Recurrent Neural Networks for Soil Moisture Prediction Leveraging Soil Matric Potential Data

(Article begins on next page)

Recurrent Neural Networks for Soil Moisture Prediction Leveraging Soil Matric Potential Data

Nicola Dilillo *, Antonio C. Marceddu *, Mattia Barezzi †, Umberto Garlando †, Renato Ferrero *

* Department of Control and Computer Engineering (DAUIN)

† Department of Electronics and Telecommunications (DET)

Politecnico di Torino

Turin, Italy

Email: {nicola.dilillo, antonio.marceddu, mattia.barezzi, umberto.garlando, renato.ferrero}@polito.it

***Abstract*—Soil moisture is a parameter of paramount importance for a variety of applications, such as predicting floods and droughts, monitoring agricultural crop performance, and managing water supply. It can be measured in several ways, such as Volumetric Water Content (VWC) and soil matric potential. In this paper, an experiment was carried out by placing soil matric potential sensors at the depths of −20 cm and −40 cm within the root layer of an adult apple tree orchard to gather data every 10 minutes and to train some Recurrent Neural Network (RNN) models to predict future matric potential values at the depths of interest. Base RNN, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) networks were employed to accomplish the aforementioned goal. Additionally, feature selection analysis was used to determine the best parameters to feed the models. The trained models give an accurate short-term prediction of 10 minutes, with a R^2 of 0.9947, and a long-term prediction of 3 hours, with a R^2 of 0.7922 at −20 cm.**

***Index Terms*—Soil Matric Potential, Neural Networks, Precision Agriculture, Recurrent Neural Networks, Long Short Term Memory**

I. Introduction

Soil moisture is a critical parameter in various environmental and agricultural processes, as it is essential in agricultural monitoring, drought and flood forecasting, forest fire prediction, water supply management, and other natural resource activities. Accurate measurement, monitoring, and prediction of soil moisture is critical to understanding the complex interactions between the land surface, atmosphere, biosphere, and beyond.

In agricultural monitoring, soil moisture provides valuable insights into crop growth, development, yield, and the overall health of agricultural ecosystems [1], [2]. Farmers and agricultural researchers can thereby optimize irrigation schedules, improve water use efficiency, and increase crop resilience to adverse weather conditions.

Soil moisture is then critical for drought and flood prediction, as it can serve as an early warning indicator of impending extreme events [3], [4]. Soil moisture deficits can signal the onset of drought conditions, often preceding other more standard indicators, such as precipitation anomalies or vegetation health indices. Conversely, excess soil moisture can indicate an increased risk of flooding, enabling timely and effective flood management strategies.

In the context of forest fire prediction, soil moisture allows the determination of vegetation susceptibility to ignition and the potential spread of wildfires [5], [6]. Fire management agencies can better anticipate and mitigate wildfire risks by monitoring soil moisture levels.

In addition, soil moisture is essential for water supply management and other natural resource activities. Through it, water resource managers can make more informed decisions regarding water allocation, reservoir operation, and groundwater management. This information is also valuable for ecosystem restoration efforts, land use planning, and climate change adaptation strategies.

Depending on the measured property, soil moisture measurements can be described as matric potential (or soil water tension) and Volumetric Water Content (VWC). The former, typically expressed as a negative pressure in kPa, identifies the potential energy the plant needs to overcome to extract water from the soil using its roots. The drier the soil, the more negative this value becomes; conversely, the value approaches zero for wetter soil. The latter measures the amount of water in the soil, expressed as a percentage of the total soil volume. This measure is dimensionless and can range from 0% (completely dry soil) to 100% (soil is fully saturated) based on the porosity of the soil. In particular, matric potential can be used as an input variable to decide whether to irrigate or evaluate plant's water stress in professional apple orchards [7] in such a way as to maintain the optimal range of soil water tension in the root layer of the cultivar.

