DISCERN: Leveraging Knowledge Distillation to Generate High Resolution Soil Moisture Estimation from Coarse Satellite Data

Abdul Matin, Paahuni Khandelwal, Shrideep Pallickara, Sangmi Lee Pallickara
Department of Computer Science
Colorado State University
Fort Collins, USA
{Abdul.Matin,Paahuni.Khandelwal,Shrideep.Pallickara,Sangmi.Pallickara}@colostate.edu

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386179

***Abstract*—Accurate estimation of soil moisture is crucial for efficient agricultural management and environmental monitoring. However, the task of predicting soil moisture levels becomes challenging in regions with limited data availability. In this study, we propose a knowledge distillation-based deep learning approach to enhance soil moisture prediction with machine learning apporach using the low resolution but wide coverage soil moisture Active Passive (SMAP) satellite data.**

Our framework leverages the knowledge distillation, where a high-capacity teacehr model (VGG13) which is pre-traineed on a large dataset (SMAP) and a lightweight student model (ResNet8) which is then trained on sensor-based highly accurate but extremely sparse station data. The student model benefits from the distilled knowledge of the teacher model, acquiring a deeper understanding of the underlying patterns and relationships in the data.

The space-efficient student model significantly reduces the inference time with high prediction accuracy and demonstrates the potential benefit to agricultural management, water resource planning, and ecological studies by providing accurate and reliable soil moisture predictions in data-scarce regions. Our findings reveal how to identify performant settings for achieving the best trade-off between accuracy and model complexity.

***Index Terms*—knowledge distillation, smap, soil moisture, vgg, resnet**

I. INTRODUCTION

Soil moisture (SM) refers to the amount of water existing within the pore spaces of the soil. SM is a vital observation in various geoscientific domains such as plant sciences, meteorology, monitoring of floods and droughts, climate change, and precision agriculture [1, 2]. Measurements of SM have been performed using in-situ sensors and remote sensing. In-situ sensors have traditionally been the dominant way to measure SM, and often provide the most accurate and detailed measurements. For example, the dielectric probe is considered one of the most reliable methods achieving an accuracy of over 96% [3]. However, landscape heterogeneity entails installation of sensors every 1-100 m to capture high variability in SM with soils type, composition, and climactic conditions. The resulting costs stemming from deployment and maintenance of sensor networks precludes practicality. Satellite-based remote sensing is an effective way to monitor SM over a large spatial extent. Soil Moisture Active Passive (SMAP) [4], Soil Moisture Ocean Salinity (SMOS) [5], and ESA's Sentinel missions [6] provide SM datasets. However, estimating SM from satellite measurements can be limited by the attenuation of the atmosphere, high cloud coverage, and sparse revisit frequencies. Also, the spatial resolutions available in current satellite systems are very coarse to capture variability of SM for local or field scale decision making (Fig. 1 (a)). Hydrological models are widely used for estimating SM by simulating physical processes by considering interactions between components of the water cycle such as runoff, evapotranspiration, and precipitation [7]. However, topographical variability and distribution of challenging features, alongside the significant lack of ground observations makes modeling, parametrization, and application of such hydrological models difficult to estimate SM at the desired resolution [8]. Recently, there have been several studies using machine learning approaches to capture nonlinear relationships across ancillary conditions and complexity of soil characteristics [8, 9]. Constructing machine learning models to estimate high-resolution SM map poses several computational challenges. First, the number of ground-based SM measurements is significantly low. The world largest Soil Moisture network – the International SM Network (ISMN) – federates data from 71 networks and more than 2800 stations globally [10]. Although these networks provide critical datasets for SM research and modeling, the distribution density of these stations is too sparse to capture spatial variability of SM values for training machine learning models. Second, SM estimation is closely related to ancillary observations in the surrounding area. Integrating ambient observations with mismatching temporal and spatial resolution is highly challenging to achieve acceptable accuracy. Third, deployment of a deep learning model to estimate fast evolving SM measures requires substantial computational resources to generate timely outputs. Finally, highly complex machine learning models with extensive inference times are not practical for real-world applications.

In this paper, we present our novel approach, DISCERN

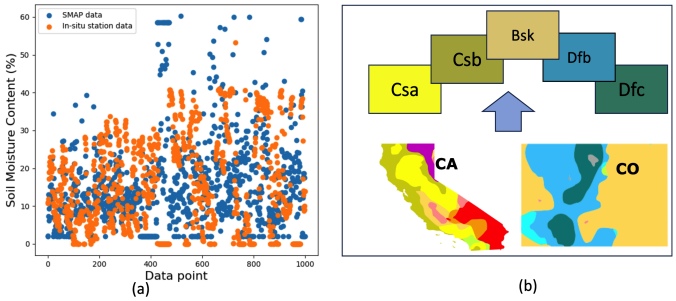

Fig. 1: (a): 1000 random SMAP and In-situ station SM data points in Colorado, year: 2020-2021 (b): Candidate teacher model accuracy

