Predictive Modeling of Soil Moisture: A Review of Benchmark Datasets, Their Strengths, and Limitations

Kamrul Hasan, Arnold Muiruri
College of Computing
Grand Valley State University
Allendale, MI, USA
e-mail: {hasanka | muiruria}@mail.gvsu.edu

***Abstract*—Groundwater depletion, primarily driven by unsustainable irrigation practices in agriculture, has become a pressing global issue. Accurate soil moisture monitoring and prediction are essential for supporting sustainable water resource management. This review contributes to an ongoing research effort aimed at developing a predictive soil moisture modeling framework by integrating signals from sparsely distributed ground-based sensors with satellite-derived datasets, including NASA's Soil Moisture Active Passive (SMAP) products. As a part of this study, a case analysis involving several International Soil Moisture Network (ISMN) stations in the United States is conducted to evaluate the agreement between in-situ and satellite-derived measurements. While both data sources reveal consistent seasonal trends, significant discrepancies in magnitude highlight concerns regarding the reliability of these data as a universal benchmark. The paper provides a comprehensive review of recent advances and persistent challenges in soil moisture prediction, emphasizing the role of ISMN data. The overarching goal is to guide the development of robust, high-resolution tools for precision agriculture and sustainable groundwater management.**

Keywords-soil moisture prediction; remote sensing; international soil moisture network; data fusion; machine learning.

I. Introduction

Groundwater levels are declining at an alarming rate across the globe due to various factors, with excessive irrigation practices being one of the primary ones [1][2]. According to the 2018 U.S. Census of Agriculture, approximately 50% of the irrigated land in the United States depends exclusively on groundwater, while an additional 16% relies on a combination of groundwater and surface water. Alarmingly, nearly half of the monitoring sites across 28 U.S. states have reported significant groundwater depletion since 1980, indicating unsustainable usage patterns [3].

To address this growing crisis, it is imperative to optimize agricultural water consumption. An ongoing research project at Grand Valley State University (GVSU), Michigan, conducted under the Precision Agriculture Research Lab, aims to address this challenge. The focus of the project is on predicting soil moisture by integrating data from sparsely distributed in-situ moisture sensors with satellite-based observations, such as NASA's Soil Moisture Active Passive (SMAP) mission [4] and the European Space Agency's Climate Change Initiative (ESA CCI) [5].

Soil moisture monitoring and predictions can play a pivotal role not only in minimizing water waste but also in enabling informed decision-making for farmers and policy makers.

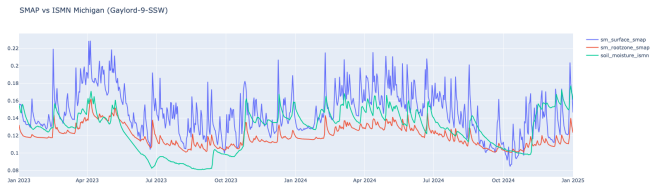

Figure 1. Average daily soil moisture by SMAP (surface-level and rootzone) vs ISMN at Gaylord-9-SSW (Michigan, U.S.). Null values were imputed through forward-fill (rolling average with window-size=3).

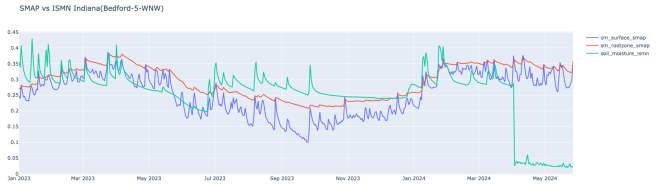

Figure 2. Average daily soil moisture by SMAP (surface-level and rootzone) vs ISMN at Bedford-5-WNW (Indiana, U.S.). Null values were imputed through forward-fill (rolling average with window-size=3).

Effective soil moisture management supports long-term soil health, prevents erosion, and ensures sustained agricultural productivity. In addition to precision agriculture, it enables better drought monitoring, flood forecasting, and land-atmosphere interaction modeling [6][7]. Although soil moisture prediction has been widely investigated, the development of consistent and reliable benchmark datasets remains an ongoing challenge. Figure 1 presents the aggregated daily average soil moisture measurements from January 2023 to January 2025 at the Gaylord-9-SSW station in Michigan, USA, an example site within the ISMN, a publicly accessible global database that consolidates in-situ soil moisture observations from numerous monitoring networks. By offering standardized data formats and automated quality control protocols, the ISMN serves a vital role in validating satellite-derived soil moisture products and land surface models, and has become a widely adopted reference in hydrological and climate research due to its comprehensiveness and accessibility [8].

