Fusion of Forecasting and Retrieval for Uncertainty Aware Soil Moisture estimation

Grigorios Tsagkatakis[†,*], Amer Melebari[*], James D. Campbell[*], Archana Kannan[*], Parnia Shokri[*], Erik Hodges[*], Mahta Moghaddam[†,*], Panagiotis Tsakalides[†,*]
[†]Institute of Computer Science, Foundation for Research and Technology - Hellas, Greece.
[‡]Computer Science Department, University of Crete, Greece.
[*] Department of Electrical and Computer Engineering, University of Southern California, USA.

***Abstract*—There is a need for accurate and timely estimation of environmental variables such as soil moisture. Instantaneous retrieval using observations from remote sensing platforms. Challenges in the retrieval and scarcity of measurements. Data Assimilation employs physical models that integrate the remote sensing observation in order to produce a spatio-temporal complete and reliable estimation, however, at the cost of processing time, which induces latency in the generation of the data. In this work, we propose an uncertainty-aware machine learning framework for the integration of forecasting and retrieval estimations of geophysical parameters under a Bayesian framework. Experimental results using observations from the CYGNSS global navigation satellite system-reflectometry (GNSS-R) platform and estimations from the SMAP L4 product indicate that the proposed approach achieves superior results in soil moisture retrieval, both in terms of retrieval accuracy and uncertainty estimation.**

***Index Terms*—Data assimilation, uncertainty quantification, machine learning, soil moisture, SMAP, CYGNSS**

I. Introduction

Accurately and reliably estimating geophysical parameters is essential for numerous environmental and agricultural applications. Among these, the precise estimation of surface Soil Moisture (SM) stands out as a critical component for understanding and managing agricultural productivity, optimizing water resource allocation, and predicting droughts [1]. To achieve a reliable estimation of SM, three main sources of data are available, namely, observations from *remote sensing* platforms, measurements from *in situ* sensor networks, and estimates generated by physics-driven *data assimilation* processes [2, 3].

While SM retrieval from in situ measurements offers high accuracy, this type of approach lacks spatial coverage. Remote sensing technologies have revolutionized the field by providing large-scale soil moisture estimates. Satellites such as the SMAP [4] and the SMOS [5] satellites utilize radiometers and radar to measure soil moisture with high spatial resolution. More recently, global navigation satellite system (GNSS) missions like Cyclone GNSS (CYGNSS) [6] have emerged as a complementary approach, leveraging reflected Global Positioning System (GPS) signals to infer soil moisture with enhanced temporal resolution [7]. Despite their abilities, these methods face challenges such as temporal gaps and sensitivity to surface conditions.

Data assimilation techniques are pivotal in geophysical parameter estimation by integrating observational data with numerical models to produce more accurate and reliable estimates. Methods such as the Kalman Filter and its variants are commonly employed to update model states based on new observations [8]. The integration of satellite-derived soil moisture data from missions such as SMAP and SMOS into land surface models has significantly enhanced streamflow predictions and agricultural management practices [9]. However, data assimilation (DA) processes face several challenges, including high computational complexity, model mis-specifications, and the accurate representation of uncertainties.

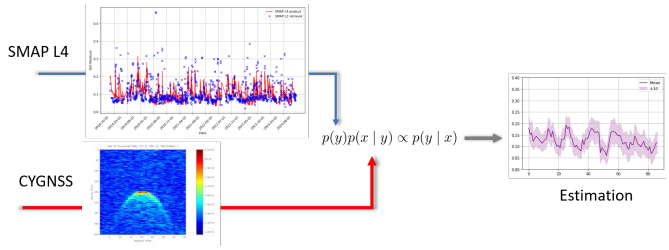

Fig. 1. Block diagram of the proposed framework. The proposed scheme accepts historical estimations from data assimilation models and measurements from remote sensing to generate the soil moisture predictions.

More specifically, we propose a framework to integrate both sources of estimation—DA and remote sensing retrieval—into a unified approach, as illustrated in Fig. 1. On one side, we propose leveraging forecasting to predict future states using historical data derived from DA. On the other side, we employ remote sensing retrieval to extract critical information about current conditions from observational data. It is worth noting that despite the significance of these tasks, they have traditionally been treated as independent processes, thereby limiting their potential to exploit mutual information effectively [10].