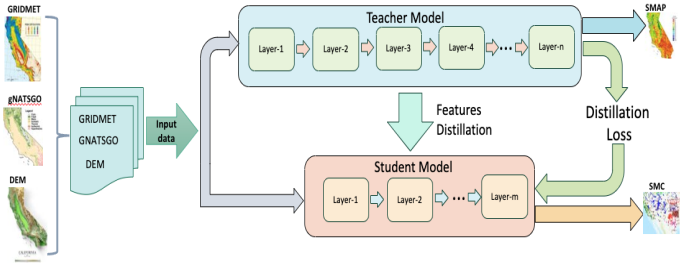

Fig. 2: Overview of the DISCERN Architecture

that estimates SM using a machine learning model with the SM measurements from limited number of ground stations. To address the aforementioned challenges, we use knowledge distillation that transfers knowledge from a large and more complex model (the teacher) to a simpler, space-efficient model (the student) that is computationally frugal for time-sensitive applications. Our teacher model captures nonlinear relationships between the SM and weather as well as topographical conditions using geospatially contiguous satellite measurements at coarse resolutions. Later, the knowledge about this relationship is transferred to the student model that targets SM estimation with higher precision while maintaining accuracy over regions with extremely sparse (or no) in-situ observations.

A. Research Questions

Research questions that we explore in this study include, RQ-1: How can the model capture nonlinear relationships between soil moisture and ancillary conditions over regions with sparse in-situ observations? RQ-2: How can the model generate accurate soil moisture estimations while ensuring reduced inference times? RQ-3: How can we facilitate faster training of models for regions with distinctive topographical and climatic characteristics?

B. Overview of Approach

In this study, our goal is to accurately estimate SM at scale while accounting for their significant geospatial variability. We accomplish this using deep learning models that are trained on sparsely labeled data samples. To this end, we first integrate multi-modal datasets encompassing climate information, soil properties, elevation data, and satellite images. These datasets are available for the spatial scales that we consider (the continental United States or CONUS); however, there are significant variations in both the spatial resolutions and temporal resolutions at which the data are available. We incorporate SMAP satellite images, offering a low-resolution (L3_SM_A 3km product which is model derived from 36km product) view of SM across a broad geospatial extent. We supplement these with a limited number of (spatially sparse) high-precision SM readings from ground stations for the target dataset during model training.

To address the challenge posed by the extreme sparsity of target data samples, our approach entails learning the nonlinear relations between SM and environmental conditions using coarse-resolution SMAP satellite images. We transfer the acquired insights to the final SM estimation model via Knowledge Distillation (KD) techniques. KD involves two distinct models – the teacher and student. Typically, the teacher model has a more complex network structure that captures the feature interactions more comprehensively. On the other hand, the student model(s) are often smaller and simpler while distilling the complex relationships by encapsulating knowledge from the teacher model. The teacher model (VGG13, [11]) in DISCERN learns interactions between SM and ancillary observations by leveraging available (but low resolution) SMAP satellite images. Our student model (ResNet8 [12]) extracts knowledge from the teacher model using multiple KD methods such as feature map distillation, integrating soft prediction from the teacher model, as well as the loss calculated from the SM readings from the ground stations. Our empirical evaluation demonstrates that our approach improves the validation accuracy by 40% and test accuracy by 22% for the unseen regions. We employ our KD methods to train an ensemble of models over a larger geospatial extent. Due to the significantly coarse spatial resolution of SMAP images and the complexity of our teacher model, it is impractical to train the student model across a substantially large geospatial extent, especially considering the computing resources that this would entail. As a result, our nested DISCERN training scheme involves a teacher model encompassing multiple climatic regions and student models per climatic region. In our empirical benchmarks, we contrast accuracy and training efficiency with the standalone traditional deep learning approach. Our results show that DISCERN can converge with better accuracy earlier than normal approach. Moreover, it can be very useful to estimate SM for the case where we might not have enough stations to train a model. DISCERN can be trained on a different data-rich area with similar climate classification. For instance, DISCERN can predict SM content in California with 94% accuracy while trained on the similar climate zone in Colorado.

C. Paper Contributions and Transitional Impacts

Our methodology for estimating SM includes the following contributions:

1) A novel scheme to achieve highly accurate SM estimations through machine learning with extremely sparse target SM samples.
2) Design of knowledge distillation techniques, that leverage multi-modal target datasets for both the teacher and

student models.

3) A lightweight model trained using knowledge distillation that captures the spatial variability of SM at a high resolution.

Transitional Impacts: Accurate SM estimation at high spatial resolutions is critical for applications in many domains such as agriculture, meteorology, microbial ecology, and plant disease forecasting. The proposed methodology has broad applicability in constructing machine learning models using a combination of extremely sparse (spatially) target datasets with high precision and highly available target datasets with lower precision, allowing for effective capturing spatial variability.

D. Paper Organization

The remainder of the paper is structured as follows: Section II covers Background and Dataset. Section III explains the methodology. In Section IV, we delve into the experimental setup, model performance analysis, model sensitivity analysis, and evaluation of nested training using KD. Section V discusses related works, while conclusions and future directions are outlined in Section VI.

II. Background and Related Work

A. Soil Moisture Estimation

Soil moisture content is defined as the amount of water present in the soil in volume. Although SM is a critical parameter in various domains such as agriculture, hydrology, and climate science, using highly accurate in-situ sensors is not a suitable solution for obtaining SM due to the high spatial variability of SM measurements.

