26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

How much data is enough? Evaluating training length in physics-informed LSTM for soil moisture

Chenting Jiang [a,b], **Marcus Hardie** [a,b], **Quan Bai** [c]

[a] *Tasmanian Institute of Agriculture, University of Tasmania, Private Bag 98, Hobart, Tasmania 7001, Australia,* [b] *Cooperative Research Centre for High Performance Soils, Callaghan, NSW, Australia,* [c] *School of Information and Communication Technology, University of Tasmania, Hobart, Tasmania 7001, Australia*
Email: chenting.jiang@utas.edu.au

Abstract: Accurate forecasting of soil moisture and estimation of soil water retention curves (SWRCs) are essential for irrigation scheduling and soil-water management. Recent advances in AI-based models, particularly hybrid approaches like physics-informed LSTMs (PI-LSTMs), have shown strong potential for hydrological modelling. Yet, most remain limited to controlled settings and fail to address real-world deployment needs. A key challenge is determining how much site-specific soil data users must provide to train reliable, localised models. This study evaluates how different training durations (3, 6, 9, and 12 months) affect the performance of a previously developed the LSTM-PINN model, tested across three soil types and depths. Using a consistent 1-month test window, we assessed both soil moisture forecasts and SWRC estimates. Results show that 6-9 months of training provide the most reliable forecasts, while 12 months best match lab-measured SWRCs. This range also enables retraining strategies to adapt to evolving field conditions, such as seasonal shifts or extreme rainfall events. Our findings provide actionable guidance for deploying AI-driven soil models in agricultural settings, highlighting that moderate data requirements are sufficient for both forecasting and parameter estimation.

Keywords: *Soil moisture forecasting, Soil water retention curve, Physics-informed LSTM*

1. INTRODUCTION

Modern soil moisture sensors are now widely deployed on farms, yet much of their potential remains untapped. Growers often face difficulties in interpreting data and seeing clear return on investment (Department of Agriculture, Fisheries and Forestry 2024). This creates a persistent gap between data availability and actionable decision-making. In practice, many sensor systems only display raw values, without offering predictive insights or context-specific thresholds, such as field capacity, plant available water, or soil water retention curves (SWRCs), which are critical for guiding irrigation and water management.

AI-based models have shown promise in forecasting soil moisture and estimating hydraulic parameters, However, many existing models are trained in controlled environments and lack real-world generalisability. Purely data-driven methods may achieve accuracy but often disregard physical constraints (Bandai & Ghezzehei 2021). Moreover, soil properties are shaped by local texture, structure, and management practices, making universal training across sites impractical.

To address these challenges, we build on our previously developed the LSTM-PINN model (Jiang et al. 2025), which integrates physical laws with data-driven learning. This hybrid model not only improves forecasting accuracy but also allows for site-specific estimation of SWRCs directly from sensor data. In this study, we examine how different lengths of training data affect model performance across three soil types and depths. Our goal is to identify a training window that balances accuracy, robustness, and practical retainability, ultimately enabling AI tools that transform raw soil moisture data into context-aware irrigation support.

2. METHODOLOGY AND RESULTS

The experiments were designed to evaluate the LSTM-PINN framework across three agricultural sites with both homogeneous and heterogeneous soil profiles (Table 1). This model was trained separately for each site to reflect site-specific soil properties. At each site, hourly soil moisture at the three depths, together with rainfall and evapotranspiration, were used as model inputs. To investigate the effect of training history, four scenarios were tested using 3, 6, 9, and 12 months of data for model training, with a subsequent fixed 1-month test window per site (June for S1 and S3, December for S2) reserved for performance evaluation. Testing was restricted to the same calendar month across scenarios to control for comparability, while acknowledging potential seasonal bias. The LSTM-PINN used a hybrid loss consisting of data, physical, and boundary terms with adaptive weights. This design enabled systematic assessment of the model's ability to forecast soil moisture and estimate SWRCs across varied soil conditions.

Results showed that a training length of 6-9 months generally delivered the best forecasting performance. At the 6-month setting, average R^2 values across all three sites exceeded 0.90, indicating high forecasting accuracy, while 9 months yielded the highest R^2 at S1 and S3 (Table 1). In contrast, training with only 3 months of data was insufficient to capture seasonal variability, while 12 months often led to overfitting and reduced generalisability. The reduced performance of S2 in December may reflect both sandy soil properties and seasonal effects.