This paper aims to predict, in the short term, matric potential at −20 cm and −40 cm in order to lay the foundation for future work related to agricultural monitoring and water supply management. This will be achieved using the base architecture of Recurrent Neural Network (RNN), Gated Recurrent Unit

This work was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 - Component 2 - Investment 3.1 - Call for tender No. n. 3264 of 28/12/2021 of Italian Ministry of Research funded by the European Union – NextGenerationEU - Project code: IR0000027, Concession Decree No. 128 of 21/06/2022 adopted by the Italian Ministry of Research, CUP: B33C22000710006, Project title: iENTRANCE, and within the Agritech National Research Center receiving funding from the European Union NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

(GRU), and Long Short-Term Memory (LSTM) models and exploiting feature selection in order to obtain better prediction models.

After this introductory section, soil moisture, matric potential, and soil water tension are considered the same concept in the following part of the manuscript. The remainder of the paper is organized as follows. Section II describes the data collected in the experiment. Section III discusses the sequential neural networks adopted in the analysis. The results are presented in Section IV, and some conclusions are drawn in Section V.

II. Data Preprocessing

A. Data Source

The analyzed cultivar is situated in Verzuolo (CN), Italy (44°36′33.49″ N 7°31′17.82″ E) where a professional orchard of apple (variety Crimson Snow) is cultivated on a sandy loam soil texture. The plant is in the adult stage (the cultivar was planted in 2016).

Soil sensors are placed at −20 cm and −40 cm with respect to the ground surface. These values are used since the root layer of this cultivar is within 0-50 cm.

Soil Matric Potential and Soil Temperature are monitored at depths of −20 cm and −40 cm utilizing matric potential sensors called TEROS 21 (METER Group, Inc., Pullman, USA). These sensors are connected to a low-power, long-range IoT (Internet of Things) node able to read TEROS 21 sensors, sending its output via LoRa (Long Range) radiofrequency protocol [8]. The data usage period goes from 18 June 2023 to 18 October 2023, with data collected every 10 minutes.

In addition, meteorological data are gathered hourly in the same period as the soil sensors. These data are obtained from one of the professional weather stations provided by the Regional Environmental Protection Agency and Regional Agrometeorological Network of the Piedmont region close to the analyzed field. In particular, in the simulations, Air Temperature (°C), Air Relative Humidity (%), and Precipitation (mm) are considered.

Before proceeding to the training phase, it is essential to preprocess the data. This is crucial to improve model accuracy, ensure consistency, reduce complexity, deal with missing values, facilitate feature engineering, enhance efficiency, and improve model interpretability.

B. Timeseries Missing Values

Data have been collected directly from crops to create this dataset. This has been done by exploiting the LoRa (Long Range) radiofrequency technology and the LoRaWAN Medium Access Control (MAC) layer protocol, which is unreliable because no acknowledgment mechanisms exist. Radio interference, coverage issues, duty cycle limitations, gateway capacity, and environmental disturbances are possible reasons why an IoT node might occasionally drop some packets.

Even if information is sent every 10 minutes, in some cases, a disturbance occurs, and information is lost, generating missing values in the time series. In this scenario, the highest loss is within two missing consecutive transmissions, causing a maximum gap of 30 minutes. It is essential to address it adequately to maintain the integrity of this analysis.

Fig. 1. The correlation matrix of the available parameters. It is possible to see a high correlation between Soil Temperature at −20 cm and −40 cm. Significant correlations are also present between Soil Moisture at −20 cm and −40 cm and between Soil Moisture and Air Temperature.

There are several techniques to solve this problem. For this case, the mean imputation has been chosen as the easiest but most suitable because this dataset does not present critical missing points.

C. Feature Selection

Feature selection is an essential step in machine learning because it involves selecting the most relevant variables used to create predictive models.