Traditionally, physics-based models have been widely adapted to estimate SM content [13, 14]. These models simulate the movement of water in the vadose zone by considering physical processes such as infiltration, evapotranspiration, and drainage. Environmental parameters, such as soil properties, weather conditions, and vegetation characteristics, are critical for obtaining accurate values using physics-based models.

Remote sensing techniques, including microwave and thermal remote sensing, have been actively used to capture spatially distributed estimates of SM over large areas [4–6]. These techniques leverage the interaction between electromagnetic radiation and soil properties to infer SM content. For instance, microwave remote sensing exploits the sensitivity of SM to the dielectric properties of the soil, which affects the propagation and scattering characteristics of microwave radiation. Thermal remote sensing, on the other hand, utilizes the difference in thermal properties between wet and dry soil to estimate moisture content. However, the spatial and temporal heterogeneity of soil properties necessitates the need for SM measurements at higher spatial resolution. Soil Moisture Active Passive (SMAP), a widely known remote sensing observatory is designed to carry two instruments that map SM and determine the freeze or thaw state of the same area being mapped. SM content can be mapped via the radiometer data at a spatial resolution of 36 km every 2-3 days. Details about such satellite datasets are discussed in the following section.

TABLE I: Datasets integrated for training DISCERN deep learning model

Dataset	Source	Spatial Resolution	Temporal Resolution
gNATSGO	NRCS USDA	10m	one time
gridMET	NRCS USDA	4km	daily
DEM	USGS	30m	one time
SMAP	NASA	3km	daily
In-situ station	NOAA	single point	daily

B. Dataset and Study Area

In this study, one of the primary challenges is the heterogeneous spatial and temporal resolution of different datasets that we integrate. As depicted in Table I, we used the Gridded National Soil Survey Geographic (gNATSGO) Database [15] which is a 10m resolution with multiple bands (9 bands were used for this study) representing different properties of soil. The Gridded Surface Meteorological (gridMET) dataset [16] offers daily surface meterological data such as temperature, wind, humidity at the sptial resolution of 4km, 1/24 degree. We used 10 different bands of gridMET data.

We also integrate the Digital Elevation model (DEM) [17], which is a depiction of topographic characteristics that specifically represents the natural terrain, excluding trees, structures, and other surface features covering entire US. We use elevations at 30m resolution. We use SMAP [4] which directly measures the amount of water in the top 5cm of surface soil everywhere on Earth. SMAP was first launched in January 2015 and data operation started from April 2015. It has several products starting from half orbit 36km. In our study we use L3_SM_A product which is a 3km daily product. The last dataset that we use is in-situ sensor based weather station data that we primarily collect from National Oceanic and Atmospheric Administration (NOAA) data repository for 146 stations across California and Colorado. The temporal range for our study is 2020 to 2021. According to the experimental scenario, these states are divided into multiple sub-regions based on climatic characteristics. We organize the datasets into two scenarios. The first parts of experiments were performed on scenario-1 which includes data from Colorado that has an area of 269,837 km^2. For scenario-2, we utilized the Köppen Climatic Classification system [18] to segment our study area. The Köppen Classification is a widely employed climate classification system that offers hierarchical climatic regions. we chose five different climate classes (Bsk, Csa, Csb, Dfb, Dfc) and incorporate datatset for those classes from California and Colorado area.

C. Related Work

1) Physics based Approaches: Over the past few decades, considerable research has been devoted to simulating and predicting SM dynamics based on historical data and meteorological variables [19, 20]. Physics-based models (PBMs), primarily based on Richards' equation [21], have demonstrated favorable performance in various complex scenarios, benefitting from reliable descriptions of physical processes [13]. However, these traditional approaches often struggle to capture complex nonlinear relationships inherent in SM dynamics.

2) Machine Learning Approaches: The data-driven machine or deep learning techniques have widely been used

to map the relationship between input features including geographical and meteorological variables and SM output.

Santi et al. [22] introduced an ANN-based approach to estimate daily SM at a spatial resolution of 10 km. Their model utilized backscatter, local incidence angle, azimuth angle, Latitude, Longitude information from Advanced Scatterometer (ASCAT), and SM data from the International SM Network (ISMN) for training.

In a similar vein, Lee et al. [23] investigated twenty-five different variations of an ANN-based deep learning model for estimating daily SM at a spatial resolution of 4 km.

3) Knowledge Distillation: The concept of Knowledge distillation (KD) was first introduced by Hinton et al. in 2015 [24]. Knowledge distillation has been successfully applied to a wide range of tasks and domains. DistilBERT [25] is a distilled version of BERT, which applies knowledge distillation to compress large-scale language models while maintaining their performance. Zhang et al. [26] explored attention-based self-distillation, where the student model learns to attend to important features by distilling attention maps from the teacher model.

Li et al. [27] investigated cross-modal distillation for few-shot learning, leveraging knowledge transfer between different modalities to enhance the student model's ability to generalize and learn from limited labeled data. Recently, Wang et al. [28] proposed Semantic Calibration for Cross-Layer Knowledge distillation (SemCKD) to solve the performance reduction due to negative regularization from the mismatch of layer semantics between teacher and student networks.