Formally, this paper proposes a novel framework that fuses forecasting and retrieval through a Bayesian approach. The Bayesian framework provides a natural mechanism for combining prior knowledge with observational evidence, enabling

This work was supported by the TITAN ERA Chair project (contract no. 101086741) within the Horizon Europe Framework Program of the European Commission.

joint modeling of both tasks while quantifying uncertainty. By integrating probabilistic forecasting outputs with retrieval processes, our method not only improves prediction accuracy but also enhances interpretability through uncertainty estimation.

The contributions of this work are threefold:

1) We introduce a Bayesian fusion framework that jointly models forecasting and retrieval tasks, enabling seamless integration and mutual reinforcement.
2) We demonstrate the effectiveness of the proposed approach on *[specific dataset or application]*, showcasing improved performance and uncertainty quantification.
3) We provide a detailed analysis of the implications of uncertainty modeling, highlighting its benefits for reliable decision-making.

II. Proposed Framework

A. Formulation

Formally, we assume the availability of a set of N time-indexed input-output pairs encoded in the dataset $\mathcal{D} = \{(\mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)}, y_t^{(i)})\}_{i=1}^N$, where t represents the time index.

In the case of **retrieval**, the inputs $\mathbf{x}_t^{(i)} \in \mathbb{R}^{d_x}$ and $\mathbf{z}_t^{(i)} \in \mathbb{R}^{d_z}$ represent the remote sensing and potentially ancillary observations, respectively. The target variable $y_t^{(i)} \in \mathbb{R}$ represents the surface soil moisture value at time t. This formulation focuses on estimating soil moisture values based on instantaneous observations $(\mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)})$, making it suitable for real-time applications.

For **forecasting**, the model exclusively utilizes the historical time series of soil moisture values to predict future values. The input at time t is defined as $y_{t-\tau+1:t}^{(i)}$, which represents the soil moisture values over a lag window of size τ. The forecasting target is $y_{t+\Delta t}^{(i)} \in \mathbb{R}$, representing the soil moisture value at a future time $t + \Delta t$. Here, Δt is the forecasting horizon. By focusing solely on the temporal dependencies within the soil moisture time series, this approach enables efficient and accurate predictions of future soil moisture dynamics.

B. Probabilistic Regression

In traditional regression tasks, point-like estimation predicts a single deterministic value $\hat{y}$ for the target y. These predictions are often obtained by assuming a fixed model that minimizes a loss function, such as the mean squared error (MSE), which corresponds to the maximum likelihood estimation (MLE) under specific probabilistic assumptions. For instance, when the conditional distribution $p(y \mid \mathbf{x}, \boldsymbol{\theta})$ is assumed to follow a Gaussian distribution with a fixed variance σ^2, the MLE of the mean μ is the value that minimizes the squared error:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{N} (y^{(i)} - \mu)^2. \tag{1}$$

In this case, the point estimate $\hat{\mu}$ corresponds to the conditional mean of the target variable given the inputs, while the variance σ^2 is often assumed constant and is not explicitly modeled.

While point-like estimation is effective for deterministic predictions, it fails to account for the uncertainty associated with the predictions, which is particularly important in applications like soil moisture estimation where observational noise, model error, and data variability play a significant role.

To address these limitations, **probabilistic regression** extends this framework by modeling the entire conditional distribution $p(y \mid \mathbf{x}, \boldsymbol{\theta})$ rather than a single point. For example, when using a Gaussian distribution, the parameterization involves estimating both the mean μ and the standard deviation σ of the distribution:

$$p(y \mid \mathbf{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{2}$$

where $\boldsymbol{\theta} = [(\mu, \sigma)]$.

C. Probabilistic Regression for Retrieval and Forecasting

For **retrieval**, the task involves estimating the parameters μ and σ of the conditional distribution $p(y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\theta}_r)$, given the remote sensing observations $\mathbf{x}$, ancillary data $\mathbf{z}$ and model parameter $\boldsymbol{\theta}_r$. The MLE approach is applied to optimize the likelihood of the observed data:

$$\mathcal{L}_{\text{retrieval}} = -\sum_{i=1}^{N} \log p(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \boldsymbol{\theta}_r) \tag{3}$$

which for a Gaussian distribution simplifies to

$$\mathcal{L}_{\text{retrieval}} = \sum_{i=1}^{N} \left(\frac{1}{2\sigma^2}(y^{(i)} - \mu)^2 + \log\sigma\right) \tag{4}$$

This loss encourages the model to fit both the central tendency μ and minimize the dispersion σ of the soil moisture estimates.

