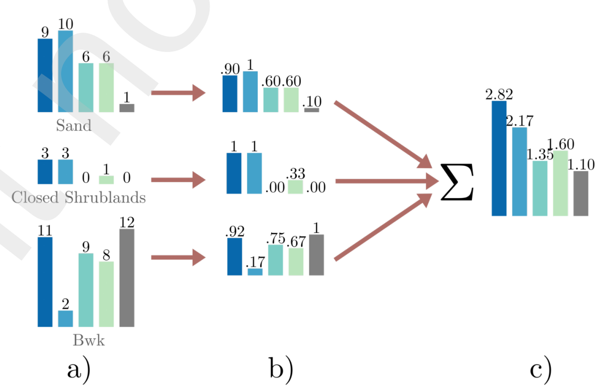


Figure 15: Weighting schema for unbiased top performers. a) All models predict on all sites belonging to a specific category. Each time a model outperforms every other model it gets a point. b) Points are then normalized. This ensures under-represented categories have equal importance in assessing model performance. c) The normalized points are summed providing a final assessment of model performance on all categories.

This study serves to assess the feasibility and advantage of using an ensemble of models to predict SWC at higher resolutions. In the case of the two probabilistic ensembles (Dense and Prob), they

Model	Metric	*Ens.*	Sand	Clay	Koep	MCD12	Free	pH	Texture
	R	**0.632**	0.621	0.615	0.607	0.618	0.631	0.613	0.558
Dense	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
	R	**0.629**	**0.629**	0.620	0.592	0.618	0.623	0.613	0.596
Prob	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**
	R	0.594	0.594	**0.598**	0.586	0.594	0.594	0.586	0.589
WDL	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

Table 7: Average station performance for each ensemble member and the ensemble as a whole on the validation dataset.

represent exceedingly simple models. The purpose of these ensembles is to permit equal representation for all unique land characteristics in the training process as to prevent overfitting to a dominant characteristic. However, perhaps the weighting scheme for one land characteristic may be a sufficient representation of the data and an ensemble is redundant.

First we compare the average performance of each ensemble member against the ensemble in the validation dataset. This is seen in Table 7. Here, we can see that for the Dense ensemble, the ensemble is only marginally better than its unweighted member. Whereas for the Prob and WDL ensembles, the Sand and Clay weighted members outperformed their respective ensembles. In all instances the ensembles average performance is not significantly improved upon when compared to the unweighted member.

To ensure that there isn't a dominant subclass that is easy to predict for both ensemble and members, we compare the ensembles performance on static domains against every ensemble member. In other words, for each texture/land cover/Koeppen class listed in Table 1, we compare the prediction performance of individual ensemble members versus the full ensemble on that subset of data. For each site a model outperforms the other, their score for that class increases. The two scores for that class are normalized so that the model that outperforms on the most sites receives a value of 1. This process is illustrated in Fig. 15. This is done for each metric (R, ubRMSE, Bias). These final scores are summed and these final sums represent the total normalized performance ratio for that ensemble vs ensemble member pairing. These final normalized performance ratios for each ensemble-member pairing are visualized in Fig. 17.

When looking at these unbiased performances across subclasses, 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 architecture is likely too simple to overfit a characteristic, and the GLM of the WDL seems to be adept at guiding predictions and preventing overfitting. From a purely numerical context, there does not exist a clear ensemble advantage.

Lastly, we compare the spatial predictions of the ensemble vs the unweighted ensemble member. Here there exists a much starker difference in behaviour. Namely, the Dense ensemble predictions seem to capture more of the land surface characteristics than the single ensemble member. This is seen in Figure 16. Although not directly quantifiable, it is clear that the Ensemble is able to incorporate more of the land surface characteristics into it's prediction than the unweighted ensemble member. This however, is not the case for the Prob architecture. The single ensemble member for Prob seemed do distinguish the same land characteristic fidelity as the ensemble. For the WDL architecture, ensemble member prediction is noisier than the ensemble. Further analysis will need to be conducted to asses whether these behaviours constitutes a substantial improvement of one over the other.