The majority of knowledge distillation (KD) studies have demonstrated enhanced performance across various classification tasks, particularly in the realm of image classification. Our innovative approach involves harnessing KD to create finely detailed SM maps using low-resolution satellite data. To the best of our knowledge, this approach stands as a unique contribution to this field.

III. Methodology

Our proposed approach encompasses model, data integration, and model training strategies working in concert with each other to achieve one goal of generating accurate SM estimations in a timely manner.

A. Data Wrangling:[RQ1]

Our methodology involves training complex deep learning models that require large amounts of training samples. As depicted in Table 1, we utilize a set of diverse datasets that are acquired from in-situ sensors, satellites, models and surveys. To align datasets from multiple sources, we selected a temporal range (2020-2021) and geospatial extents (Colorado and California, USA) that are common across the training data samples. Since DISCERN involves two types of models with diverse spatial resolutions, we perform two sets of data cleaning processes. For the teacher model, we downscaled gNatsgo (10m), DEM(30m) datasets, and then merged gridMET(4km) with DEM dataset to align with the resolution of SMAP satellite images (3km). To align mismatching temporal ranges, we first preprocess the SMAP data. Then we considered only those dates for which we had valid SMAP data. After preprocessing, we had more than 13 million input output data samples for the teacher model. Meanwhile, we use the highest possible resolution for the student model training. The detailed descriptions of data preprocessing for each experiment are included in the section 4.

B. Designing a Deep Network to Estimate SM: [RQ1, RQ2]

One of the primary challenges to develop a deep learning model to estimate SM lies in the scarcity of SM readings from ground stations, which are crucial as labels during model training. To overcome this challenge, DISCERN leverages the non-linear relationships between the soil, topographical and weather conditions of the area using low-resolution SM map obtained from the SMAP satellite. Subsequently, the insight gained are transferred to a light-weight deep learning model, ultimately trained using SM measures from the ground stations. By harnessing the knowledge and expertise encoded in the teacher model of DISCERN, the student model can achieve comparable performance while requiring reduced computational resources and reduced memory footprints thus enabling the deployment of deep learning networks in resource-constrained environments. In our proposed model, we used VGG13 [11] as the teacher model, while comparatively lightweight networks, ResNet8 [12] as the student model. The MobileNetV2 [29] was used as a student model for an experiment to compare the performance of ResNet8 as student model. To select the teacher model we contrasted the accuracy and performance for all the candidate teacher model shown in Fig. 3 to find out the most fitted one. We have tested complex networks ResNet34 and VGG13 as the teacher model. The network is dense and there are around 21.28M learnable parameters. A speciality of ResNet is the shortcut connection between layers which helps skip one or more layers while training to maintain generalization of the model. The other deep model, VGG13 has 13 convolution layers followed by three fully connected layers with 9.41M learnable parameters. Considering the model performance and memory footprint, VGG13 has been selected as the teacher for DISCERN to conduct further analysis.

C. Teacher Model to capture relationship between SM and environmental conditions in the surrounding area

The input for the teacher model consists of meteorological and weather related information that have impact on SM content. We combine all the nine bands of gNATSGO, 10 bands of gridMET and single band of DEM data into a 3D array. So the dimension of each input sample becomes 32x32x10 which are then can be fed to the teacher model for training.

We use L1 loss function (equation 2) for VGG13 based on empirical analysis. The loss is calculated against the model prediction and SMAP data and then the parameter weights are adjusted accordingly with a multiplication factor

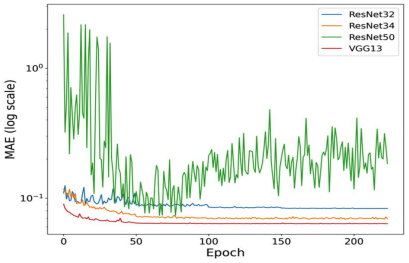

Fig. 3: Accuracy of all the candidate teacher models

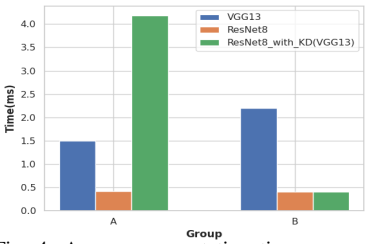

Fig. 4: Average computation time per sample (a): Training, (b): Inference

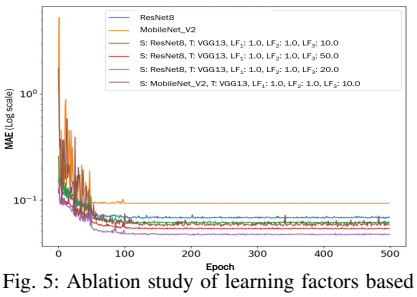

Fig. 5: Ablation study of learning factors based on Accuracy (Area: Colorado)

of learning rate. The insights acquired by the teacher model are accessible in diverse formats. We integrate both outputs from the intermediate layers of the teacher model (section 3-B-(3)) and its final prediction outputs (section 3-B-(4)) during the training of the student model.

