DeepSoil: A Science-guided Framework for Generating High Precision Soil Moisture Maps by Reconciling Measurement Profiles Across In-situ and Remote Sensing Data

Paahuni Khandelwal
Colorado State University
Fort Collins, Colorado, USA
Paahuni.Khandelwal@colostate.edu

Sangmi Lee Pallickara
Colorado State University
Fort Collins, Colorado, USA
Sangmi.Pallickara@colostate.edu

Shrideep Pallickara
Colorado State University
Fort Collins, Colorado, USA
Shrideep.Pallickara@colostate.edu

Abstract

Soil moisture plays a critical role in several domains and can be used to inform decision-making in agricultural settings, drought forecasting, forest fire predictions, and water conservation. Soil moisture is measured using in-situ and remote-sensing equipment. Depending on the type of equipment that is used, some challenges must be reconciled, including the density of observations, the measurement precision, and the resolutions at which these measurements are available. In particular, in-situ measurements are high-precision but sparse, while remote sensing measurements benefit from spatial coverage, albeit at lower precision and coarser resolutions. The crux of this study is to produce higher-precision soil moisture estimates at high resolutions (30m). Our methodology combines scientific models, deep networks, topographical characteristics, and information about ambient conditions alongside both in-situ and remote sensing data to accomplish this. Domain science infuses several aspects of our methodology. Our empirical benchmarks profile several aspects and demonstrate that our methodology accounts for spatial variability while accounting for both static (soil properties and elevation) and dynamically varying phenomena to generate accurate, high-precision 30m resolution soil moisture content maps.

CCS Concepts

• **Information systems** → *Geographic information systems*; • **Computing methodologies** → **Neural networks**; *Spatial and physical reasoning*.

Keywords

science-guided learning, KGML, big data, spatiotemporal phenomena, soil moisture, and deep neural networks

ACM Reference Format:
Paahuni Khandelwal, Sangmi Lee Pallickara, and Shrideep Pallickara. 2024. DeepSoil: A Science-guided Framework for Generating High Precision Soil Moisture Maps by Reconciling Measurement Profiles Across In-situ and Remote Sensing Data. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3678717.3691261

SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1107-7/24/10
https://doi.org/10.1145/3678717.3691261

1 Introduction

There has been a proliferation of in-situ and remote sensing equipment in several geospatial domains such as atmospheric sciences, agriculture, environmental and ecological monitoring, etc. These systems are used extensively to monitor and understand diverse phenomena[3, 69]. The devices report measurements at different frequencies, precision, and spatial resolutions.

Remote sensing systems provide excellent spatial coverage encompassing vast spatial extents (often earth scale). However, the data are available at coarse resolutions, and the reported measurements are commensurately low-precision and contribute to averaging effects across the spatial extent. The coarse measurements coupled with the lower temporal revisit intervals preclude effective use in decision-making.

In-situ sensing environments provide high-precision measurements and are typically equipped with networking capabilities and batteries allowing data generation to occur at higher frequencies. Given that in-situ devices provide point measurements their spatial coverage is poor. The costs for deploying in-situ sensors to ensure reasonably dense spatial coverage are often too prohibitive.

Scientific models encapsulate domain science and knowledge within the community to describe, analyze, and forecast phenomena. A challenge with scientific models is their parameterization to account for regional variations coupled with ancillary measurements to seed and calibrate them.

The crux of this study is the generation of higher-precision soil moisture content (SMC) maps at high spatial resolutions (30m) daily that can facilitate decision-making. Soil moisture, which is the amount of water available in the soil, plays a critical role in several domains including agriculture and forestry. The broader science issue that we consider is how to combine remote sensing, in-situ data, and scientific models to produce high-precision (similar to in-situ) maps at higher resolutions (much higher than remote sensing), with substantial spatial coverage (similar to remote sensing), and at moderate temporal frequencies (higher than remote sensing but lower than what in-situ devices provide). This research is based on the overarching principle of knowledge-guided machine learning (KGML) as further explained in the related work.

1.1 Challenges

Generation of soil moisture maps by combining in-situ, remote sensing, and scientific models face several challenges.

(1) **Regional heterogeneity:** Soil moisture is subject to regional variations stemming from ambient conditions. As such, methods should account for such heterogeneity that drive subtle variations.
(2) **Multimodal data:** Generation of such maps relies on the inclusion of ancillary data relating to topography, weather, and soil characteristics. These multimodal data are often available at different resolutions.
(3) **Voluminous data:** Cumulatively, the data are voluminous and available at different dimensionality, and rates, and may have temporal and spatial sparsity.

1.2 Research Questions

Specific research questions that we explore include:

RQ-1: How can we combine the benefits of remote sensing and in-situ sensing schemes?

RQ-2: How can we leverage domain science to regulate model training and ensure scientific consistency in models?

RQ-3: How can we reconcile the competing pulls of computational tractability vs. refining models that are tuned to the particular spatial extent?

1.3 Approach Summary

The two sources of ground-truth SM measurements are at different precisions and spatial resolutions are: (1) NASA's SMAP satellite produces medium-precision, coarse-resolution (36km i.e., each pixel is a 1000 sq km) estimates that are unable to account for field-scale variations. (2) In-situ stations maintained by state agencies in the U.S. provide high-precision measurements albeit at a limited number (1200) of locations nationwide. Our model reconciles these limitations while estimating SMC precisely at finer-scale spatial resolutions and larger spatial extents is critical.

We combine data from in-situ, remote sensing, and scientific models alongside ancillary data to generate higher-precision soil moisture maps. To accomplish this, we rely on a novel mixture of - (1) data wrangling, (2) creation of science-infused training datasets, (3) leveraging multi-modal data from ancillary phenomena that contribute to variations in soil moisture such as meteorological, topological, and soil properties, (4) designing deep neural networks with carefully calibrated layers to account for spatiotemporal variations, (5) leveraging clustering to segment the U.S. into hydrologically similar spatial extents, and (6) training deep networks using multi-part, science-guided loss functions.

Our data-wrangling schemes rely on leveraging data from diverse sensing equipment alongside related phenomena. These data are available at different resolutions and temporal frequencies. Furthermore, the data have a mix of *static* components – such as soil properties, hydraulic conductivity, and topographical information – and *dynamically varying* components such as measurements, weather, vegetation, etc. We spatiotemporally align these datasets, partition them along pre-configured boundaries, and disperse them over our cluster of machines to ensure data locality. Next, we generate fixed-size tiles from these datasets that are suited for being input to our models.

We supplement these with data from a well-known scientific model, HydroBlocks[14], to curate training datasets. HydroBlocks outputs are generated at 30m resolutions. While HydroBlocks encapsulates domain science, a disadvantage of HydroBlocks is that there are errors in its estimations. We reconcile errors in the HydroBlocks data using ground-truth data from in-situ sensors and remote sensing in the loss function during model training.

Our deep networks consider observations (in-situ and remote) alongside ancillary information such as topography, elevation, weather conditions, etc., that, when collated together, get us closer to generating accurate, high-resolution estimates. Our deep network, DeepSoil, is a variant of the U-Net architecture [56]. It is a convolutional neural network commonly used for image-to-image translation tasks. U-Net is ideal for tasks like semantic segmentation, image colorization, and medical image analysis, as it preserves spatial details and contextual information through skip connections, making it highly effective in maintaining image quality during translation. We modify this architecture to generate high-resolution soil moisture maps, by adding spatial feature maps to further enhance image quality.

We calibrate custom loss functions to ensure effective training of our deep networks. Our loss functions are multi-part, science-guided, and account for several aspects such as the mean squared error (MSE), incorporating SMC of in-situ observations, and enforcing fuzzy bounds on the expected variation of soil moisture within a tile. Different components of the loss function are weighted dynamically to account for their contributions and significance during model training. Our loss functions allow the training to be robust and numerically stable while enhancing the model's capability for generalization. This is accomplished by regulating the model training using the van-Genuchten model [63], which characterizes the Water Retention Curve (WRC), capturing the relationship between water content and soil water properties.

Finally, our methodology involves a scaling component that relies on an ensemble of models to generate the soil moisture maps. Rather than train each of these constituent models exhaustively, we first partition the area of interest (CONUS or the contiguous United States) into a set of clusters based on their land cover, climatic classifications, and soil characteristics.

1.4 Contributions & Translation Impact

Our methodology targets the generation of high-precision soil moisture maps from observational and other ancillary data to produce high-resolution maps at temporal frequencies that can be used to inform decision-making. Our specific contributions include the following:

(1) We reconcile multimodal data from not just sensing devices, but other ancillary data.
(2) Our methodology combines the spatial coverage of remote sensing with the precision of the in-situ measurements.

(3) This study represents a novel combination of scientific models, deep networks, equations governing physics of the phenomena used to regulate how the model learns, and multimodal data working in tandem to generate soil moisture maps.
(4) We demonstrate how ancillary-related data can be combined with scientific/domain-theoretic models to inform the generation of soil moisture maps. In particular, we show how such models can be trained and how their loss functions can be calibrated and refined to guide the training of models.