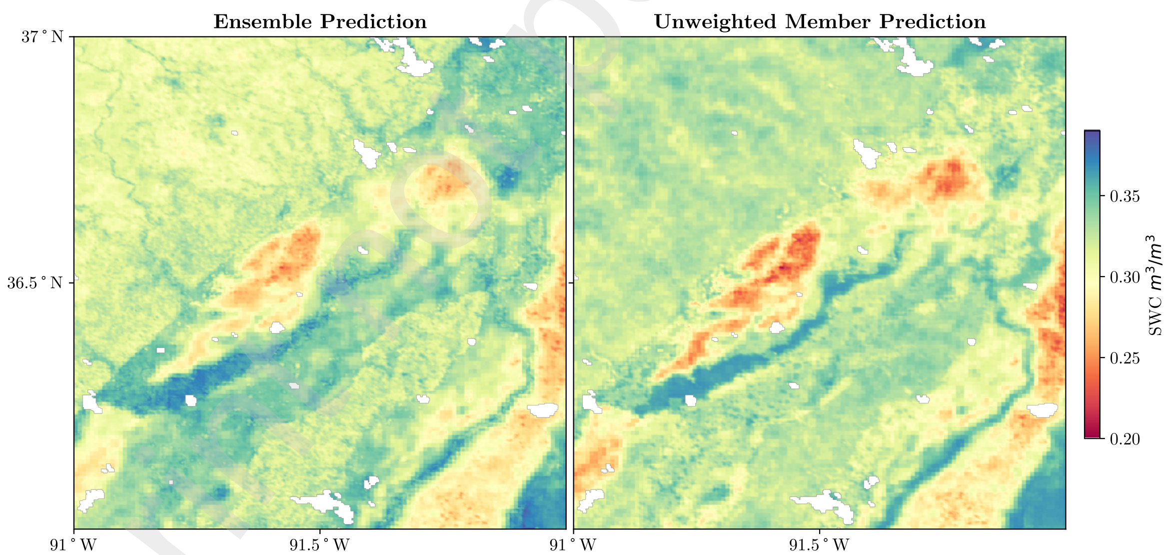


Figure 16: Spatial Predictions comparing the Dense ensemble vs the unweighted (Free) ensemble member

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

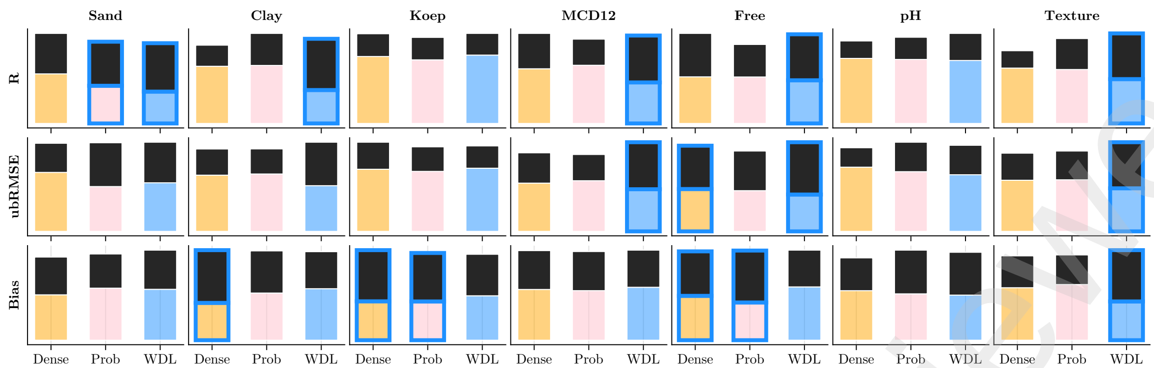


Figure 17: Head to head comparison of Ensembles (Bottom label) vs their member constituents (Top label) with normalized performances. Bars highlighted in blue indicate an instance where an ensemble member outperformed the ensemble on that metric (Left label). An explanation of this head to head competition is seen in Figure 15

6 Conclusion

The work conducted in this paper served to demonstrate that an ensemble of simple ML architecture can yield acceptable SWC downscaling results. Analysis revealed that these ensembles can reliably do this with strong generalizability. However, certain ensemble members can outperform or achieve parity with the full ensemble on the validation dataset. This suggests there is no/little benefit one would achieve from an ensemble that one would not also achieve with a rigorous sample weighting scheme. Despite this, Comparison of the spatial predictions between Ensembles vs these seemingly similarly performing members showed that ensembles appear to capture more of the land surface characteristics. More analysis is needed to assess whether or not this is advantageous and by how much. Multi-variable analysis of ensemble 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 ensemble which needs curated dated to learn from and so struggles on the random crossfold data.