As shown in Fig. 1, a very high correlation exists between the Soil Temperature −20 cm (°C) and the Soil Temperature −40 cm (°C). Apart from this couple, there is a very low correlation between all the data, besides a slightly higher correlation between Soil Moisture −20 cm (kPa) and Soil Moisture −40 cm (kPa) and between Air Temperature (°C) and Air Humidity (%). However, neither pair has a high enough correlation to be considered critical.

In light of this analysis, it was decided to proceed with multiple training iterations. Firstly, all available input variables were used, amounting to 8 features. Subsequently, the Soil Temperature at −40 cm (°C) was excluded, reducing the total number of features to 7.

D. Normalization

Several normalization techniques, such as Min-Max and Z-Score, were applied to the input data to find the one that

TABLE I
THE LAYERS OF THE NEURAL NETWORKS USED FOR THE PREDICTION OF SOIL MOISTURE VALUES.

Layer Type	Description
RNN/GRU/LSTM	Base layer with h outputs
Fully Connected (FC)	Linear layer with h inputs and outputs equal to the number of features to select

TABLE II
THE HYPERPARAMETERS USED TO TRAIN THE DIFFERENT MODELS.

Parameter	Value
Learning Rate	0.0005
Batch Size	4
Hidden Layers	64

maximized the results of the neural networks under the same initial conditions. The first, with values between 0 and 1, was the one that yielded the best performance. The results depicted in the paper were obtained using such a normalization technique.

III. Methodology

A. Neural Network Models

Three different sequential neural networks were tested: base RNN, GRU, and LSTM. Both can model sequential data for sequence recognition and prediction, maintaining the memory of previous inputs [9]. This makes these models the perfect candidate to work with time series.
Even if base RNNs, GRUs, and LSTMs are used for the same purpose, they differ in architecture and efficiency. GRUs have gating mechanisms and are more parameter-efficient, offering a sophisticated approach to modeling temporal dynamics and dependencies. LSTMs, instead, utilize an additional memory cell and more complex gating mechanisms, including input, output, and forget gates, which allow them to capture long-term dependencies better and mitigate the vanishing gradient problem more effectively.
The models adopted are rather simple and are summarized in Table I.

B. Hyperparameter Settings

Different combinations of hyperparameters were considered during the analysis. The table II summarizes the best values found, which were used for subsequent training of the neural networks. During training, the learning rate was subject to adjustment through a callback function that halves its value if the validation loss does not improve for three consecutive training epochs. The Early Stopping callback was also adopted to avoid overfitting and improve training time. It was calibrated to stop the training process when the performance on a validation set did not improve after 5 epochs.

C. Prediction

Models were trained by combining different length input sequences and several future point steps ahead to predict, using the following values:

- Input sequence lengths of 6 and 12, corresponding to 60 and 120 minutes, respectively.
- Future points with steps 1, 6, 12, and 18, corresponding to 10, 60, 120, and 180 minutes.

For this work, all the models used have been set to predict values of Soil Moisture that are at the maximum −0.4 kPa, the highest values recorded in the dataset.

D. Evaluation Metric

The following metrics were used to evaluate the performance of the models thoroughly:

- The Mean Square Error (MSE) and the related Root Mean Square Error (RMSE) tend to approach 0 in the case where the obtained network models the problem optimally. They are computed as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad RMSE = \sqrt{MSE} \quad (1)$$

- The coefficient of determination (R^2), on the contrary, tends to approach 1 when the obtained network models the problem optimally. It is computed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \quad (2)$$

E. Model Settings

Several networks were trained with the goal of searching for the best model able to describe the problem. At this scope, the dataset was first split into 80% for training and 20% for testing. A 5-fold cross-validation was then applied to the first dataset to obtain a variable validation and training dataset in order to seek a quantization of both that would maximize the results. As anticipated in Section II-C, three different trainings were performed:

- The first uses all the available features, which is recalled as 8.
- The second avoids the usage of Soil Temperature at −40 cm, for a total of 7 features.

The next section will highlight the results obtained regarding the prediction of Soil Matric Potential at −20 cm and −40 cm.