To assess the consistency between ground-based and satellite-derived soil moisture measurements, we compare average daily values from NASA's SMAP products with corresponding data

from the ISMN. As illustrated in Figure 1, both datasets exhibit similar seasonal trends, with the primary differences occurring in the magnitude rather than the overall pattern. A comparable analysis at a second ISMN site, Bedford-5-WNW in Indiana, is shown in Figure 2. In this case, the discrepancy between SMAP and ISMN measurements is more pronounced than at the Gaylord-9-SSW station. These findings raise important questions regarding the reliability of these data as a benchmark for soil moisture modeling: *To what extent can ISMN be trusted for model evaluation? What are its inherent strengths and limitations? And are there viable alternatives that offer improved consistency or coverage?* This review primarily focuses on the following key aspects related to soil moisture prediction:

- Identifying the challenges involved in building accurate soil moisture prediction models.
- Examining the difficulties associated with collecting reliable data.
- Evaluating existing benchmarks for soil moisture prediction, with particular emphasis on their strengths and limitations in supporting robust model development.

The paper is organized as follows. Section II outlines advances and challenges in soil moisture prediction. Section III reviews ISMN data, emphasizing its strengths, limitations, and applications. Finally, Section IV summarizes the review with key observations and recommendations.

II. Advances and Challenges in Soil Moisture Prediction

Soil moisture prediction has evolved significantly over the past two decades, driven by advances in remote sensing, data assimilation, and machine learning. Traditional approaches primarily relied on physics-based land surface models (LSMs), such as the Noah LSM and the Community Land Model (CLM), to simulate water and energy fluxes at the land-atmosphere interface [9][10]. These models use meteorological inputs and land surface parameters, but their performance is often constrained by uncertainties in input data, parameterization, and the scale mismatches between model outputs and observational datasets [11].

Machine Learning (ML) and Deep Learning (DL) methods have recently emerged as powerful alternatives or complements to traditional models. Data-driven algorithms, including random forests, support vector machines, and artificial neural networks, have been employed to estimate soil moisture from remote sensing and meteorological data [12][13]. Deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated strong capabilities in modeling complex spatiotemporal patterns in soil moisture dynamics [14]. Additionally, hybrid approaches that integrate physical modeling with ML have gained attention for improving generalizability and interpretability [15][16].

Satellite missions such as NASA's SMAP, ESA's Soil Moisture and Ocean Salinity (SMOS), and the AMSR series have facilitated the development of predictive models at multiple spatial scales, contributing to applications from global hydrological assessment to localized precision farming [17][18]. However, most satellite-derived products are available at coarse spatial resolutions (e.g., 1 km or greater), limiting their usefulness in field-level agricultural decision-making [19].

Despite these technological advancements, several key challenges hinder the development of accurate and reliable soil moisture prediction models. A major issue is the scarcity and spatial sparsity of high-quality ground truth data, which is critical for both model training and validation [20]. The heterogeneity of environmental variables, such as soil properties, vegetation cover, land use patterns, and topography, further complicates model generalization across different regions [21]. Equally critical are the challenges associated with data collection. In-situ soil moisture measurements, such as those provided by ISMN, offer valuable ground truth but are often spatially sparse and unevenly distributed, particularly in under-monitored regions [22]. Variations in sensor type, calibration, and installation practices introduce inconsistencies, while sensor failure or communication issues can lead to temporal gaps. Satellite-based data, while offering broader coverage, are impacted by cloud cover, vegetation, and surface roughness, reducing measurement reliability in many settings [23][24]. Arid and semi-arid regions, where accurate soil moisture monitoring is most crucial, are particularly affected due to low signal-to-noise ratios [25].

Addressing these multifaceted challenges calls for multi-disciplinary strategies involving improved sensor networks, data harmonization, uncertainty quantification, and interpretable modeling frameworks. The integration of adaptive machine learning algorithms with heterogeneous data sources is critical to developing high-resolution, accurate soil moisture predictions that can transform sustainable water resource management and data-driven agriculture.