For SWRCs, longer training durations improved alignment with laboratory measurements. The 12-month setting showed the closest match and outperformed estimates from Rosetta, a widely used pedotransfer function model (Figure 1). Overall, 6-9 months of data appear sufficient for reliable forecasting, while training beyond 9 months can further enhance the precision of soil hydraulic parameter estimation.

Table 1 Soil moisture forecasting accuracy (R^2) across sites, textures, depths, and months of training (3, 6, 9, 12).

Site	Soil texture	Depth (cm)	3 months	6 months	9 months	12 months
S1	Loam	-5	0.968	0.988	0.986	0.988
	Clay loam	-15	0.968	0.982	0.990	0.987
	Sandy clay loam	-25	0.990	0.972	0.982	0.965
Average R^2	—	—	**0.975**	**0.981**	**0.986**	**0.980**
S2	Sandy loam	-5	0.959	0.956	0.958	0.955
	Sandy loam	-15	0.897	0.850	0.933	0.949
	Sand	-25	0.782	0.908	0.493	0.366
Average R^2	—	—	**0.880**	**0.905**	**0.795**	**0.757**
S3	Clay	-5	0.836	0.830	0.920	0.876
	Clay	-15	0.978	0.984	0.984	0.996
	Clay	-25	0.965	0.986	0.991	0.994
Average R^2	—	—	**0.926**	**0.933**	**0.965**	**0.955**

Colour key: dark green ($R^2 \geq 0.95$) – Excellent, highly reliable forecasts; light green ($0.90 \leq R^2 < 0.95$) – good, generally reliable; yellow ($0.80 \leq R^2 < 0.90$) – moderate, use with caution; orange ($R^2 < 0.80$) – poor, often unreliable.

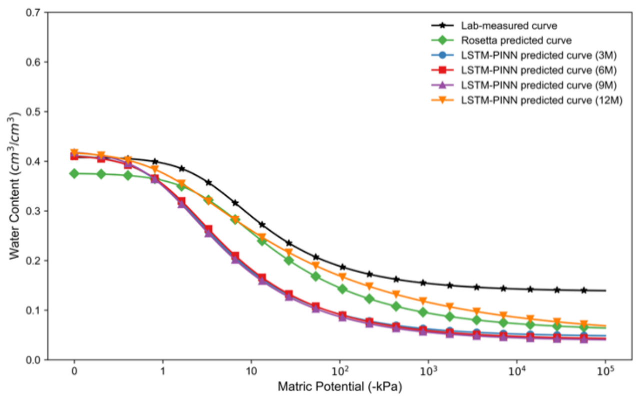

Figure 1 Water retention curves at site S1 (depth = -25 cm).

3. CONCLUSIONS

This study demonstrates a scalable strategy for leveraging raw sensor data to deliver both short-term forecasts and long-term hydraulic insights with minimal training requirements. A training window of 6-9 months was found to balance accuracy with efficiency, enabling periodic retraining to account for seasonal variation and field changes. The proposed workflow involves initial training on moderate datasets followed by incremental updates as new observations become available. This supports the generation of decision-ready outputs, including soil moisture forecasts and water retention curves. These outputs can be directly translated into irrigation plans, defining immediate actions, setting automated trigger points, and guiding monitoring intervals. By transforming raw sensor readings into adaptive, site-specific insights, this approach lays the foundation for intelligent and scalable soil moisture management at the farm scale.

ACKNOWLEDGEMENTS

This work has been supported by the Cooperative Research Centre for High-Performance Soils whose activities are funded by the Australian Government's Cooperative Research Centre Program.

REFERENCES

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

Department of Agriculture, Fisheries and Forestry (2024) *Modern Soil Moisture Monitoring for Drought Resilience*. Australian Government, Canberra

Jiang C, Bai Q and Hardie M (2025) A neighbour-aware LSTM-PINN model for physically consistent prediction of soil moisture and water retention curves. *Authorea*, 8 August. https://doi.org/10.22541/au.175466594.45672666/v1