1) Student Model trained using the SM measurements from the sparse ground stations: The goal of DISCERN's student model is to provide precise SM estimates over a smaller geospatial extent (point level). To capture the spatial variability of SM the student model is trained with environmental conditions and measures such as satellite imagery gNATSGO , climate data from GridMET, and terrain attributes to produce detailed soil maps at a finer-scale spatial resolution. To match the fine spatial resolution of the results, the training samples retain their original spatial granularity. The student model predicts a single decimal value for a single point. To align with the SMAP data, input data samples are taken from the same geospatial bounds where the ground station location falls in. The highest resolution of the input dataset gNATSGO is 10m. To incorporate knowledge distillation, the dimension of input data tensor is kept the same as Teacher model which is 32x32x10.

For the student model, we used ResNet8 and MobileNetV2. It has 78.46K learnable parameters which is quite lower than VGG13 or ResNet34. We also used MobileNetV2 (for data scenario-1) which is relatively heavier than ResNet8 but lighter than VGG13 and ResNet34.

2) Knowledge Transfer using Feature-Map Distillation: Feature-map distillation is a method that involves transferring the knowledge from feature maps of the intermediate layers of a teacher model to a student model. This process enables the student model to capture and replicate the high-level mapping of input and output data acquired by the trained teacher model. Feature-map distillation has demonstrated promising outcomes across diverse domains, spanning computer vision and natural language processing [30].

In this study, DISCERN transfers the output of each layer from the teacher model, which is trained with coarse SMAP satellite images, to the corresponding layer of the student model. We pair the output of each layer based on the model structures and inform the student model training using the loss function proposed in [30].

$$L_{FMD} = \sum_{s_l, t_l} Dist(T^t(F^t_{t_l}), T^s(F^s_{s_l})), \tag{1}$$

where $F_{t_l}{}^t$ and $F_{s_l}{}^s$ stand for the output of each target layer and student layer respectively. The functions $T^t(.)$ and $T^s(.)$ are the methods to transform the feature maps of all the candidate teacher-student layer pairs into pairwise similarity matrices [31]. L_{FMD} is the summation of the loss among all candidate teacher-student layer pairs calculated by Dist(.) function. Based on the similarity matrices between teacher-student layer pairs, the set of layers that have positive or higher than threshold value are dynamically selected for feature distillation. If there is a mismatch in the layer structure between teacher and student model, it is resolved using the dimension reshaping technique. For example, an extra convolution operation is performed to project the required number of features from teacher layer to student layer.

*3) **Loss Function**:* To incorporate learned factors from the teacher model during training with highly accurate SM observations, we introduce a loss function termed $L_{DISCERN}$ integrating three distinct sources: SM readings obtained from sparse ground stations, soft predictions from the teacher model, and the feature distillation loss (section 3-B-3). Initially, as both models function in a regression manner, we have opted for the widely used L1 loss function to assess the disparity between the output of the final layer and the ground truth.

$$\ell(x, y) = L = \{l_1, ..., l_N\}^T, l_n = |x_n - y_n|, \tag{2}$$

where, x is input, y is target and N is the batch size. The loss for the student model, L_S is computed using equation (2), capturing disparity between the student models' prediction and the target station data. Furthermore, the same equation is utilized to calculate the logit loss, L_{KD}, which represents the distinction between the soft prediction of the teacher model and the ground truth of the student model. Lastly, in each batch, feature maps are produced by each layer of the student model, and the discrepancy between the feature maps of the corresponding layer in the teacher model is characterized as L_{FMD} (section 3-B-3). Therefore, the overall loss for each batch is computed as,

$$L_{Discern} = \sum LF_1 L_S + LF_2 L_{KD} + LF_3 L_{FMD}, \tag{3}$$

where LF_1. LF_2,LF_3 defines the factor to be considered for each of the loss L_S, L_{KD} and L_{FMD} respectively. The average loss computed from equation (3) at the end of each epoch is employed to update the weights using a predefined learning factor and subsequently backpropagated to the input layer.

4) Nested Training of the Student Models by Extending Knowledge Distillation: [RQ3] : Our nested training scheme involves a teacher model that covers a broader geographical area and a set of student models based on climate regions. As depicted in Fig. 1 (b), we group regions with similar sets of climate zones. For instance, California and Colorado encompass areas primarily characterized by semi-arid and Mediterranean climates (Csa, Csb, Bsk, Dfb, and Dfc in Köppen's climate classification). The teacher model is trained using only labels retrieved from SMAP at a 3km resolution.

Subsequently, we train an ensamble of student models based on the climate zones, all using the same teacher model. Because there is no interdependence among our student models, we can train them in parallel to achieve scalability. For climate zones with extremely sparse labels, DISCERN employs the student model trained under the most similar climate conditions. We utilize the hierarchical structure inherent in the Köppen climate classification system. Köppen classifies the world into five main groups and assigns subgroups beneath each main group, with each group and subgroup represented by a letter. Subgroups within the same main group share the same prefix.