III. ISMN Data: Strengths, Challenges, and Applications

The ISMN has emerged as a critical resource for collecting and harmonizing in-situ soil moisture data across global observation networks. It serves as a foundational resource for validating, calibrating, and benchmarking satellite- and model-derived soil moisture datasets. Its importance lies in the harmonized collection and open dissemination of in-situ soil moisture data from a wide array of monitoring networks across different climate zones, land cover types, and soil structures [8][22]. The ISMN enables intercomparison of remote sensing products (e.g., SMAP, SMOS, AMSR2) by providing a global standard against which these data sources can be evaluated [20]. It also supports the assessment and development of downscaling algorithms and machine learning models by offering high-quality ground truth measurements [26]. Moreover, the temporal consistency and metadata richness of ISMN facilitate long-term hydrological studies and trend detection, which are crucial for climate resilience planning and agricultural decision-making. By improving the accuracy and robustness of predictive models, ISMN plays a critical role in the advancement of soil moisture science and its practical

applications in water resource management, agriculture, and disaster mitigation.

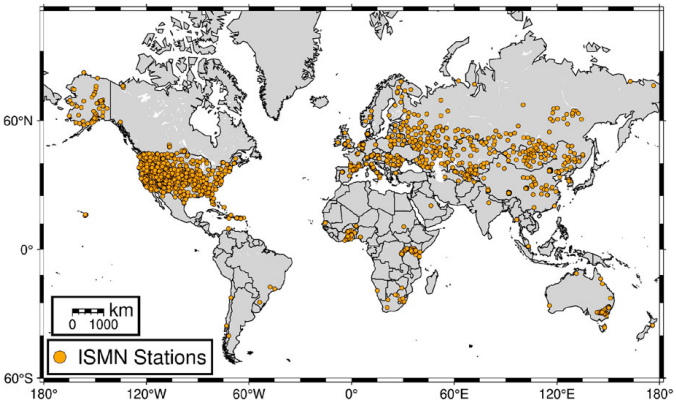

Figure 3. ISMN Stations Wold wide – an exact extract from [27].

ISMN aggregates soil moisture measurements from a variety of sources, standardizing and applying quality control procedures to improve accessibility and usability [8]. However, the ISMN data can still exhibit inconsistencies due to differences in sensor types, installation depths, and environmental heterogeneity [20]. Draper et al. emphasized the importance of preprocessing ISMN data before using it for validation or modeling tasks [28].

Despite its value, ISMN presents several challenges when used in predictive soil moisture modeling. The spatial distribution of ISMN stations, as shown in Figure 3, is highly uneven, with denser coverage in North America and Europe and sparse representation in Africa, Asia, and South America. This limits global-scale modeling and regional calibration, especially in underrepresented ecosystems. Station metadata, including soil depth, vegetation, and land use, is sometimes incomplete or inconsistent, complicating efforts to standardize data inputs for machine learning and physical models [20]. Discrepancies also arise from heterogeneity in sensor types, calibration protocols, and measurement depths across networks, introducing uncertainty into inter-station comparisons and satellite validation studies [22]. Moreover, data gaps due to sensor maintenance or environmental interference pose problems for time series continuity. These limitations necessitate pre-processing steps such as harmonization, gap-filling, and filtering, which introduce additional complexity into model development pipelines. Despite these challenges, ISMN remains a critical benchmark for validating satellite retrievals and downscaling methods, though its shortcomings highlight the importance of complementing it with other data sources and standardization frameworks.

The increasing availability of ISMN data has enabled its integration into machine learning and deep learning models for high-resolution soil moisture estimation. Xu et al. [29] used ISMN data to validate a wide and deep neural network that improved the spatial resolution of SMAP satellite data across the U.S. Similarly, Celik et al. [30] and Lee et al. [31] developed deep learning models incorporating ISMN observations to improve performance in heterogeneous landscapes by reducing dependency on physical modeling assumptions. In the agricultural domain, Custódio and Prati [32] applied ensemble machine learning models to IoT-supported irrigation systems, using soil moisture as a key variable. Their results, validated with real-time field data, support the use of AI for operational water resource management.

While the ISMN is the most prominent repository for in-situ soil moisture measurements, several alternative datasets and platforms also play crucial roles in soil moisture research and modeling. One key alternative is the USDA Soil Climate Analysis Network (SCAN), which provides high-resolution, near-real-time soil moisture data across agricultural zones in the United States [33]. Similarly, the FLUXNET network offers point-based data through eddy covariance towers, which include soil moisture as part of broader ecosystem flux measurements [34]. In terms of satellite-derived products, SMAP and ESA's SMOS missions provide global, gridded soil moisture datasets at regular intervals [35]. The Advanced Scatterometer (ASCAT) onboard EUMETSAT MetOp satellites also offers a long-term record of soil moisture estimations with relatively high temporal resolution [36]. Additionally, regional in-situ networks such as the OzNet (Australia), REMEDHUS (Spain), and ARM Southern Great Plains (USA) serve as valuable sources for local model calibration and validation. These alternatives, while often complementary to ISMN, highlight the diversity of data sources available for soil moisture modeling and reinforce the importance of integrated approaches that combine satellite, in-situ, and model-based observations.