Translational Impact: Similar issues arise in other scientific domains where remote sensing sources may have higher spatial coverage, but lower precision while in-situ measurements may have poor spatial coverage but higher precision. Combining such diverse observational phenomena to produce higher-precision maps at high resolutions has applications in tracking greenhouse gas emissions, ecological monitoring, health hazards due to airborne pollutants, and other phenomena where spectral measurements are available alongside in-situ observations. Finally, we note that there will always be subtle regional variations relating to ambient conditions; it is important to account for these variations to ensure model accuracy. The proposed methodology describes how to accomplish this objective.

2 Datasets Leveraged in DeepSoil

2.1 Input and Target Datasets

Our methodology is designed to produce daily soil moisture maps at 30m spatial resolutions. A key focus is on accurately estimating soil moisture values in the top 5 cm of soil. We have developed a model that accounts for a rich set of features from phenomena that influence soil dynamics. We categorize these data sources into two main groups: 'Static' and 'Dynamic' Phenomena. These datasets play a pivotal role in training our multi-step deep neural network. Static features represent attributes or characteristics that are relatively constant for a given region but have variability *across* spatial regions. In contrast, dynamic phenomena include features that change on a daily/hourly basis. In our model, DeepSoil, we integrate 11 distinct datasets, each described in this section and tabulated in Table 3 (see Appendix A.3); we also leverage these datasets for inferences and estimates that we produce over the states of Colorado and Oklahoma.

1) Static Phenomena

Soil Properties: We incorporate two distinct soil properties datasets: the Gridded National Soil Survey Geographic (gNATSGO) Database [61] and the Probabilistic Remapping of SSURGO (POLARIS) soil properties [15] dataset, both covering the CONUS.

From the gNATSGO database, we selectively utilize a subset of key properties at 30m spatial resolution. These include the available water storage estimate (in mm) and the thickness of soil components (in cm). The available water storage estimate is a vital factor in our soil moisture estimation, offering insights into the actual soil water content at a given location. Although it exhibits dynamic characteristics and is influenced by factors such as precipitation, plant water uptake, and evaporation, it serves as a baseline for the expected available water storage in a particular region. This is very useful, especially in regions where in-situ observation stations are lacking. Additionally, we incorporate derived features at coarser (90m) spatial resolution, encompassing available water content, field capacity, and soil porosity. The available water content represents the range of moisture available for plant uptake. In cropping settings, the available water content is between the soil's maximum water-holding capacity (field capacity) and the point at which plants can no longer extract water effectively (wilting point). Soil porosity provides information about soil pore volumes that impact water-holding capacity, aeration, and root penetration.

The POLARIS dataset provides additional details (at 30m resolution) about soil properties, including soil texture, pH, organic matter content, and mineral composition. We specifically leverage data on soil texture, such as silt, clay, and sand percentages. Additionally, we utilize bulk density (g/cm^3), residual soil water content (m^3/m^3), organic matter (%), and saturated hydraulic conductivity ($log_{10}(cm/hr)$) to account for soil mass per unit volume and its ability to transmit water. One such known relationship is between soil organic matter and soil porosity, where a high percentage of organic matter increases soil porosity value, allowing the soil to store more water and air for plant roots, thus increasing soil moisture content [26]. Also, soil moisture, soil matric potential, and hydraulic conductivity hold non-linear relationships with soil potential (energy exerted by plants to extract water from the soil) which varies depending on soil texture and composition (soil particle size), pore size, organic matter content, etc. These interactions are also expressed using soil retention curves (WRCs) and hydraulic conductivity functions (HCFs).

Satellite Imagery: We also include cloud-free satellite imagery obtained from the Landsat 8 Collection 2 Level-2 atmospherically corrected surface reflectance product [41]. This imagery provides a high-resolution view of the land surface at 30m across the red, blue, green, and near-infrared (NIR) spectra. These spectral bands enable us to compute the Normalized Difference Vegetation Index (NDVI), a valuable indicator for assessing vegetation health. Lower NDVI values suggest moisture-stressed vegetation, while higher values indicate denser and healthier green vegetation.

Land Cover and Climatic Conditions: We incorporate two additional datasets to enhance our model's contextual grounding. The first dataset is the National Land Cover Database (NLCD) [25] that provides a comprehensive classification of the Earth's surface into 16 distinct categories, including forests, urban areas, agriculture, and water bodies. This dataset enables precise characterization of land covers within our study area.

The second dataset we leverage is the Köppen climate classifications [33]. The Köppen classifications offer insights into long-term climatic patterns across different geographic regions. In particular, these classifications segment CONUS into 30 classes organized under five primary climatic groups: tropical, dry, temperate, continental, and polar.

Elevation Dataset: Relative slopes have an impact on pooling, water movement, and soil moisture variations within a region. We include the 3D Elevation Program's (3DEP) 1

Arc-second Digital Elevation Models (DEMs) dataset [62]. This dataset represents the bare-earth topographic surface, offering elevation values at a 30m spatial resolution.

2) Dynamic Phenomena

Meteorological Conditions: We leverage the Gridded Surface Meteorological (GridMET) dataset [5] for access to daily surface meteorological data covering CONUS. The GridMET data is at 4 km resolution and includes several meteorological variables that exhibit dynamic changes over time such as precipitation accumulation, maximum and minimum relative humidity, temperature, etc.

Leaf Area Index (LAI): To capture the dynamic aspect of vegetation and its influence on soil moisture, we incorporate the MODIS/ Terra+Aqua Leaf Area Index/FPAR (MCD15A3H.061) dataset [46] available every 4 days at 500m resolution. The leaf area index plays a pivotal role in regulating soil moisture dynamics. As vegetation grows, it can extract water from the soil through transpiration. Higher leaf area index values may indicate increased water uptake, potentially leading to reduced soil moisture levels, while lower values suggest reduced vegetation. The dataset quantifies the leaf area in a specific region; the leaf area index changes with plant growth and seasonal variations.

Satellite-Based Soil Moisture Maps: We also rely on the Soil Moisture Active Passive (SMAP) satellite data [21] to obtain a time-series of soil moisture measurements at a spatial resolution of 9 km within the top 5 cm of the soil. This dataset is categorized as a Level 3 product, presenting daily composites derived from Level 2 surface soil moisture data (36 km). The dataset offers a relatively coarse estimate of soil moisture, providing an overview of moisture trends within a region [22, 31], although with limited precision in capturing field-scale soil moisture.

In-situ Soil Moisture Measurements [RQ-1]: Our objective in this research extends beyond generating high-resolution (30m) soil moisture maps using deep neural networks (DNNs). A key goal is to ensure accuracy and precision that are close to in-situ station data. Notably, HydroBlocks exhibits errors when compared to in-situ observations. A key source of these errors seems to be limitations in accounting for the influence of precipitation events on soil moisture dynamics. In-situ soil moisture station data is the most accurate soil moisture estimate available; given that it is a point measurement the estimate is accurate within proximate spatial regions (typically 10-30m), making our task particularly challenging.

To produce 30m spatial soil moisture maps that closely align with in-situ data, we harness soil moisture measurements from **15 distinct soil moisture observation networks**, as detailed in Table 2 (see Appendix A.3). In total, there are **1216 observation stations**, each continuously measuring soil moisture levels on an hourly/daily basis. The precise locations of these stations are depicted in Fig. 1.

Physics-Based Soil Moisture Maps [RQ-2]: SMAP-Hydro Blocks (SMAP-HB) [14, 65] are high-resolution soil moisture maps generated through a physics-based approach using the Tau-Omega Radiative Transfer Model (RTM), reverse RTM models and Bayesian Merging Scheme. This methodology starts by identifying hydrologically similar

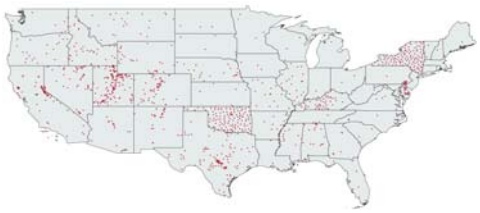


Figure 1: Geographic distribution of in-situ soil moisture monitoring stations operated by state and federal agencies within the United States.

units within a region, capturing interactions based on meteorological, topological, water flow, and hydrological properties of the soil. Next, RTM-based temperature products are computed and integrated with SMAP L3 brightness temperature observations through a Bayesian merging approach. Finally, an inverse RTM is applied to convert these temperature products into downscaled soil moisture maps. This satellite-based surface soil moisture dataset covers an entire continent, serving as our primary target dataset for training our deep neural networks. Although not representing the ground truth data, it provides accurate soil moisture (7% MAE) that closely approximates real-world conditions. The HydroBlocks product offers data at a 30m spatial resolution, with a temporal frequency of 6 hours, encompassing CONUS but only for the period spanning 2015 to 2019.