IV. EMPIRICAL EVALUATIONS

We performed experiments to assess the effectiveness of KD technologies in enhancing model performance in areas with sparse labels and to compare different KD strategies. Our evaluation comprises (1) model performance analysis, (2) model sensitivity analysis, and (3) performance analysis of nested KD learning. We have designed specific evaluation scenarios for each group of experiments, and the data preprocessing for each experiment is detailed in the following section.

A. Experiment Setup

1) Data Preprocessing: Our data preprocessing involves several steps, including aligning the spatial resolutions of the datasets through downscaling (e.g., gNatsgo and DEM), as well as cropping aligned images to focus on the target regions. For the model accuracy evaluation (Section IV-B), we generated approximately 13 million training samples to train the teacher model using observations from the state of Colorado. The spatial resolutions were adjusted to match the resolution of the SMAP satellite images (3km).

The in-situ station SM data was used as the ground truth for student models. We selected samples from each input dataset based on their geospatial proximity to the avaialable ground station location. To support the model sensitivity (Section IV-C) evaluation and nested model (Section IV-D) performance we generated approximately 500,000 samples for the teacher model and 73,000 samples for the student model from datasets covering the states of Colorado and California. This broader geographic scope led us to use a total of 123 unique small geospatial bounds (each containing at least one in-situ weather station) for generating training samples for the student model. Areas with more than 20% invalid or null values in gNatsgo, gridMET or SMAP were excluded. We reduced the spatial extent for the teacher model to account for possible variability in soil properties. We ensured that the test dataset had no overlapping spatial extent with the dataset used for model training to avoid information leakage. We conducted a 70-30% train-test split for teacher model and consider samples from the unique areas for testing the teacher model to accomplish the experiments.

TABLE II: Model Performance with and without Knowledge Distillation (Area: Colorado)

Model	Ground Truth	Validation MAE(%)		Test MAE(%) for unseen region		Improvement(%)	
		without KD	With KD	without KD	with KD	Validation	Test
VGG13	SMAP	5.1		6.1			
ResNet8	Station SMC	6.7	4.6	20.6	17.7	31	14
MobileNetV2	Station SMC	9.2	5.5	20.7	16.2	40	22

2) Model Training Details: As illustrated in Fig. 2, we utilize the outputs and feature maps from the teacher model for training the student model. Initially, we train the teacher model until it reaches a reasonable performance level. To train the student model, we pass the input samples through both the teacher and student models. We then calculate the distillation loss using the features from intermediate layers and the predictions from the output layer of the pre-trained teacher model. Additionally, we compute another loss based on the predictions of the student model. Please note that we use the same ground truth values to calculate the loss for both the teacher and student models. Finally, we calculate the total loss using Equation (3) and backpropagate it to the student model exclusively to adjust the learning weights.

3) Evaluation Metrics: To evaluate the proposed framework, we used the Mean Absolute Error (MAE). MAE between predicted SM (as in %) and target SM readings (as in %) was used to measure the model accuracy. Also, we used MAEs to contrast the effectiveness of DISCERN to other deep learning models. We tracked the average MAE for all the test samples. In addition, we calculated the percentage of improvement, I_{acc} for a better understanding of the contribution of the KD using the following formula.

$$I_{acc} = \frac{MAE_{Model_{w/oKD}} - MAE_{Model_{withKD}}}{MAE_{Model_{w/oKD}}} \times 100 \quad (4)$$

B. Model Performance Analysis: [RQ1, RQ2]

1) Model Accuracy: In this evaluation scenario, we employed VGG13 as the teacher model and ResNet8 as the student model using 70% of all the available data covering the entire area of Colorado. As shown in Table II, DISCERN demonstrates a validation accuracy of 95.4%. Generally, SM ranges from 10% to 45%, except during or after watering. Considering that the most reliable in-situ SM sensor technology, the dielectric probe method, estimates surface SM with an error level of 4% [3], our model's performance is highly reliable for various applications. We also compared the model accuracies with and without the KD approach. The validation accuracy improved by up to 40% when we employed the KD approach for model training.

To assess the effectiveness of our approach in regions that were not part of the model training, we evaluated our model's accuracy in areas remote from the ones used for training. To create the testing data samples for this experiment, we divided

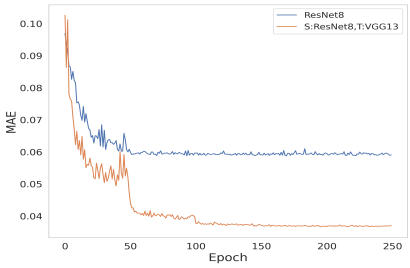

Fig. 6: Model Accuracy of ResNet8 and DISCERN (Area: California and Colorado)

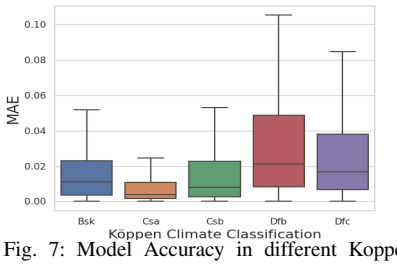

Fig. 7: Model Accuracy in different Koppen Climate Zone (Area: California and Colorado)

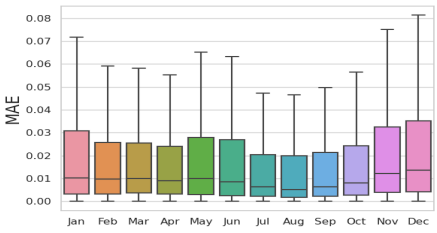

Fig. 8: Model Accuracy in each month (Area: California and Colorado)

the entire Colorado refion into unique 18x12km sub-regions and trained and tested our model with data samples from groups of these sub-regions without any overlap. As shown in Table II, the model's accuracy improved by 14-20% when we applied KD technology to the unseen area.