IV. Conclusion

Accurate soil moisture prediction is vital for mitigating groundwater depletion in irrigation-dependent regions. This review highlights the potential of integrating satellite data with sparse in-situ measurements, though concerns remain regarding the consistency of benchmark datasets like ISMN. Case studies reveal seasonal alignment with SMAP, yet discrepancies in magnitude question ISMN's reliability as a ground truth. Key challenges include sparse station coverage, sensor inconsistencies, and the coarse resolution of satellite products. Moving forward, improving data quality, harmonization, and leveraging explainable AI and high-resolution models will be essential for developing robust, interpretable soil moisture prediction systems to support sustainable agricultural water management.

Future work must prioritize the refinement of benchmark datasets through enhanced quality control, data harmonization, and sensor calibration strategies. Simultaneously, advances in data fusion, explainable AI, and high-resolution modeling hold the potential to significantly improve prediction accuracy and practical utility.

References

[1] L. F. Konikow, "Groundwater depletion in the united states (1900–2008)", *US Geological Survey Scientific Investigations Report*, vol. 2013, no. 5079, p. 63, 2005.

[2] S. Foster and J. Chilton, "Groundwater: The processes and global significance of aquifer degradation", *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1440, pp. 1957–1972, 2003.

[3] J. Mercier, *Groundwater levels are dropping across the u.s.—faster than we can replenish*, Accessed: 2025-07-18, 2023.

[4] D. Entekhabi *et al.*, "The soil moisture active passive (SMAP) mission", *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010. DOI: 10.1109/JPROC.2010.2043918.

[5] W. Dorigo *et al.*, "Esa cci soil moisture for improved earth system understanding: State-of-the art and future directions", *Remote Sensing of Environment*, vol. 203, pp. 185–215, 2017.

[6] I. Eniang, A. Edet, and E. Anwan, "Importance of soil moisture content in agriculture: A review", *International Journal of Agriculture Innovations and Research*, vol. 7, no. 2, pp. 2319–1473, 2018.

[7] L. Alletto, Y. Coquet, P. Benoit, D. Heddadj, and E. Barriuso, "Soil moisture monitoring for optimizing irrigation management in a maize cropping system", *Agricultural Water Management*, vol. 148, pp. 1–13, 2015.

[8] W. A. Dorigo *et al.*, "The international Soil Moisture Network: A data hosting facility for global in situ soil moisture measurements", *Hydrology and Earth System Sciences*, vol. 15, no. 5, pp. 1675–1698, 2011. DOI: 10.5194/hess-15-1675-2011.

[9] M. Rodell *et al.*, "The global land data assimilation system", *Bulletin of the American Meteorological Society*, vol. 85, no. 3, pp. 381–394, 2004.

[10] K. W. Oleson *et al.*, "Technical description of version 4.0 of the community land model (clm)", NCAR, Tech. Rep., 2010.

[11] T. Enemark, I. Sandholt, R. Fensholt, and K. H. Jensen, "Evaluating the skill of land surface models to represent the soil moisture variability over africa", *Hydrology and Earth System Sciences*, vol. 24, no. 7, pp. 3717–3738, 2020.

[12] Y. Zhang, Z. Li, H. Shen, and E. Chen, "Machine learning for downscaling remotely sensed soil moisture using multisource data", *Remote Sensing*, vol. 10, no. 3, p. 431, 2018.

[13] D. Jeong, H. Lee, and S. Lee, "Machine learning-based soil moisture prediction using remote sensing and weather data", *Environmental Modelling & Software*, vol. 148, p. 105 248, 2022.

[14] S. Khaki *et al.*, "Cnn-rnn deep learning framework for spatiotemporal soil moisture estimation", *Water Resources Research*, vol. 56, no. 2, 2020.

[15] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven earth system science", *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[16] J. Yin, Q. Zhuang, H. Tian, and M. Pan, "A physics-informed machine learning framework for hydrological modeling: Application to the mississippi river basin", *Water Resources Research*, vol. 57, no. 8, 2021.

[17] S. Chan, P. O'Neill, E. Njoku, T. Jackson, and R. Bindlish, "Soil moisture active passive (smap) enhanced level 3 radiometer global daily 9 km ease-grid soil moisture, version 1", *NASA National Snow and Ice Data Center Distributed Active Archive Center*, 2018.