3 Methodology

Our overall approach for generating 30m daily soil moisture maps is presented in Fig. 2. Our methodology encompasses:

(1) Data wrangling and harmonization: To reconcile and harness diverse multimodal data. We distribute our raw input datasets uniformly across a cluster of machines, with each machine responsible for preprocessing and modeling a subset of the data.
(2) Design of deep neural networks: The network design includes layers suited to our task alongside a calibration of the network structure and hyperparameters using the HyperBand algorithm[36].
(3) Science-guided loss functions: We have designed a novel, multi-part loss function that accounts for nonlinear interactions between soil properties to regulate how our deep neural network learns.
(4) Clustering regions based on SMC dynamics: This is used to manage the competing pulls of refinements and computational requirements. Rather than train an all-encompassing model, we cluster spatial extents based on their spatial similarities along soil and hydrological characteristics. Models are trained for clusters of regions. During inferences, each spatial extent is parameterized based on their available multimodal data.

Our model development proceeds in two phases. First, we design a model that forecasts monthly average soil moisture maps. The monthly-average model captures nonlinear relationships between soil characteristics within a region and the monthly average precipitation patterns within a spatial

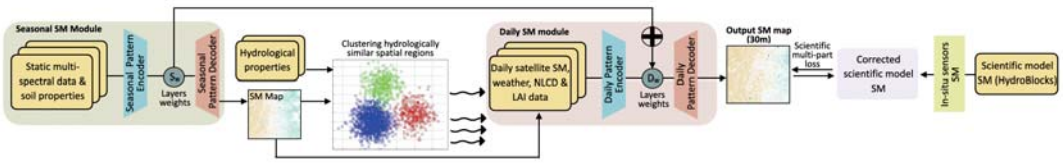


Figure 2: System Overview: Dataset preprocessing is performed while preserving data locality. k-means++ clustering, process-based models, and science-informed multipart loss functions are key components of DeepSoil.

extent. The coarse temporal granularity of the monthly-average model allows us to ensure accuracy and provides contextual grounding for the daily-moisture model.

In the second phase, we design a daily-moisture estimation model that specializes in assimilating daily variations across input features such as meteorological data (GridMET), leaf area indices, land cover, and topographical information to estimate daily soil moisture maps. We also include coarse-grained, low-precision SMAP satellite soil moisture measurements. All the input datasets fed to the daily-moisture model are available daily and at diverse spatial resolutions.

Our methodology partitions the area of interest into spatial extents, S_E. Each spatial extent, S_E, comprises 9 tiles, and each tile has 64x64 pixels, where each pixel has a 30m resolution. Our training datasets are collated in terms of spatial extents. For each spatial extent, we spatiotemporally harmonize and align multimodal data across meteorological, topographical, soil properties, etc. These data are often available at different spatial resolutions. Spatial extents are also employed to adjust soil moisture estimates generated using scientific models.

3.1 Data Preprocessing [RQ-1]

Our research involves the integration of diverse data sources with varying value ranges, spatial resolutions, and temporal frequencies. To ensure efficient model training, we harmonize all datasets to a consistent spatial resolution (30m), temporal frequency (daily), one-hot encoding (for categorical data), and normalized values (0 to 1). Missing values are marked as -1, and a mask is created to identify these locations. Datasets that are not available on a daily basis undergo the nearest temporal interpolation process to bridge temporal gaps using proximate temporal scans. For spatial consistency, we employ OpenCV's [10] inter-nearest image resizing, reducing all datasets to a uniform 64x64 pixels.

To handle large raw tiles, some of which encompass the entire CONUS, we adopt the quadkeys concept [1]. This approach recursively divides the geospatial coordinate space into non-overlapping bounding boxes, each assigned a specific identifier. Smaller quadkey lengths represent larger spatial regions. We partition all datasets with 14-character-long quadkeys. For GridMET and SMAP datasets with coarser spatial resolutions, we use 12-characters long quadkeys. This partitioning strategy aligns with memory and computational constraints during model training.

In addition to partitioning images into 64x64 spatial regions, we introduce the concept of Spatial Extent (S_E). This is a spatial region generated by merging 9 neighboring 64x64 tiles centered around an in-situ station. The resulting image encompasses a larger spatial area, measuring 192x192 pixels. This is illustrated in Fig. 11 (in Appendix A.3). When we cluster the tiles and perform the warm-start model training, we have used a similar scheme of spatial extents.

3.2 Clustering Regions [RQ-3]

Next, we identify regions that share similar soil moisture dynamics and properties. This is achieved by integrating an extensive range of environmental and climatic datasets. These datasets encompass the Köppen climate map, landcover information, elevation data, seasonal averages of soil moisture extracted from SMAP observations, and soil property data sourced from the POLARIS and gNATSGO databases.

We cluster spatial extents characterized by similar soil moisture dynamics and properties. We leverage k-Means++ clustering algorithm, known for its capacity to optimize cluster initialization. Identification of the number of clusters, k, is guided by the maximization of Silhouette scores which assess how well data points within a cluster are separated from each other compared to how close they are to data points in neighboring clusters. We classify 1216 spatial extents into 19 distinct clusters.

The number of identified clusters corresponds to the number of distinct model instances that we train. We train 19 distinct model instances, each specializing in the prediction of soil moisture dynamics within regions classified as hydrologically similar. Each cluster is comprised of 20 to 110 distinct locations. We distribute the training data across 19 machines, each corresponding to these distinct clusters, and simultaneously train the models. Each model utilizes locally stored data on its respective machine's disk.

3.3 DNN Architecture [RQ-2, RQ-3]

Our proposed architecture is a modified U-Net framework [27], a 4-block encoder-decoder DNN with skip connections between corresponding layers in both the encoder and decoder stacks. DeepSoil has a total of 12.9 million trainable parameters (see Appendix A.3 Fig. 10).

Each encoder block comprises a stack of 2D convolutional layers, batch normalization, and LeakyReLU activation functions within the inner layers. In contrast, the outer layer employs a ReLU activation function, ensuring that all negative values are set to zero. These activation functions allow our network to model and capture complex non-linear relationships between input variables such as hydraulic conductivity, porosity, etc, and output soil moisture predictions. These encoder blocks extract embeddings from the input meteorological and hydrological features. Identical decoder blocks replace the convolutional layer with Conv2DTranspose, which upsamples the latent vector (4x4x512) from the encoder back

to a 64x64 image, representing the generated soil moisture. Skip connections are introduced to recover information lost during input downsampling in the Encoder blocks. Skip connections have been known to facilitate faster training and gradient flow during backpropagation, effectively mitigating the vanishing gradient problem. Dropout layers are also strategically added to prevent overfitting during model training. In the final decoder block, we employ a concatenation operation, merging low-level features extracted from the NDVI index, landcover classification maps, and Landsat-derived bands with the upsampled image from the U-Net architecture. This enhances model performance by capturing intricate spatial relationships within the input data. This informs our model's understanding of road structures, agricultural patterns, and spatial region structures, ultimately improving its ability to produce high-quality outputs.

3.4 Science-Guided Loss Function [RQ-2]

A key contribution of this study is a novel loss function to regulate model training by accounting for the influence of key physical processes that account for the dynamics of soil water content within the top 5 cm of soil. We use multi-part loss functions that combine traditional measures of learning with domain science. Our multipart loss functions go beyond pixel-to-pixel error minimization to include consistency with well-established physical laws in hydrology, such as those derived from the van-Genuchten equations (see Equation 3). This ensures that the model's predictions are not only accurate but also aligned with well-known underlying hydrological relationships.

Our loss function accounts for various factors, including soil texture, soil dryness/wetness, and available water for uptake by the plant. The weights assigned to each component of the loss function are dynamically adjusted over training epochs to modify the impact of loss terms on accuracy across various model training scenarios. For all experiments, the predicted soil moisture values and the ground truth values are multiplied by a mask to exclude invalid or missing data points as marked in the target data.