For **forecasting**, probabilistic regression focuses on the time series of soil moisture values $y_{t-\tau+1:t}$. The model estimates the parameters μ and σ of $p(y_{t+\Delta t} \mid y_{t-\tau+1:t}, \boldsymbol{\theta}_f)$, where the input is the lagged sequence of past values, and the target is the future value $y_{t+\Delta t}$ at the forecasting horizon Δt. The corresponding loss function is:

$$\mathcal{L}_{\text{forecasting}} = -\sum_{i=1}^{N} \log p(y_{t+\Delta t}^{(i)} \mid y_{t-\tau+1:t}^{(i)}, \boldsymbol{\theta}_f) \tag{5}$$

and for a Gaussian distribution:

$$\mathcal{L}_{\text{forecasting}} = \sum_{i=1}^{N} \left(\frac{1}{2\sigma^2}(y_{t+\Delta t}^{(i)} - \mu)^2 + \log\sigma\right) + \text{const.} \tag{6}$$

This approach allows the model to predict not just the expected future value μ but also the associated uncertainty σ, providing a more informative and robust forecasting framework.

D. Bayesian Integration Framework

In this work, we adopt a Bayesian framework to fuse soil moisture estimates from forecasting and retrieval. The forecasting model provides prior information $p(y) = \mathcal{N}(\mu_f, \sigma_f^2)$, representing the expected soil moisture mean (μ_f) and associated uncertainty (σ_f). The retrieval model supplies the

likelihood $p(y \mid \mathbf{x}, \mathbf{z}) = \mathcal{N}(\mu_r, \sigma_r^2)$, capturing the information from remote sensing observations ($\mathbf{x}$) and ancillary data ($\mathbf{z}$).

The posterior distribution, representing the fused soil moisture estimate, is also Gaussian:

$$p(y \mid \mathbf{x}, \mathbf{z}) = \mathcal{N}(\mu_p, \sigma_p^2) \tag{7}$$

where the posterior variance (σ_p^2) and mean (μ_p) are given by:

$$\sigma_p^2 = \left(\frac{1}{\sigma_f^2} + \frac{1}{\sigma_r^2} \right)^{-1} \tag{8}$$

$$\mu_p = \sigma_p^2 \left(\frac{\mu_f}{\sigma_f^2} + \frac{\mu_r}{\sigma_r^2} \right). \tag{9}$$

E. NGBoost for Probabilistic Regression

We consider the natural gradient boost (NGBoost) [11], a recently proposed approach for probabilistic regression. The NGBoost algorithm is defined based on three modules, namely the type and number of base learners $f^{(n)}$, the parametric form of distribution P_θ, and the scoring rule $\mathcal{S}(P_\theta, y)$. During the training of the NGBoost, the training set $\mathbf{x}$ is fit through a sequence of base learners

$$\{f^{(n)}(\mathbf{x})\}_{n=1}^{N} \tag{10}$$

in a sequential fashion following the boosting paradigm [12]. These base learners can be decision trees or other types of (typically) shallow classifiers. The cascade of base learners is employed for estimating the parameters of the distribution P_θ. While in typical point-wise predictions, training amounts to minimizing an appropriate loss function, for the case of probabilistic regression, this role is taken up by a *scoring rule*, which compares the estimated probability distribution to the observations. For the case of Gaussian distributions, a proper scoring rule $\mathcal{S}(P, y)$ is the negative log-likelihood that is given by

$$\mathcal{S}(\theta, y) = -\frac{1}{N} \sum_{i=1}^{N} \log P_\theta(y^{(i)} | x^{(i)}) \tag{11}$$

A key novelty of NGBoost is the introduction of the Natural Gradient during the optimization, which more closely captures the gradient in the parametric space of the distribution, allowing the estimation of the steepest ascent direction. In this work, we consider NGBoost for both retrieval [13] and forecasting, assuming a 5-day forecast horizon.

III. Analysis Ready Dataset

The Analysis Ready Dataset considers surface soil moisture measured between zero and 5 cm depth. We consider the following locations in the USA: JR-1, JR-2, and JR-3 in NM, Kendall and Lucky Hills in AZ, and Z1 and Z4 in CO. All these sites are part of the SoilSCAPE network [14, 15].