[18] M. Saberi, H. Moradkhani, W. Bardsley, and X. He, "Multiscale soil moisture modeling for agricultural drought monitoring using remote sensing and land surface modeling data", *Agricultural and Forest Meteorology*, vol. 308, p. 108 548, 2021.

[19] F. Greifeneder, L. Brocca, T. Pellarin, and W. Wagner, "Machine learning in soil moisture retrievals: Progress, challenges, and future directions", *Reviews of Geophysics*, vol. 59, no. 2, 2021.

[20] A. Gruber, W. T. Crow, W. A. Dorigo, and W. Wagner, "Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network", *Vadose Zone Journal*, vol. 12, no. 2, 2013.

[21] N. Nicolai-Shaw, L. Gudmundsson, M. Hirschi, and S. I. Seneviratne, "Improving soil moisture modeling in heterogeneous landscapes", *Hydrology and Earth System Sciences*, vol. 23, pp. 1217–1235, 2019.

[22] W. A. Dorigo *et al.*, "The international soil moisture network: Serving earth system science for over a decade", *Hydrology and Earth System Sciences*, vol. 25, no. 2, pp. 5749–5804, 2021. DOI: 10.5194/hess-25-5749-2021.

[23] S. Kim, T. J. Jackson, R. Bindlish, M. H. Cosh, and V. Lakshmi, "Validation of smos soil moisture retrievals over the continental us using ground-based observations", *Remote Sensing of Environment*, vol. 121, pp. 383–398, 2012.

[24] Y. Ma *et al.*, "Validation of smap soil moisture products with in situ observations from the international soil moisture network", *Remote Sensing*, vol. 8, no. 6, p. 524, 2016.

[25] H. Wu, X. Li, and J. Peng, "Evaluating soil moisture retrievals from amsr2 using ground-based observations over semiarid regions in china", *Remote Sensing*, vol. 9, no. 5, p. 499, 2017.

[26] Y. Liu, R. M. Parinussa, W. A. Dorigo, R. A. M. de Jeu, and W. Wagner, "A comparative study of four algorithms for downscaling remotely sensed soil moisture over the contiguous us", *Remote Sensing of Environment*, vol. 179, pp. 1–19, 2016.

[27] S. K. Do, T.-N.-D. Tran, M.-H. Le, J. Bolten, and V. Lakshmi, "A novel validation of satellite soil moisture using sm2rain-derived rainfall estimates", *Frontiers in Remote Sensing*, vol. 5, p. 1 474 088, 2024.

[28] C. S. C. Draper *et al.*, "Comparison of near-surface soil moisture from smos and ascat with in situ measurements from the u.s. scan network", *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3070–3088, 2011. DOI: 10.1016/j.rse.2011.03.026.

[29] L. Xu, X. Zhang, J. Wang, and Y. Zhu, "Downscaling SMAP soil moisture using wide and deep learning with spatial feature engineering", *Remote Sensing*, vol. 14, no. 4, p. 987, 2022. DOI: 10.3390/rs14040987.

[30] M. Celik, Y. Yilmaz, and O. Kose, "Multisource deep learning model for soil moisture prediction in diverse climates", *Environmental Modelling & Software*, vol. 155, p. 105 430, 2022.

[31] S. Lee, M. Choi, and E. Kim, "A deep learning framework for smap soil moisture retrieval without radiative transfer models", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[32] P. Custódio and A. Prati, "Adaptive ensemble models for soil moisture estimation in iot-supported irrigation", *Agricultural Water Management*, vol. 295, p. 108 579, 2024.

[33] G. L. Schaefer, M. H. Cosh, and T. J. Jackson, "Usda soil climate analysis network (scan)", *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 12, pp. 2073–2077, 2007.

[34] D. Baldocchi *et al.*, "FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities", *Bulletin of the American Meteorological Society*, vol. 82, no. 11, pp. 2415–2434, 2001. DOI: 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

[35] Y. H. Kerr *et al.*, "The SMOS mission: New tool for monitoring key elements of the global water cycle", *Proceedings of the IEEE*, vol. 98, no. 5, pp. 666–687, 2010. DOI: 10.1109/JPROC.2010.2043032.

[36] W. Wagner *et al.*, "ASCAT soil moisture product: Operational validation, hydrometeorological applications, and beyond", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 3, pp. 1452–1466, 2013. DOI: 10.1109/JSTARS.2013.2262336.