Overall, our unified loss function is a weighted combination of the L2 error (MSE) and a science-based loss shown in Equations (1) and (2), that enforce the scientific consistency of relationships across soil characteristics, soil matric potential, and soil moisture dynamics during model training. Here, γ represents a hyperparameter whose value is reduced dynamically while training to reduce the impact of the basic loss component (MSE) and increase the weight of the scientific loss term. We dynamically adjust γ based on the training epoch and the total number of epochs with the objective of enhancing model performance. Our methodology divides the training process into steps, gradually decreasing γ to fine-tune the learning rate over time. During initial training epochs, the model prioritizes backpropagating errors from pixel-wise squared errors of SMC from the HydroBlocks dataset. Towards the end of the training, the scientific loss term attains the highest weight as γ values decrease to 0 to consider the relations among the predicted SMC and other soil properties. This approach is an effective optimization technique for achieving convergence; crucially, as our performance benchmarks demonstrate, this allows DeepSoil to achieve better generalization on unseen locations.

$$\mathrm{L_{Total}} = \gamma \cdot \mathrm{L_{MSE}} + (1-\gamma) \cdot \mathrm{L_{sci}} \tag{1}$$

3.4.1 ***Mean Squared Errors.*** We define a custom MSE loss between the predicted and target soil moisture maps. This criterion calculates the pixel-wise average of the squared difference between the corresponding valid elements of the predicted and target soil moisture maps. The squaring operation, allows the MSE to be more sensitive to outliers or large errors. The MSE plays a significant role during the initial training epochs when errors are typically higher.

$$L_{\mathrm{MSE}} = \frac{1}{B}\sum_{k=1}^{B}\frac{1}{N}\sum_{i=1}^{N}\left(w_i \cdot \left|x_{k,i}\cdot mask_{k,i} - y_{k,i}\cdot mask_{k,i}\right|^2\right) \tag{2}$$

Here, L_{MSE} represents the average MSE over the training dataset. B represents the number of images in the batch. N is the total number of pixels in each image. $x_{k,i}$ is the value of the pixel at index i in the predicted image of the k-th image in the batch. $y_{k,i}$ is the value of the pixel at index i in the target image of the k-th image in the batch. w_i is the weight associated with the pixel at index I. For the 16 neighboring pixels around the given index where in-situ is located, w_i has a higher value, giving them more importance than remaining pixels, and $mask_{k,i}$ provides valid pixel locations.

3.4.2 ***Soil Water Retention Curves (WRC).*** As reported by the authors of HydroBlocks, the SMC dataset exhibits notable errors that can range from approximately 7% to 15%, when compared to in-situ station measurements of soil moisture. Given our reliance on HydroBlocks as a scientifically consistent target dataset, addressing the inherent disparities between these target values and ground-truth measurements is necessary.

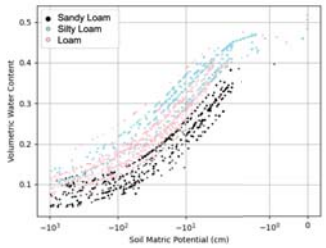


Figure 3: Using van-Genuchten to capture nonlinear relationships between water content and soil matric potential (in logarithmic scale) for different soil types using 1200+ in-situ sensors observation over a year.

The soil Water Retention Curve (WRC) represents a well-known relationship between soil characteristics - soil matric potential (ψ) and water content (θ) [48]. Soil matric potential, also known as capillary pressure, denotes the pressure or energy required to extract water from the soil against gravity, and the force with which water is withheld within

soil pores and particles. (ψ) serves as a direct measure of the available water for plants to uptake from soil. As soil moisture decreases, capillary forces increase, causing water to be retained more tightly within the soil pores. A decrease in soil moisture thus results in a more negative matric potential (ψ), indicating a greater energy requirement for water extraction from the soil and vice versa. The relationship is accurately captured in a widely-used numerical equation called the van-Genuchten equation [63], which is based on observed field/lab values of soil properties. Our model training relies on this scientific relation between soil properties, soil matric potential, and SMC to adjust the model's trainable parameters. Key parameters characterizing this relationship include soil type, residual water content, saturated water content, the α parameter influencing curve shape (shown in Fig. 3), and the n parameter representing soil pore-size distribution. This relationship can be mathematically formulated -

$$\psi = \frac{1}{\alpha} \cdot \left[\left(\frac{\theta_s - \theta_r}{\theta - \theta_r} \right)^{n/n-1} - 1 \right]^{1/n} \tag{3}$$

here, ψ is soil matric potential, α is inverse of air-entry value, θ_s is saturated water content, θ_r is residual water content, θ is volumetric water content, n, m are some constants defining curve. Our scientific loss term then becomes -

$$L_{sci} = \psi_{predicted} - \psi_{target} \tag{4}$$

We capture this non-linear relationship in our science-guided loss function to calculate ψ using Equation (3) by leveraging some of the variables from POLARIS datasets available at 30m resolution and ground-truth SMC using in-situ sensors. The soil matric potential serves as a regularization term for our DeepSoil model by calculating how far the desired ψ is from the ψ calculated using model-generated SMC. In regions where SMC is unavailable through in-situ sensors, our target dataset, HydroBlocks SMC is used to calculate ψ using the van-Genuchten equation. By integrating both HydroBlocks data and in-situ measurements, we leverage the strengths of both these products. This approach helps us mitigate inaccuracies inherent in HydroBlocks by weighting them appropriately using convolutional layers of our deep neural network and refining the model using more accurate in-situ data. This enhances the overall accuracy of our predictions.

DeepSoil learns the relation between soil matric potential and soil moisture content by learning patterns over diverse soil conditions at different locations.

By integrating station-based error adjustments through loss function into our approach, we achieve a higher level of precision and reliability in the target soil moisture maps that we used to train our model against. This scientific approach enables us to capture localized variations and address discrepancies between ground truth measurements and physics-modeled values (HydroBlocks), ultimately increasing the accuracy of our predictions.

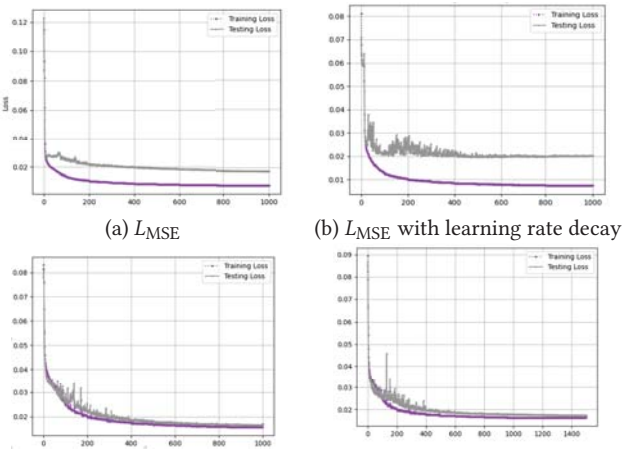
4 Empirical Benchmarks & Evaluation

To assess the performance of our models, we performed empirical benchmarks over a 19-node cluster. Each node in our distributed cluster is equipped with an Intel Xeon E5-2620v3 processor, 64 GB of RAM, and a Quadro P2200 GPU with 5GB of memory and 1280 cores. Our approach involves the uniform distribution of datasets, organized by spatial clusters identified through quadkey assignments. Data originating from different sources but sharing the same quadkey are placed on the same machine to ensure data locality and minimize network transfers during model training.

We use the PyTorch framework and the Adam Optimizer to train all our models. We train our models for 1000-1500 epochs, leveraging a dataset comprising 13,000 randomly selected training samples over 1219 spatial extents. The input dataset used for training consists of soil moisture maps from the year 2019, focusing on areas close to in-situ stations, as visualized in Fig. 1. We evaluate the performance of our models using data from unfamiliar geographical locations.

4.1 Loss function components [RQ-2]

We begin by assessing the significance of incorporating a science-guided loss function into our model training process. Our combined loss function comprises two primary components: the traditional MSE loss (L_{MSE}) and the science-guided loss component (L_{sci}). Each of the component's weights is dynamically adjusted while training. We start by evaluating the contribution of each of these loss terms.



(a) L_{MSE} (b) L_{MSE} with learning rate decay

(c) L_{MSE} and L_{sci} (HydroBlocks) (d) L_{MSE} and L_{sci} (In-situ sensors)

Figure 4: MAE for models with different combinations of MSE and science-guided loss components. The testing errors are measured using MAE between predicted and target soil moisture maps at unseen locations.

In Fig. 4, we depict training and testing loss profiles under four different scenarios - (1) Training only with the MSE loss function (also BASELINE model i.e. without using science-guided learning), (2) Training with MSE loss while gradually reducing learning rate by a factor of 0.8 every 100 epochs using step scheduler, (3) Training with MSE loss and scientific loss using ψ prediction based on HydroBlocks' model outputs, and (4) Training with MSE loss and scientific loss using ψ prediction based on in-situ sensor observations.

The optimal model training results are achieved when the loss function combines (L_{MSE} and L_{sci}) using in-situ observations while reducing γ parameter using step decay.

This combination yields an outstanding overall testing loss of 0.017 (1.7%) and a testing PSNR of 34.48 dB. While a model trained only with the MSE loss exhibits relatively low testing MAE (1.9%) and PSNR accuracy of (33.92 dB), the images produced are slightly blurry and lack finer details.