For the Data Assimilation, we consider the SMAP Level 4 product [16], which provides estimates of surface and root-zone soil moisture by integrating observations from the SMAP satellite with a land surface model using advanced data assimilation techniques [17, 18]. Specifically, the L4 product incorporates SMAP L-band radiometer measurements into the NASA Catchment Land Surface Model (CLSM) via an ensemble Kalman filter (EnKF) [19]. The SMAP L4 product offers 3-hourly, 9-km resolution estimates of surface (0-5 cm) and root-zone (0-100 cm) soil moisture. It has been operational since March 31, 2015.

To perform instantaneous retrievals, we consider remote sensing observations from the CYGNSS platform. CYGNSS is a NASA Earth Venture mission employing eight low-Earth orbit satellites to perform a GNSS-reflectometry (GNSS-R) measurements. Each satellite is equipped with two nadir-looking antennas for GNSS-R data collection, with four receiving channels enabling simultaneous acquisition of multiple signals. The key measurement output is a delay-Doppler map (DDM), which provides information about the reflected GPS signals. Since August 2019, the processing integration time for each DDM has been 0.5 seconds (2 Hz), offering improved temporal resolution. In this work, we assume a region of 5 km around the specular point to represent the spatial footprint of the DDM. In addition, there are ancillary data, which are the elevation and the ground slope estimated from the one-arc-second SRTM digital elevation model (DEM) [20].

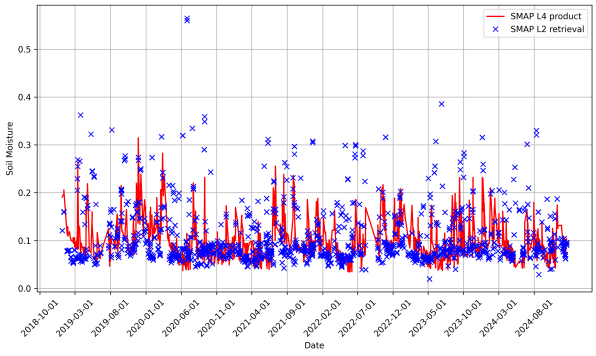

Fig. 2. Plot for SM at the *JR-1* site estimated by DA (SMAP L4 product) and instantaneous retrieval (SMAP L2 product).

Fig. 2 presents a time series of SM estimations from the SMAP L4 DA model from one site (JR-1) over the period used for generating the training data. For reference, the figure also includes the instantaneous retrievals encoded in the SMAP L2 product. The plot demonstrates the impact of temporal evolution in SM encoded in the DA, as well as the discrepancies between DA and retrieval, enforcing the need for accurate uncertainty estimation.

IV. Experimental Results

To quantify performance, we consider two metrics, one capturing estimation accuracy and one for quantification of uncertainty. For the case of accuracy, we consider unbiased RMSE between *ground truth* values in the DA (not utilized during forecasting), and different methods for SM estimation. To quantify the quality of uncertainty estimation, we use the expected calibration error (ECE), which measures the alignment between predicted uncertainty and the empirical

accuracy of the estimates. Specifically, ECE evaluates whether the confidence intervals of the predictions reliably represent the actual error distribution. In both cases, lower values indicate better performance. For the case of forecasting, we assumed a 5-day forecast horizon.

Table I presents the average values for the unbiased RMSE for the three approaches and all sites. The results indicate that in all cases, the best performance in terms of mean-value prediction is typically achieved by either the retrieval or the combined approach.

TABLE I
UNBIASED RMSE (URMSE) METRICS FOR EACH SITE.

Site	Retrieval	Forecast	Combined
JR-1	**0.0281**	0.0318	0.0282
JR-2	**0.0197**	0.0242	0.0203
JR-3	0.0248	0.0310	**0.0247**
Kendall	0.0411	0.0462	**0.0387**
Lucky Hills	0.0357	0.0450	**0.0356**
Z1	**0.0442**	0.0608	0.0522
Z4	0.0322	0.0404	**0.0309**

In addition to estimation accuracy, Fig. 3 presents the values for ECE for all sites and methods under consideration. These results clearly indicate that the superiority of the proposed approach is consistently obtaining the best performance. Furthermore, one can also observe that in some cases, the performance of retrieval is better than forecasting, while in other cases, the inverse holds.