Final summary:

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

6.1 Acknowledgements

The research was generously support by funding from the Max Planck Institute of Biogeochemistry with support from the International Max Planck Research School for Global Biogeochemical Cycles. 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.

Competing Interests

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

References

[1] Laibao Liu, Lukas Gudmundsson, Mathias Hauser, Dahe Qin, Shuangcheng Li, and Sonia I. Seneviratne. Soil moisture dominates dryness stress on ecosystem production globally. *Nature Communications*, 11(1):4892, December 2020.

[2] Zheng Fu, Philippe Ciais, I. Colin Prentice, Pierre Gentine, David Makowski, Ana Bastos, Xiangzhong Luo, Julia K. Green, Paul C. Stoy, Hui Yang, and Tomohiro Hajima. Atmospheric dryness reduces photosynthesis along a large range of soil water deficits. *Nature Communications*, 13(1):989, December 2022.

[3] Benjamin D. Stocker, Jakob Zscheischler, Trevor F. Keenan, I. Colin Prentice, Sonia I. Seneviratne, and Josep Peñuelas. Drought impacts on terrestrial primary production underestimated by satellite monitoring. *Nature Geoscience*, 12(4):264–270, April 2019.

[4] Marco Bittelli. Measuring Soil Water Content: A Review. *HortTechnology*, 21(3):293–300, June 2011.

[5] Kaijun Wu, Gabriela Arambulo Rodriguez, Marjana Zajc, Elodie Jacquemin, Michiels Clément, Albéric De Coster, and Sébastien Lambot. A new drone-borne GPR for soil moisture mapping. *Remote Sensing of Environment*, 235:111456, December 2019.

[6] Dara Entekhabi. *SMAP Handbook Soil Moisture Active Passive.* JPL Publication JPL, 2014.

[7] Peggy E. ONeill, Steven Chan, Eni G. Njoku, Tom Jackson, and Rajat Bindlish. SMAP Enhanced L3 Radiometer Global Daily 9 km EASE-Grid Soil Moisture, Version 3, 2019.

[8] Noemi Vergopolan, Justin Sheffield, Nathaniel W. Chaney, Ming Pan, Hylke E. Beck, Craig R. Ferguson, Laura Torres-Rojas, Felix Eigenbrod, Wade Crow, and Eric F. Wood. High-Resolution Soil Moisture Data Reveal Complex Multi-Scale Spatial Variability Across the United States. *Geophysical Research Letters*, 49(15):e2022GL098586, August 2022.

[9] Bibi S. Naz, Wolfgang Kurtz, Carsten Montzka, Wendy Sharples, Klaus Goergen, Jessica Keune, Huilin Gao, Anne Springer, Harrie-Jan Hendricks Franssen, and Stefan Kollet. Improving soil moisture and runoff simulations at 3 km over Europe using land surface data assimilation. *Hydrology and Earth System Sciences*, 23(1):277–301, January 2019.

[10] Brahima Koné, Arona Diedhiou, Adama Diawara, Sandrine Anquetin, N'datchoh Evelyne Touré, Adama Bamba, and Arsene Toka Kobea. Influence of initial soil moisture in a regional climate model study over West Africa – Part 1: Impact on the climate mean. *Hydrology and Earth System Sciences*, 26(3):711–730, February 2022.

[11] Brahima Koné, Arona Diedhiou, Adama Diawara, Sandrine Anquetin, N'datchoh Evelyne Touré, Adama Bamba, and Arsene Toka Kobea. Influence of initial soil moisture in a regional climate

model study over West Africa – Part 2: Impact on the climate extremes. *Hydrology and Earth System Sciences*, 26(3):731–754, February 2022.

[12] Andreas Colliander, Joshua B. Fisher, Gregory Halverson, Olivier Merlin, Sidharth Misra, Rajat Bindlish, Thomas J. Jackson, and Simon Yueh. Spatial Downscaling of SMAP Soil Moisture Using MODIS Land Surface Temperature and NDVI During SMAPVEX15. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2107–2111, November 2017.

[13] Nitu Ojha, Olivier Merlin, Christophe Suere, and Maria José Escorihuela. Extending the Spatio-Temporal Applicability of DISPATCH Soil Moisture Downscaling Algorithm: A Study Case Using SMAP, MODIS and Sentinel-3 Data. *Frontiers in Environmental Science*, 9:555216, March 2021.