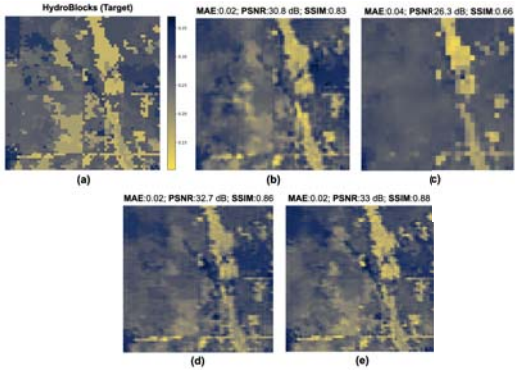


Figure 5: Sample images generated by models trained with different configurations: (a) Target SMC map; (b) L_{MSE}, (c) L_{MSE} with learning rate decay, (d) L_{MSE} and L_{sci} using HydroBlocks-based soil matric potential. and (d) L_{MSE} and L_{sci} using in-situ sensor data.

An important observation is that while the loss value of a model trained with the MSE loss is considerably low, the image quality is quite low when assessing using measures such as PSNR and SSIM (structural similarity index measure; higher the better), which quantifies the fidelity of images. This is evident in Fig. 5, where 5(b) and (c) corresponds to an image generated by a model trained only with MSE in the loss function. The image appears blurry and lacking in finer, sharper spatial details such as road structures and landcover boundaries. Overall, the training and testing loss show that **much better generalization achieved using the scientific loss component, and the model does not overfit**. This shows that DeepSoil is able to cope with limited training samples (due to the low number of in-situ sensors) while achieving a solid performance over entirely unseen locations. As can be seen, the SMC maps are much sharper than pure MSE-based loss functions.

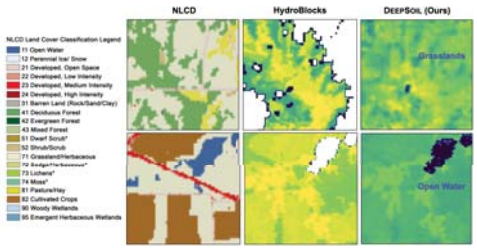


Figure 6: Emergent behavior in DeepSoil: Regeneration of missing pixels based on landcover types.

We also observed an emerging behavior within our DNN models, as depicted in Fig. 6. These models were trained with a science-guided loss function, leveraging inputs from landcover maps, NDVI indices extracted from Landsat imagery. We observed that approximately 1.1% of pixels in the HydroBlocks ground-truth dataset were missing, primarily associated with water bodies. While these missing pixels were intentionally excluded during the loss calculation process to decrease training errors, our DNN models exhibited a remarkable, emergent ability to regenerate these water bodies effectively. During the testing phase, the model indicated the ability to regenerate the missing pixels within the image and accurately distinguish soil moisture values based on the land cover type.

Table 1: Comparing SMAP, HydroBlocks, and variations of DeepSoil w.r.t to ground station data.

Model	MAE	Std. Deviations
SMAP	0.26	+/- 0.06
HydroBlocks	0.077	+/- 0.067
BASELINE (L_{MSE})	0.061	+/- 0.043
L_{MSE} with learning rate decay	0.16	+/- 0.062
L_{MSE} and L_{sci} (HydroBlocks)	0.0638	+/- 0.046
L_{MSE} and L_{sci} (In-situ sensors)	0.046	+/- 0.0331

4.2 Clustering [RQ-3]

We did an accuracy analysis of our model, comparing its performance with and without the incorporation of our k-means++ clustering scheme. Specifically, we evaluate two different model training approaches: one involving a single global model trained on all locations in the training set but with fewer time steps and the other utilizing models trained on clustered spatial extents. The clustering process is based on a combination of soil, land, and hydrological properties, resulting in 19 distinct clusters.

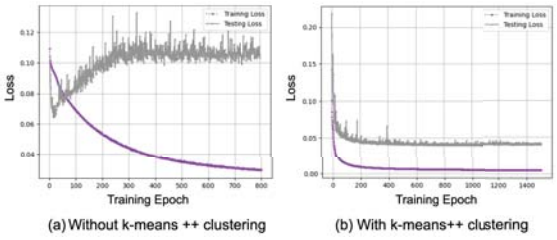


Figure 7: Training and testing MAE in models with and without clustering.

As illustrated in Fig. 7, the model trained without clustering indicates significantly higher errors. It converges slowly during training but also demonstrates poor performance on unseen data, with MAE errors increasing to 10-12%. Conversely, the models trained using multiple clusters exhibit faster convergence and consistently achieve lower test loss, with an average MAE of 5%. These models, when trained on clustered spatial extents, become more specialized as they focus on a smaller number of distinct spatial regions. This experiment highlights the advantages of our clustering-based approach in enhancing model performance and achieving better generalization overall.

4.3 30m Resolution SMC Maps [RQ-1, RQ-2]

Finally, we assess the overall effectiveness of our approach by contrasting soil moisture prediction errors with different variations of loss functions in DeepSoil, HydroBlocks, and the satellite-based SMAP 36 km data product. These assessments are performed by contrasting estimates with in-situ, ground truth measurements. This evaluation is performed over 100+ testing locations spanning March-November of 2019, which were *unseen* during the model training phase. We report MAE and standard deviations in Table 1 for locations where ground-truth stations are available.

Notably, DeepSoil has an average of **a 4.6% soil moisture error when compared to in-situ station observations** over 19 clusters. We achieve a (substantial) **82.3% reduction in SMC percentage error over SMAP** 9 km and **40.25% reduction in error over HydroBlocks**. As we experiment with different loss terms, the errors reduce significantly while incorporating scientific knowledge using van-Genuchten equation. Our model outperforms the BASELINE model that is based on regular MSE-based loss with a 24.59% reduction in errors. DeepSoil reduces the weight of MSE errors and increases the overall impact of the scientific equation over training epochs in the loss function. For fair comparisons, we also performed experiment by training model using pure MSE-based learning but with reducing learning rates over epochs (using step learning rate decay). From Table 1, this variation leads to the worst performance with an increase in errors by 162.2% and 247.8% compared to BASELINE and our science-guided model respectively.

Lastly, we compare the model performance and show how incorporating in-situ sensors in scientific-loss term helps in reducing errors in HydroBlocks. Overall, we achieve 27.8% decrease in errors when in-situ sensor observations are leveraged to drive the model training. These results demonstrate the accuracy and precision of our model in predicting soil moisture. Our approach has a demonstrably high correlation with in-situ stations.

We also conducted an extensive evaluation encompassing the entire state of California and Oklahoma (depicted in Fig. 8 and 9). We compared our predictions with satellite-based SMC from SMAP at its original sensing resolution of 36 km. We generated the SMC map for the months of April (for CA, 2024) and July (for OK, 2023) because these months influence planting and irrigation plans. April marks the start of the growing season in temperate regions such as California. Similarly, mid-summer (July), marks a critical period for crops such as corn as they are in pivotal stage of development and are highly sensitive to available water content. Lack of water intake can lead to significant loss of crop yield. This evaluation involved utilizing datasets of over 200,000+ 64x64 pixel tiles, amounting to approximately 83 GB data. To contrast the scales involved: our model generated images for the area of 163,000 mi^2 for California with 35097x38100 pixels (versus SMAP's 23x26 pixels) and 24849x11930 pixels for Oklahoma (versus SMAP's 9x21 pixels).

SMAP data products include several flags regarding the poor quality of the data, especially for surfaces with permanent ice and snow, urban areas, wetlands, and proximity of large water bodies or coastlines [32, 50, 51]. Our experiment includes coastal regions and urban areas where SMAP is known to have high errors [16]. Our model performs well with the limited number of in-situ stations in this target region and demonstrates an **accuracy of higher than 99.7% across the areas unseen by the model**. In particular, this result highlights the effectiveness of our approach in the regions where SMAP underperforms significantly.

On the other hand, in Oklahoma, in-situ stations are near-uniformly dispersed, providing rich data coverage for soil moisture mapping. DeepSoil consistently demonstrates its capacity to generate high-quality soil moisture predictions that closely align with SMAP observations that are very coarse in spatial resolution. Crucially, our approach captures subtle spatial variations in soil moisture, particularly in distinguishing between riverbeds and surrounding land.

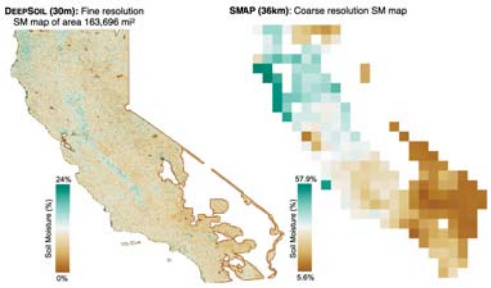


Figure 8: Visualizations contrasting soil moisture maps for April 6th, 2024 over California generated by DeepSoil (30m) vs SMAP (36km sensed resolution).