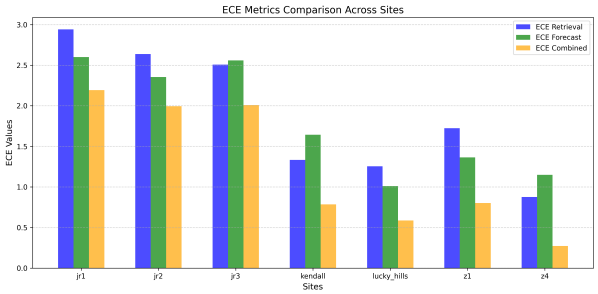

Fig. 3. ECE across different sites for retrieval, forecast, and combined approaches.

Last, to gain a deep understanding of the performance of different methods, Fig. 4 presents an exemplary case of SM estimation from CYGNSS-based retrieval, DA forecasting, and the proposed fusion approach from one site ('JR-1') on the validation set. The plots clearly demonstrate the dramatic improvement in terms of uncertainty of the proposed approach compared to the retrieval and forecasting.

V. CONCLUSIONS

In this study, we presented a novel framework that integrates forecasting and retrieval methods for surface soil moisture estimation. By leveraging a Bayesian fusion approach, our framework combines the temporal consistency of probabilistic forecasting with the observational precision of GNSS-R-based retrieval. This integrated model not only improves prediction accuracy but also provides robust uncertainty estimates, enabling more informed and reliable decision-making. Future work will explore how physical models can be introduced to the forecasting and/or the retrieval part of the process.

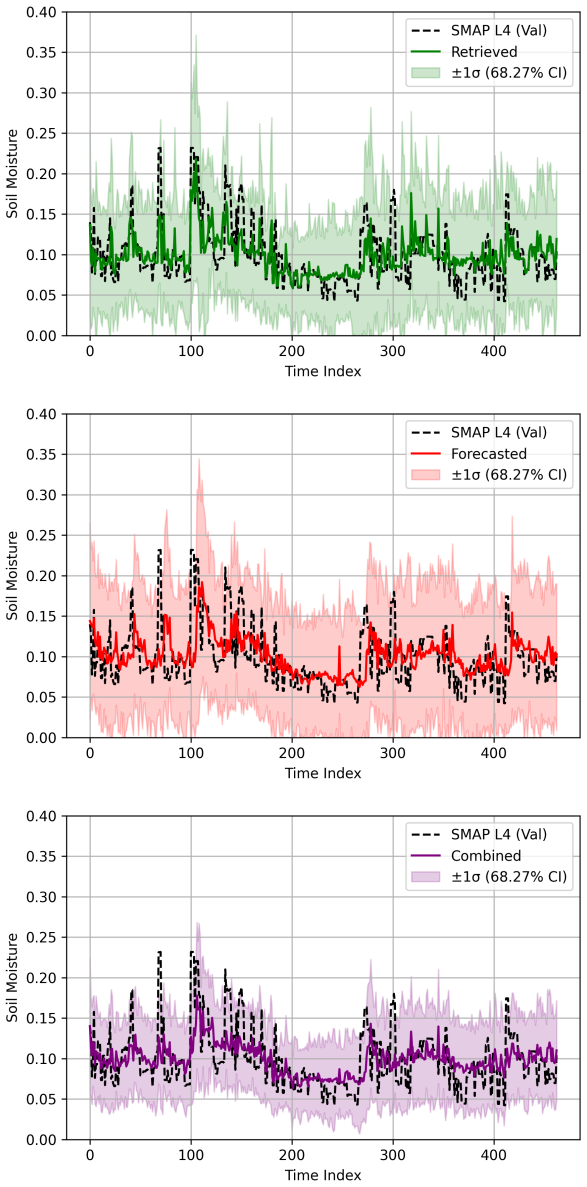

Fig. 4. SM estimation, both point estimates and uncertainty, for retrieval (top), forecasting (middle), and joint (bottom).

REFERENCES

[1] S. I. S. et al., "Investigating soil moisture–climate interactions in a changing climate: A review," *Elsevier Earth-Science Reviews*, vol. 99, no. 3-4, 2010.