2) Model Training Time, Inference Latency, and Memory Requirement: In our methodology, the teacher model is a sophisticated deep-learning network, while the student model adopts a space-efficient architecture. In Fig. 4, we compared the training time (for a single sample) and inference latency for VGG13 (teacher model) and ResNet8 (student model) with and without the KD approach.

Because the teacher models have a higher number of learnable parameters (VGG13 has approximately 9.41M parameters in our evaluation), they require more time for training and consume more memory. In contrast, student models are relatively lightweight (ResNet8 has approximately 78.5K parameters in our evaluation) and require less time for training, as well as frugal memory consumption [32]. Due to the involvement of VGG13 while training a student model, the training time for the student model was longer compared to the case without the KD approach. However, our model achieves equivalent inference latency (0.4 milliseconds) compared to ResNet8 without KD. Overall, our approach provides highly accurate predictions with significantly lower latency (5.5 times lower than VGG13) compared to the complex teacher model (VGG13 at 2.2 milliseconds).

3) Effectiveness of the KD components: To understand how each component of KD in the distillation loss contributes to model training, we conducted a microbenchmark by applying different sets of weights within our loss function while using ResNet8 as the student model. In Fig. 5, we compare the effectiveness of the KD approaches that DISCERN collectively employs. In the figures, LF1, LF2, LF3 are the three multiplicative factors which decide how much loss is considered from each of student model loss, teacher model output loss and feature distillation loss respectively (equation. 3). Empirically we can find the appropriate combination that gives the best improvement.

C. Model Sensitivity Analysis: [RQ1, RQ3]

To assess model performance under various conditions, we conducted sensitivity analysis across seasons and climatic zones. To enhance diversity and have sufficient training samples, we expanded our target area to include both Colorado and California in this evaluation. For this, we utilized VGG13 as the teacher model and ResNet8 as the student model. With the extended spatial extent, our student model achieved a similar overall accuracy, with a mean absolute error (MAE) of 5.9% (Fig. 6). Furthermore, the use of KD improved the student model's accuracy by 37% compared to ResNet8 without KD.

We evaluated the model performance over different climactic zones specified in the Koppen climate classification system. We compared the model performance across five climate zones that are most popular in Colorado and California. Our model performs the best in the Cs (Csa and Csb, dry summer climate) and Bsk (Semi-arid climate). These climate zones are the majority of Colorado and California. The individual average MAE of each climate class was lower than 2% throughout these regions (Fig. 7). The model accuracy in the Dfb (warm summer humid continental climate) region was relatively lower than other climate zones due to the relatively low number of samples from the area.

Fig. 8 presents a monthly sensitivity analysis, which assesses the model's ability to predict SM effectively across different months of the year. The analysis highlights the model's strongest performance during the months of July, August, and September when the weather patterns in California and Colorado exhibit notable consistency. In contrast, predictions during the winter months exhibit a comparatively wider range of mean absolute error (MAE) variation when compared to other seasons. These variations can be attributed to unique outliers primarily caused by snow effect. During this time, usually the soil surface is covered with snow most of the time and the external factors like temperature, wind have little effect on the SM [33].

D. Performance Evaluation of Nested Training[RQ3]

DISCERN harnesses insights captured by the sophisticated teacher model using coarse-resolution observations. In this section, we assess the effectiveness of DISCERN in training diverse student models, which is applicable for model training over a large spatial extent. As described in Section 3-4, our nested training involves a single teacher model and shared by multiple student model based on the climate zones.

To evaluate our model's capability to handle areas with extremely sparse labels, we assessed model accuracy using teacher models trained in different areas or under similar climatic zones. Fig. 7) demonstrates the consistently high accuracy observed across these diverse experiments. Notably, we observed that a single teacher model pre-trained with large

geospatial extent can assist multiple smaller student models regularizing to local climate zone effectively.

V. Conclusion

Accurate prediction of SM holds considerable significance for effective agricultural management and environmental monitoring. The challenge of predicting SM in regions characterized by limited data availability necessitates innovative solutions. This study describes a novel approach through the utilization of knowledge distillation-based deep learning techniques.

By harnessing the knowledge distillation framework, we have amalgamated the capabilities of a high-capacity VGG13 model as the teacher and a lightweight ResNet8 model as the student [RQ1]. The teacher model, enriched by its pre-training on extensive Soil Moisture Active Passive (SMAP) data, imparts its intricate insights to the student model, which in turn, is trained on sparsely available sensor-based station data.