We compared our predictions with the soil volumetric water content (VWC) data produced by Oklahoma State University at 800m spatial resolution[49], as illustrated in Fig. 9(c). This map's generation process involves utilizing in-situ SMC measurements from the Oklahoma Mesonet for a regression kriging algorithm. This algorithm integrates soil texture information, specifically the percentage of sand, along with the precipitation index. The map depicts elevated levels of VWC in the western region of the state, with several localized hotspots (high VWC). However, this diverges from both SMAP satellite data and our prediction. These results, which includes the years 2023 and 2024 show that the model performs consistently well across years, demonstrating robustness to climate variations and its ability to capture non-linear interactions from daily meteorological data and satellite-based coarse resolution soil moisture content.

5 Related Work

Estimation of soil moisture maps at high precision is critical. Research efforts can be broadly categorized as being based on machine learning techniques, conventional physics-based models, and hybrid methods that seek to integrate both.

Numerical models: Physics-based approaches use fundamental equations to predict soil moisture dynamics. For example, ParFlow, Hydrus-1D/2D, pedotransfer functions [14, 17, 35, 42, 44, 53, 54, 60] are numerical hydrology models that employ Richards' equations to simulate 2D/3D subsurface flows. Richards' equation [55] is a partial differential equation that estimates water flow through unsaturated

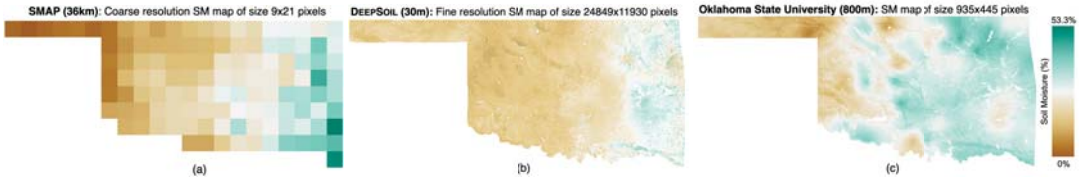


Figure 9: Visualizations contrasting soil moisture maps over Oklahoma for 17th July 2023 generated by - (a) SMAP at 36 km resolution, (b) DeepSoil at 30m spatial resolution, and (c) A kriging-based data product from OSU.

soils as a function of space and time. The equation considers various factors, including soil properties, hydraulic conductivity, and initial moisture content. These 2D/3D models account for surface and subsurface water flow, as well as land surface processes (CLM) such as evapotranspiration and snow. While it is very effective for simulations, its resolution is limited by the resolution (spatial) of input datasets. Furthermore, parameterizing the governing equations can be a challenge alongside the substantial computational requirements. These models also have known deficiencies in estimating soil moisture values over water-stressed regions.

Machine Learning Approaches: Most recent efforts have focused on utilizing time-series deep neural networks [28, 30, 37], such as the dynamic NN in Adeyemi et al. [6], offering a one-day-ahead irrigation-oriented soil moisture predictions using LSTM networks while leveraging dynamic and static climate features. Peng et al.'s [52] introduces the External Parameter Orthogonalization-Partial Least Squares model, improving soil moisture predictions by mitigating the effect of parameters such as soil salinity. In a comparative study, Senyurek et al. [58] assess the performance of Artificial Neural Networks (ANN), Random Forests (RF), and Support Vector Machines (SVM) for soil moisture prediction. Other efforts involve downsampling satellite-based measurements from Landsat and SMAP-based soil moisture using RFs[67, 68]. Ge et al. [23] propose a method combining UAV data with Extreme Learning Machines for automated parameter selection. While these techniques demonstrate solid performance, often achieving soil moisture errors within the range of 5-10%, their applicability at larger scales poses a significant challenge. Point-based prediction and evaluations at smaller spatial scales restrict the generalizability and scalability of these models.

Physics-informed machine learning models: To address these limitations efforts have leveraged knowledge-guided machine learning (KGML) in other domains. For example, Karpatne et al. [18, 29] propose an approach that integrates physics-based principles and domain knowledge into machine learning models while predicting lake water temperature with minimal in-situ measurements. This involves modifications to loss functions, network architectures, and ingesting numerically simulated input datasets. Such knowledge-guided efforts have shown promise in various applications, including weather and climate modeling [70] and carbon cycling [38]. Some physics-informed machine learning models [7, 8, 19, 24] formulate water flow by utilizing numerical models outputs to solve Richard's equation in the loss function. In super resolving tasks, machine learning models like GANs, Pix2Pix networks, and ESRGAN are often coupled with PDE-constrained loss functions by penalizing errors based on energy spectrum differences [40, 64]. We adopt a KGML approach that combines the U-Net architecture with a science-informed loss function. This combination allows us to predict high-precision soil moisture maps at 30m resolution at scale.

6 Conclusions & Future Work

In this study, we described our methodology to combine remote sensing, in-situ sensing, and scientific models to produce high-precision soil moisture maps. Our results demonstrate the effectiveness of our model for accurate and precise soil moisture mapping across diverse geographic regions.

RQ1: Our approach dynamically assigns higher weights to loss values in the surroundings of in-situ stations. This tailored approach ensures that the model training is guided by ground truth data. Despite the sparsity of in-situ data, our model successfully integrates essential information, enabling the generation of large-scale soil moisture maps. Our empirical evaluation demonstrates that our approach achieves soil moisture predictions with a very low 4% MAE when compared to in-situ station measurements.

RQ2: Our weighted multipart loss function assimilates scientific insights and conventional loss functions. This also enhances the model's convergence by tailoring soil moisture estimates to account for the interrelated factors influencing them using well-established scientific equations. Our loss function also contributes to a perceptible improvement in the quality of the generated images.

RQ3: Our approach of clustering spatial extents based on soil properties, monthly soil moisture averages, land-cover types, and other land-surface properties proved to be effective in reducing training errors when compared to a single global model trained for CONUS. Additionally, the clustered approach significantly reduces the required memory footprint and training times by converging faster. This allows our models to scale while tuning individual models for specific spatial regions.

In future work, we plan to extend our approach to predict soil moisture at deeper soil depths, specifically sub-surface soil in the range of 5-30 cm using Richard's Equation to account for hydraulic conductivity in unsaturated soils.

Acknowledgments

This research was supported by the National Science Foundation (1931363, 2312319), the National Institute of Food Agriculture (COL014021223) and an NSF/NIFA Artificial Intelligence Institutes AI-CLIMATE Award [2023-03616].

References

[1] 2018. QuadTiles. https://wiki.openstreetmap.org/wiki/QuadTiles

[2] 2019. *Montana Mesonet, Montana Climate Office 2023. Data accessed from https://climate.umt.edu/mesonet/ on January 2023.*

[3] 2019. *National Soil Moisture Network.* http://nationalsoilmoisture.com/About.html

[4] 2019. *Texas Water Development Board.* https://www.texmesonet.org/

[5] John T Abatzoglou. 2013. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology* 33, 1 (2013), 121–131.

[6] Olutobi Adeyemi, Ivan Grove, Sven Peets, Yuvraj Domun, and Tomas Norton. 2018. Dynamic neural network modelling of soil moisture content for predictive irrigation scheduling. *Sensors* 18, 10 (2018), 3408.

[7] Toshiyuki Bandai and Teamrat A Ghezzehei. 2021. Physics-informed neural networks with monotonicity constraints for Richardson-Richards equation: Estimation of constitutive relationships and soil water flux density from volumetric water content measurements. *Water Resources Research* 57, 2 (2021), e2020WR027642.

[8] Toshiyuki Bandai and Teamrat A Ghezzehei. 2022. Forward and inverse modeling of water flow in unsaturated soils with discontinuous hydraulic conductivities using physics-informed neural networks with domain decomposition. *Hydrology and Earth System Sciences* 26, 16 (2022), 4469–4495.

[9] Jesse E Bell, Michael A Palecki, C Bruce Baker, William G Collins, Jay H Lawrimore, Ronald D Leeper, Mark E Hall, John Kochendorfer, Tilden P Meyers, Tim Wilson, et al. 2013. US Climate Reference Network soil moisture and temperature observations. *Journal of Hydrometeorology* 14, 3 (2013), 977–988.

[10] Gary Bradski, Adrian Kaehler, et al. 2000. OpenCV. *Dr. Dobb's journal of software tools* 3, 2 (2000).

[11] Fred V Brock, Kenneth C Crawford, Ronald L Elliott, Gerrit W Cuperus, Steven J Stadler, Howard L Johnson, and Michael D Eilts. 1995. The Oklahoma Mesonet: a technical overview. *Journal of Atmospheric and Oceanic Technology* 12, 1 (1995), 5–19.

[12] Jerald A Brotzge, J Wang, CD Thorncroft, E Joseph, N Bain, N Bassill, N Farruggio, JM Freedman, K Hemker Jr, D Johnston, et al. 2020. A technical overview of the New York State Mesonet standard network. *Journal of Atmospheric and Oceanic Technology* 37, 10 (2020), 1827–1845.