[14] Jingyao Zheng, Haishen Lü, Wade T. Crow, Tianjie Zhao, Olivier Merlin, Nemesio Rodriguez-Fernandez, Jiancheng Shi, Yonghua Zhu, Jianbin Su, Chuen Siang Kang, Xiaoyi Wang, and Qiqi Gou. Soil moisture downscaling using multiple modes of the DISPATCH algorithm in a semi-humid/humid region. *International Journal of Applied Earth Observation and Geoinformation*, 104:102530, December 2021.

[15] Juan M. Sánchez, Joan M. Galve, José González-Piqueras, Ramón López-Urrea, Raquel Niclòs, and Alfonso Calera. Monitoring 10-m LST from the Combination MODIS/Sentinel-2, Validation in a High Contrast Semi-Arid Agroecosystem. *Remote Sensing*, 12(9):1453, May 2020.

[16] Nitu Ojha, Olivier Merlin, Beatriz Molero, Christophe Suere, Luis Olivera-Guerra, Bouchra Ait Hssaine, Abdelhakim Amazirh, Ahmad Al Bitar, Maria Escorihuela, and Salah Er-Raki. Stepwise Disaggregation of SMAP Soil Moisture at 100 m Resolution Using Landsat-7/8 Data and a Varying Intermediate Resolution. *Remote Sensing*, 11(16):1863, August 2019.

[17] Peyman Abbaszadeh, Hamid Moradkhani, and Xiwu Zhan. Downscaling SMAP Radiometer Soil Moisture Over the CONUS Using an Ensemble Learning Method. *Water Resources Research*, 55(1):324–344, January 2019.

[18] Mengyuan Xu, Ning Yao, Haoxuan Yang, Jia Xu, Annan Hu, Luis Gustavo Goncalves de Goncalves, and Gang Liu. Downscaling SMAP soil moisture using a wide & deep learning method over the Continental United States. *Journal of Hydrology*, 609:127784, June 2022.

[19] Hongfei Zhao, Jie Li, Qiangqiang Yuan, Liupeng Lin, Linwei Yue, and Hongzhang Xu. Downscaling of soil moisture products using deep learning: Comparison and analysis on Tibetan Plateau. *Journal of Hydrology*, 607:127570, April 2022.

[20] Carsten Montzka, Kathrina Rötzer, Heye Bogena, Nilda Sanchez, and Harry Vereecken. A New Soil Moisture Downscaling Approach for SMAP, SMOS, and ASCAT by Predicting Sub-Grid Variability. *Remote Sensing*, 10(3):427, March 2018.

[21] Ahmed Samir Abowarda, Liangliang Bai, Caijin Zhang, Di Long, Xueying Li, Qi Huang, and Zhangli Sun. 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, March 2021.

[22] Wei Xu, Zhaoxu Zhang, Zehao Long, and Qiming Qin. Downscaling SMAP Soil Moisture Products With Convolutional Neural Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4051–4062, 2021.

[23] Yulin Cai, Puran Fan, Sen Lang, Mengyao Li, Yasir Muhammad, and Aixia Liu. Downscaling of SMAP Soil Moisture Data by Using a Deep Belief Network. *Remote Sensing*, 14(22):5681, November 2022.

[24] Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1):150066, December 2015.

[25] Tomislav Hengl, Jorge Mendes De Jesus, Gerard B. M. Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N. Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, Mario Antonio Guevara, Rodrigo Vargas, Robert A. MacMillan, Niels H. Batjes, Johan G. B. Leenaars, Eloi Ribeiro, Ichsani Wheeler, Stephan Mantel, and Bas Kempen. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2):e0169748, February 2017.

[26] Dai Yamazaki, Daiki Ikeshima, Ryunosuke Tawatari, Tomohiro Yamaguchi, Fiachra O'Loughlin, Jeffery C. Neal, Christopher C. Sampson, Shinjiro Kanae, and Paul D. Bates. A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853, June 2017.

[27] W. A. Dorigo, W. Wagner, R. Hohensinn, S. Hahn, C. Paulik, A. Xaver, A. Gruber, M. Drusch, S. Mecklenburg, P. Van Oevelen, A. Robock, and T. Jackson. 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, May 2011.

[28] T. A. Boden, M. Krassovski, and B. Yang. The AmeriFlux data activity and data system: an evolving collection of data management techniques, tools, products and services. *Geoscientific Instrumentation, Methods and Data Systems*, 2(1):165–176, June 2013.