Notably, the adoption of the streamlined student model not only enhances prediction accuracy but also significantly reduces inference time, rendering it conducive to real-time applications [RQ2]. Our approach improves throughput of the model inference that is critical for various applications such as agricultural, water resource planning and ecological studies. By leveraging high level knowledge representation of mapping topographical and climate characteristics with SM in larger spatial extent from the trained DISCERN, the student model can be trained anytime with significantly reduced latency on a different spatial extent preserving high accuracy[RQ3].

In summary, this research underscores the capacity of knowledge distillation to overcome challenges posed by limited data availability.

VI. Acknowledgement

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

References

[1] R. Liao *et al.*, "Development of a soil water movement model for the superabsorbent polymer application," *Soil Science Society of America Journal*, vol. 82, no. 2, pp. 436–446, 2018.

[2] M. Feki *et al.*, "Impact of infiltration process modeling on soil water content simulations for irrigation management," *Water*, vol. 10, no. 7, p. 850, 2018.

[3] P. K. Sharma *et al.*, "Assessment of different methods for soil moisture estimation: A review," *J. Remote Sens. GIS*, vol. 9, no. 1, pp. 57–73, 2018.

[4] D. Entekhabi *et al.*, "The soil moisture active passive (smap) mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[5] Y. H. Kerr *et al.*, "Overview of smos performance in terms of global soil moisture monitoring after six years in operation," *Remote Sensing of Environment*, vol. 180, pp. 40–63, 2016.

[6] D. Phiri *et al.*, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.

[7] S. Su-fang *et al.*, "Soil moisture forecast model based on meteorological factors in jinhua city," *Chinese Journal of Agrometeorology*, vol. 30, no. 02, p. 180, 2009.

[8] Y. Cai *et al.*, "Research on soil moisture prediction model based on deep learning," *PloS one*, vol. 14, no. 4, e0214508, 2019.

[9] N. Li *et al.*, "Research of adaptive genetic neural network algorithm in soil moisture prediction," *Computer Engineering and Applications*, vol. 54, no. 01, pp. 54–59, 2018.

[10] W. Dorigo *et al.*, "The international soil moisture network: Serving earth system science for over a decade," *Hydrology and earth system sciences*, vol. 25, no. 11, pp. 5749–5804, 2021.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] R. A. Freeze and R. Harlan, "Blueprint for a physically-based, digitally-simulated hydrologic response model," *Journal of hydrology*, vol. 9, no. 3, pp. 237–258, 1969.

[14] J. R. Nimmo, "The processes of preferential flow in the unsaturated zone," *Soil Science Society of America Journal*, vol. 85, no. 1, pp. 1–27, 2021.

[15] S. S. Staff, "Gridded national soil survey geographic (gnatsgo) database for the conterminous united states," *United States Department of Agriculture, Natural Resources Conservation Service*, 2020.

[16] J. T. Abatzoglou, "Development of gridded surface meteorological data for ecological applications and modelling," *International Journal of Climatology*, vol. 33, no. 1, pp. 121–131, 2013.

[17] "NASA JPL(2013).nasa shuttle radar topography mission global 1 arc second number," Accessed 2023-06-05. DOI: https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL1N.003.

[18] M. Kottek *et al.*, "World map of the köppen-geiger climate classification updated," 2006.

[19] F. Karandish and J. Šimnek, "A comparison of numerical and machine-learning modeling of soil water content with limited input data," *Journal of Hydrology*, vol. 543, pp. 892–909, 2016.

[20] L. Wang and J. J. Qu, "Satellite remote sensing applications for surface soil moisture monitoring: A review," *Frontiers of Earth Science in China*, vol. 3, pp. 237–247, 2009.

[21] L. A. Richards, "Capillary conduction of liquids through porous mediums," *physics*, vol. 1, no. 5, pp. 318–333, 1931.

[22] E. Santi *et al.*, "Application of artificial neural networks for the soil moisture retrieval from active and passive microwave spaceborne sensors," *International journal of applied earth observation and geoinformation*, vol. 48, pp. 61–73, 2016.

[23] C. S. Lee *et al.*, "Estimation of soil moisture using deep learning based on satellite data: A case study of south korea," *GIScience & Remote Sensing*, vol. 56, no. 1, pp. 43–67, 2019.

[24] G. Hinton *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] V. Sanh *et al.*, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[26] S. Henkel, "Modern hopfield networks for few-and zero-shot cancer subtype classification/submitted by stephanie henkel," 2022.

[27] Q. Li *et al.*, "Knowledge distillation on cross-modal adversarial reprogramming for data-limited attribute inference," in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 65–68.

[28] H. Zhang *et al.*, "Adaptive multi-teacher knowledge distillation with meta-learning," *arXiv preprint arXiv:2306.06634*, 2023.

[29] M. Sandler *et al.*, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[30] J. Yim *et al.*, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[31] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[32] S. Bianco *et al.*, "Benchmark analysis of representative deep neural network architectures," *IEEE access*, vol. 6, pp. 64 270–64 277, 2018.

[33] M. Litaor *et al.*, "Topographic controls on snow distribution, soil moisture, and species diversity of herbaceous alpine vegetation, niwot ridge, colorado," *Journal of Geophysical Research: Biogeosciences*, vol. 113, no. G2, 2008.