[13] Todd G Caldwell, Tara Bongiovanni, Michael H Cosh, Thomas J Jackson, Andreas Colliander, Charles J Abolt, Richard Casteel, Toti Larson, Bridget R Scanlon, and Michael H Young. 2019. The Texas soil observation network: A comprehensive soil moisture dataset for remote sensing and land surface model validation. *Vadose Zone Journal* 18, 1 (2019), 1–20.

[14] Nathaniel W Chaney, Peter Metcalfe, and Eric F Wood. 2016. HydroBlocks: A field-scale resolving land surface model for application over continental extents. *Hydrological Processes* 30, 20 (2016), 3543–3559.

[15] Nathaniel W Chaney, Eric F Wood, Alexander B McBratney, Jonathan W Hempel, Travis W Nauman, Colby W Brungard, and Nathan P Odgers. 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma* 274 (2016), 54–67.

[16] Andreas Colliander, Thomas J Jackson, Rajat Bindlish, S Chan, N Das, SB Kim, MH Cosh, RS Dunbar, L Dang, L Pashaian, et al. 2017. Validation of SMAP surface soil moisture products with core validation sites. *Remote sensing of environment* 191 (2017), 215–231.

[17] Monidipa Das and Soumya K Ghosh. 2019. FB-STEP: a fuzzy Bayesian network based data-driven framework for spatio-temporal prediction of climatological time series data. *Expert Systems with Applications* 117 (2019), 211–227.

[18] Arka Daw, Anuj Karpatne, William D Watkins, Jordan S Read, and Vipin Kumar. 2022. Physics-guided neural networks (pgnn): An application in lake temperature modeling. In *Knowledge Guided Machine Learning*. Chapman and Hall/CRC, 353–372.

[19] Ivan Depina, Saket Jain, Sigurdur Mar Valsson, and Hrvoje Gotovac. 2022. Application of physics-informed neural networks to inverse problems in unsaturated groundwater flow. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 16, 1 (2022), 21–36.

[20] Howard J Diamond, Thomas R Karl, Michael A Palecki, C Bruce Baker, Jesse E Bell, Ronald D Leeper, David R Easterling, Jay H Lawrimore, Tilden P Meyers, Michael R Helfert, et al. 2013. US Climate Reference Network after one decade of operations: Status and assessment. *Bulletin of the American Meteorological Society* 94, 4 (2013), 485–498.

[21] Dara Entekhabi, Eni G Njoku, Peggy E O'neill, Kent H Kellogg, Wade T Crow, Wendy N Edelstein, Jared K Entin, Shawn D Goodman, Thomas J Jackson, Joel Johnson, et al. 2010. The soil moisture active passive (SMAP) mission. *Proc. IEEE* 98, 5 (2010), 704–716.

[22] Kuai Fang, Ming Pan, and Chaopeng Shen. 2018. The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Transactions on Geoscience and Remote Sensing* 57, 4 (2018), 2221–2233.

[23] Xiangyu Ge, Jingzhe Wang, Jianli Ding, Xiaoyi Cao, Zipeng Zhang, Jie Liu, and Xiaohang Li. 2019. Combining UAV-based hyperspectral imagery and machine learning algorithms for soil moisture content monitoring. *PeerJ* 7 (2019), e6926.

[24] Mohammad Reza Hajizadeh Javaran, Mohammad Mahdi Rajabi, Nima Kamali, Marwan Fahs, and Benjamin Belfort. 2023. Encoder–Decoder Convolutional Neural Networks for Flow Modeling in Unsaturated Porous Media: Forward and Inverse Approaches. *Water* 15, 16 (2023), 2890.

[25] Collin H Homer, Joyce A Fry, Christopher A Barnes, et al. 2012. The national land cover database. *US geological survey fact sheet* 3020, 4 (2012), 1–4.

[26] Ashok Kumar Indoria, Kishori Lal Sharma, and Kotha Sammi Reddy. 2020. Hydraulic properties of soil under warming climate. In *Climate change and soil interactions*. Elsevier, 473–508.

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[28] Noureddine Jarray, Ali Ben Abbes, Manel Rhif, Hanen Dhaou, Mohamed Ouessar, and Imed Riadh Farah. 2022. SMETool: A web-based tool for soil moisture estimation based on Eo-Learn framework and Machine Learning methods. *Environmental Modelling & Software* 157 (2022), 105505.

[29] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering (TKDE)* 29, 10 (2017), 2318–2331.

[30] Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. 2018. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31, 8 (2018), 1544–1554.

[31] Hyunglok Kim, Sangchul Lee, Michael H Cosh, Venkataraman Lakshmi, Yonghwan Kwon, and Gregory W McCarty. 2020. Assessment and combination of SMAP and Sentinel-1A/B-derived soil moisture estimates with land surface model outputs in the Mid-Atlantic Coastal Plain, USA. *IEEE Transactions on Geoscience and Remote Sensing* 59, 2 (2020), 991–1011.

[32] Seung-bum Kim, Jakob van Zyl, Scott Dunbar, Eni Njoku, Joel Johnson, Mahta Moghaddam, Jiancheng Shi, and Leung Tsang. 2014. Algorithm Theoretical Basis Document Smap L2 & L3 Radar Soil Moisture (Active) Data Products. (2014).

[33] Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. 2006. World map of the Köppen-Geiger climate classification updated. (2006).

[34] David R Legates, Daniel J Leathers, Tracy L DeLiberty, Geoff E Quelch, Kevin Brinson, Jason Butke, Rezaul Mahmood, and Stuart A Foster. 2005. Deos: The delaware environmental observing system. In *21st Int. Conf. Interactive Information Processing Systems*.

[35] Pei Leng, Zhao-Liang Li, Qian-Yu Liao, Yun-Jing Geng, Qiu-Yu Yan, Xia Zhang, and Guo-Fei Shang. 2022. Enhanced Surface Soil Moisture Retrieval at High Spatial Resolution From the Integration of Satellite Observations and Soil Pedotransfer Functions. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–11.

[36] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research* 18, 185 (2018), 1–52.

[37] Qingliang Li, Yuheng Zhu, Wei Shangguan, Xuezhi Wang, Lu Li, and Fanhua Yu. 2022. An attention-aware LSTM model for soil moisture and soil temperature prediction. *Geoderma* 409 (2022), 115651.

[38] Licheng Liu, Wang Zhou, Kaiyu Guan, Bin Peng, Shaoming Xu, Jinyun Tang, Qing Zhu, Jessica Till, Xiaowei Jia, Chongya Jiang, et al. 2024. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nature communications* 15, 1 (2024), 357.

[39] Rezaul Mahmood, Megan Schargorodski, Stuart Foster, and Andrew Quilligan. 2019. A technical overview of the Kentucky Mesonet. *Journal of Atmospheric and Oceanic Technology* 36, 9 (2019), 1753–1771.

[40] Ashray Manepalli, Adrian Albert, Alan Rhoades, Daniel Feldman, and Andrew D Jones. 2019. Emulating numeric hydroclimate models with physics-informed cGANs. In *AGU fall meeting*.

[41] Jeffrey G Masek, Eric F Vermote, Nazmi E Saleous, Robert Wolfe, Forrest G Hall, Karl Fred Huemmrich, Feng Gao, Jonathan Kutler, and Teng-Kui Lim. 2006. A Landsat surface reflectance dataset for North America, 1990-2000. *IEEE Geoscience and Remote sensing letters* 3, 1 (2006), 68–72.

[42] Reed M Maxwell and Norman L Miller. 2005. Development of a coupled land surface and groundwater model. *Journal of Hydrometeorology* 6, 3 (2005), 233–247.

[43] Renee A McPherson, Christopher A Fiebrich, Kenneth C Crawford, James R Kilby, David L Grimsley, Janet E Martinez, Jeffrey B Basara,

Bradley G Illston, Dale A Morris, Kevin A Kloesel, et al. 2007. Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *Journal of Atmospheric and Oceanic Technology* 24, 3 (2007), 301–321.

[44] Nuno Cirne Mira, João Catalão, Giovanni Nico, and Pedro Mateus. 2021. Soil moisture estimation using atmospherically corrected C-band InSAR data. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–9.

[45] M Moghaddam, A Silva, D Clewley, R Akbar, SA Hussaini, J Whitcomb, R Devarakonda, R Shrestha, RB Cook, G Prakash, et al. 2016. Soil moisture profiles and temperature data from SoilSCAPE sites, USA. *ORNL DAAC* (2016).

[46] Ranga Myneni, Yuri Knyazikhin, and Taejin Park. 2015. MOD15A2H MODIS/Terra leaf area Index/FPAR 8-Day L4 global 500m SIN grid V006. *NASA EOSDIS Land Processes DAAC* (2015).

[47] Illinois Climate Network. 2017. Water and atmospheric resources monitoring program. *Illinois State Water Survey, Champaign, Illinois* (2017).

[48] Viliam Novák, Hana Hlaváčiková, Viliam Novák, and Hana Hlaváčiková. 2019. Soil-water retention curve. *Applied soil hydrology* (2019), 77–96.