[2] Z.-L. Li, P. Leng, C. Zhou, K.-S. Chen, F.-C. Zhou, and G.-F. Shang, "Soil moisture retrieval from remote sensing measure-

ments: Current knowledge and directions for the future," *Earth-Science Reviews*, vol. 218, 2021.

[3] E. B. et al., "Ground, proximal, and satellite remote sensing of soil moisture," *Reviews of Geophysics*, vol. 57, no. 2, pp. 530–616, 2019.

[4] D. E. 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, P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M.-J. Escorihuela, J. Font, N. Reul, C. Gruhier *et al.*, "The SMOS mission: New tool for monitoring key elements of the global water cycle," *Proceedings of the IEEE*, vol. 98, no. 5, 2010.

[6] C. S. Ruf, C. Chew, T. Lang, M. G. Morris, K. Nave, A. Ridley, and R. Balasubramaniam, "A New Paradigm in Earth Environmental Monitoring with the CYGNSS Small Satellite Constellation," *Scientific Reports*, vol. 8, no. 1, pp. 1–13, 12 2018. [Online]. Available: www.nature.com/scientificreports/

[7] H. Carreno-Luengo, J. A. Crespo, R. Akbar, A. Bringer, A. Warnock, M. Morris, and C. Ruf, "The CYGNSS Mission: On-Going Science Team Investigations," *Remote Sensing 2021, Vol. 13, Page 1814*, vol. 13, no. 9, p. 1814, 5 2021.

[8] G. J. De Lannoy, P. Rosnay, and R. H. Reichle, "Soil moisture data assimilation," Springer Nature, Tech. Rep., 2019.

[9] S. Tian, L. J. Renzullo, R. C. Pipunic, J. Lerat, W. Sharples, and C. Donnelly, "Satellite soil moisture data assimilation for improved operational continental water balance prediction," *Hydrology and Earth System Sciences*, vol. 25, no. 8, pp. 4567–4584, 2021.

[10] A. Gettelman, A. J. Geer, R. M. Forbes, G. R. Carmichael, G. Feingold, D. J. Posselt, G. L. Stephens, S. C. Van den Heever, A. C. Varble, and P. Zuidema, "The future of earth system prediction: Advances in model-data fusion," *Science Advances*, vol. 8, no. 14, p. eabn3488, 2022.

[11] T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler, "NGBoost: Natural gradient boosting for probabilistic prediction," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 2690–2700.

[12] T. Chen and C. Guestrin, "XGboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794.

[13] G. Tsagkatakis, A. Melebari, J. Campbell, E. Hodges, and M. Moghaddam, "Quantifying uncertainty in machine learning based soil moisture retrieval from gnss-r measurements," in *2024 International Conference on Electromagnetics in Advanced Applications (ICEAA)*. IEEE, 2024, pp. 492–492.

[14] M. Moghaddam, D. Entekhabi, Y. Goykhman, K. Li, M. Liu, A. Mahajan, A. Nayyar, D. Shuman, and D. Teneketzis, "A wireless soil moisture smart sensor web using physics-based optimal control: Concept and initial demonstrations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 3, no. 4, pp. 522–535, 2010.

[15] A. Melebari, A. R. Silva, R. Akbar, E. Hodges, Y. Zhao, P. Nergis, D. S. McKague, C. Ruf, and M. Moghaddam, "CYGNSS SoilSCAPE Sites: Sensor Calibration and Data Analysis," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 4628–4630.

[16] R. Reichle, G. De Lannoy, R. Koster, W. Crow, J. Kimball, Q. Liu, and M. Bechtold, "Smap l4 global 3-hourly 9 km ease-grid surface and root zone soil moisture analysis update, version 7," 2022. [Online]. Available: http://nsidc.org/data/SPL4SMAU/versions/7

[17] R. H. Reichle *et al.*, "Global land data assimilation system," *Journal of Hydrometeorology*, 2017.

[18] D. Entekhabi, S. Yueh, P. E. O'Neill, and K. H. Kellogg, *SMAP Handbook*. Jet Propulsion Laboratory, California Institute of Technology, 2014.

[19] R. H. Reichle *et al.*, "Assessment of smap l4," *Journal of Hydrology*, 2019.

[20] Earth Resources Observation and Science (EROS) Center, "Shuttle Radar Topography Mission 1 Arc-Second Global," 2018. [Online]. Available: https://doi.org/10.5066/F7PR7TFT