[49] Tyson E Ochsner, Evan Linde, Matthew Haffner, and Jingnuo Dong. 2019. Mesoscale soil moisture patterns revealed using a sparse in situ network and regression kriging. *Water Resources Research* 55, 6 (2019), 4785–4800.

[50] Peggy O'neill, Steven Chan, Eni Njoku, Tom Jackson, and Rajat Bindlish. 2014. Soil moisture active passive (SMAP) algorithm theoretical basis document Level 2 & 3 soil moisture (passive) data products. *Jet Propulsion Laboratory. California Institute of Technology* (2014).

[51] Soil Moisture Active Passive. 2014. L2 & L3 Radar/Radiometer Soil Moisture (Active/Passive) Data Products. (2014).

[52] Xiang Peng, Dan Hu, Wenzhi Zeng, Jingwei Wu, and Jiesheng Huang. 2016. Estimating soil moisture from hyperspectra in saline soil based on EPO-PLS regression. *Transactions of the Chinese Society of Agricultural Engineering* 32, 11 (2016), 167–173.

[53] Giuseppe Provenzano. 2007. Using HYDRUS-2D simulation model to evaluate wetted soil volume in subsurface drip irrigation systems. *Journal of Irrigation and Drainage Engineering* 133, 4 (2007), 342–349.

[54] Luca Pulvirenti, Giuseppe Squicciarino, Luca Cenci, Giorgio Boni, Nazzareno Pierdicca, Marco Chini, Cosimo Versace, and Paolo Campanella. 2018. A surface soil moisture mapping service at national (Italian) scale based on Sentinel-1 data. *Environmental Modelling & Software* 102 (2018), 13–28.

[55] Lorenzo Adolph Richards. 1931. Capillary conduction of liquids through porous mediums. *physics* 1, 5 (1931), 318–333.

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[57] Garry L Schaefer and Ron F Paetzold. 2001. SNOTEL (SNOwpack TELemetry) and SCAN (soil climate analysis network). *Automated Weather Stations for Applications in Agriculture and Water Resources Management: Current Use and Future Perspectives* 1074 (2001), 187–194.

[58] Volkan Senyurek, Fangni Lei, Dylan Boyd, Mehmet Kurum, Ali Cafer Gurbuz, and Robert Moorhead. 2020. Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS. *Remote Sensing* 12, 7 (2020), 1168.

[59] Martha Shulski, Stonie Cooper, Glen Roebke, and Al Dutcher. 2018. The Nebraska Mesonet: Technical overview of an automated state weather network. *Journal of Atmospheric and Oceanic Technology* 35, 11 (2018), 2189–2200.

[60] J Simunek, M Sejna, M Th Van Genuchten, J Šimnek, M Šejna, D Jacques, and M Sakai. 1998. HYDRUS-1D. *Simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media, version* 2 (1998).

[61] Soil Survey Staff. 2020. Gridded National Soil Survey Geographic (gNATSGO) Database for the Conterminous United States. *United States Department of Agriculture, Natural Resources Conservation Service* (2020).

[62] US Geologic Survey. 2017. 1 Arc-second Digital Elevation Models (DEMs)–USGS National Map 3DEP Downloadable Data Collection.

[63] M Th Van Genuchten. 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil science society of America journal* 44, 5 (1980), 892–898.

[64] Thomas Vandal, Evan Kodra, Jennifer Dy, Sangram Ganguly, Ramakrishna Nemani, and Auroop R Ganguly. 2018. Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2377–2386.

[65] Noemi Vergopolan, Nathaniel W Chaney, Ming Pan, Justin Sheffield, Hylke E Beck, Craig R Ferguson, Laura Torres-Rojas, Sara Sadri, and Eric F Wood. 2021. SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. *Scientific data* 8, 1 (2021), 264.

[66] Robert J Zamora, F Martin Ralph, Edward Clark, and Timothy Schneider. 2011. The NOAA Hydrometeorology Testbed soil moisture observing networks: Design, instrumentation, and preliminary results. *Journal of Atmospheric and Oceanic Technology* 28, 9 (2011), 1129–1140.

[67] Yufang Zhang, Shunlin Liang, Zhiliang Zhu, Han Ma, and Tao He. 2022. Soil moisture content retrieval from Landsat 8 data using ensemble learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 185 (2022), 32–47.

[68] Wei Zhao, Nilda Sánchez, Hui Lu, and Ainong Li. 2018. A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression. *Journal of hydrology* 563 (2018), 1009–1024.

[69] Jingyao Zheng, Tianjie Zhao, Haishen Lü, Jiancheng Shi, Michael H Cosh, Dabin Ji, Lingmei Jiang, Qian Cui, Hui Lu, Kun Yang, et al. 2022. Assessment of 24 soil moisture datasets using a new in situ network in the Shandian River Basin of China. *Remote Sensing of Environment* 271 (2022), 112891.

[70] Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. 2019. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.* 394 (2019), 56–81.

A APPENDIX

A.1 Data Availability

All datasets we use are publicly available and several (such as the in-situ soil moisture sensor data) have redistribution restrictions.

Data Acknowledgements. This research is made possible by the New York State (NYS) Mesonet. Original funding for the NYS Mesonet was provided by Federal Emergency Management Agency grant FEMA-4085-DR-NY, with the continued support of the NYS Division of Homeland Security & Emergency Services; the State of New York; the Research Foundation for the State University of New York (SUNY); the University at Albany; the Atmospheric Sciences Research Center (ASRC) at the University at Albany; and the Department of Atmospheric and Environmental Sciences (DAES) at the University at Albany.

A.2 Model Architecture

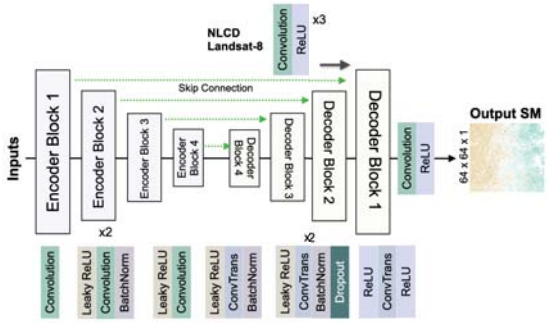


Figure 10: Soil moisture prediction architecture: the Encoder-Decoder extracts latent features from various data sources, which along with high-resolution land-cover maps and Landsat images are used to produce 30m soil moisture maps. Both input and target images are 64x64 pixels.

A.3 Dataset Description

Table 2: In-situ soil moisture observational networks used for DeepSoil predictions. Individual networks are managed by different federal and state agencies.

Soil Moisture Network/No. of stations	
TexMesonet/32 [4]	Nebraska Mesonet/18 [59]
Texas Soil Observation Network/69 [13]	Oklahoma Mesonet/114 [11, 43]
Kentucky Mesonet/42 [39]	USCRN/157 [9, 20]
Missouri Agricultural Database/3	Soil Scape/54 [45]
New York State Mesonet/121 [12]	Scan/178
Illinois Climate Network/19 [47]	Snotel/350 [57]
Montana Mesonet/14 [2]	NoahHMT/21 [66]
Delaware Environment Observing System/24	[34]

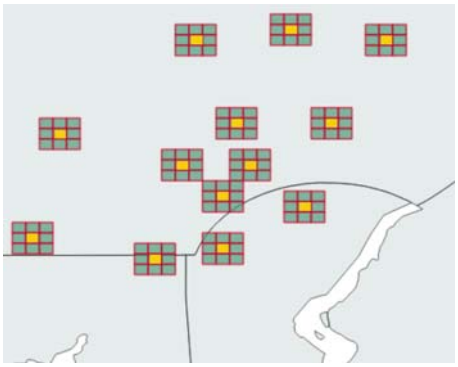


Figure 11: Example of multiple spatial extents in Northern Delaware represented as a 3x3 grid arrangement of 9 tiles. Here, each tile is 64x64 pixels in size. The center tile in each spatial extent is colored yellow and encapsulates an in-situ station. Each tile has an associated unique 14-character key.

Table 3: Datasets leveraged for training and inferences using DeepSoil.

Dataset	Training size	Inference Dataset size	Spatial Resolution	Temporal Resolution
gNATSGO	1.6 GB	8.6 GB	30m	Static
Polaris	2.4 GB	45 GB	30m	Static
Landsat	351 MB	8 GB	30m	8 days
NLCD	114 MB	2.3 GB	30m	Static
Köppen Climate	76 MB	1.7 GB	1km	Static
DEM	241 MB	4.2 GB	30m	Static
GridMET	3.3 GB	92 MB	4km	Daily
MCD15A3H (Interpolated)	41 GB	13.6 GB	500m	4 days
SMAP	1.2 GB	93 MB	9km	2-3 days
HydroBlocks	70 GB	-	30m	2-3 days
In-situ stations	1.4 GB	-	-	Daily
Total	121.68 GB	83.56 GB	-	