IV. Results

The consolidated findings are presented in Table IV, which details the diverse combinations of the input sequence number, future output points, and the count of employed features. Each model, represented by a row in the table, shows that the optimal Root Mean Square Error (RMSE) corresponds with the highest R^2 score. Consequently, this analysis will focus solely on the R^2 metric for evaluation.
A clear pattern emerges from the data: precision decreases as the prediction horizon lengthens while increasing the number of input sequences has no significant impact. This pattern emphasizes the ability of the model to accurately predict data in the near future. In terms of performance across different depths, RNNs excel in the −20 cm depth predictions, while

TABLE III
BEST VALUE PREDICTIONS FOR BASE RNN, GRU, AND LSTM MODELS USING DIFFERENT LENGHT OF INPUT SEQUENCE (SEQ IN) AND FUTURE POINT OUTPUT TO PREDICT (OUT).

			−20 cm Depth						**−40 cm Depth**					
			RMSE			**R^2**			**RMSE**			**R^2**		
Seq IN	OUT	# Features	RNN	GRU	LSTM	RNN	GRU	LSTM	RNN	GRU	LSTM	RNN	GRU	LSTM
6	1	8	**0.0068**	0.0071	0.0073	**0.9937**	0.993	0.9926	0.006	0.006	0.0058	0.9867	0.9864	0.9876
6	1	7	0.007	0.007	0.0073	0.9933	0.9932	0.9927	0.0058	**0.0052**	0.0058	0.9873	**0.9899**	0.9876
6	6	8	**0.0186**	0.0195	0.0193	**0.952**	0.9474	0.9482	0.017	0.0162	0.018	0.8912	0.9006	0.8779
6	6	7	0.0189	0.0194	0.0194	0.9505	0.9477	0.9478	0.0172	**0.0161**	0.0171	0.8882	**0.9022**	0.8887
6	12	8	**0.0305**	0.0322	0.0323	**0.8701**	0.8554	0.8546	**0.0215**	0.0228	0.0226	**0.8223**	0.8007	0.8043
6	12	7	0.0308	0.0343	0.0322	0.8679	0.8359	0.8559	0.0232	0.0226	0.0245	0.7937	0.8038	0.7697
6	18	8	**0.0405**	0.043	0.0419	**0.7689**	0.7388	0.7519	0.0299	0.0293	0.0333	0.649	0.6634	0.5668
6	18	7	0.0429	0.0413	0.0418	0.7404	0.7587	0.753	0.0445	0.0287	**0.0279**	0.2262	0.6783	**0.696**
12	1	8	0.007	0.0067	0.0066	0.9932	0.9937	0.994	0.0056	0.0059	0.0057	0.9881	0.987	0.9877
12	1	7	0.0066	0.0063	**0.0062**	0.994	0.9945	**0.9947**	0.0057	**0.0055**	0.0057	0.9875	**0.9884**	0.9876
12	6	8	0.0187	**0.0186**	0.0187	0.9512	**0.9518**	0.9512	0.0178	0.0161	0.0171	0.8793	0.9012	0.8887
12	6	7	0.0191	0.02	0.0204	0.949	0.944	0.9417	**0.0155**	0.0158	0.0182	**0.9079**	0.9047	0.8733
12	12	8	0.0317	**0.0298**	0.0311	0.8579	**0.8744**	0.8635	0.0265	0.0227	0.0247	0.7254	0.7981	0.7611
12	12	7	0.0311	0.0311	0.0319	0.8634	0.8634	0.856	0.0253	**0.0224**	0.0247	0.7493	**0.8028**	0.7605
12	18	8	0.0418	0.0419	0.0414	0.7522	0.7509	0.7563	0.0321	0.0286	**0.0273**	0.5908	0.6758	**0.7047**
12	18	7	**0.0382**	0.0414	0.0413	**0.7922**	0.7571	0.7571	0.0289	0.0302	0.0298	0.6705	0.6384	0.6486

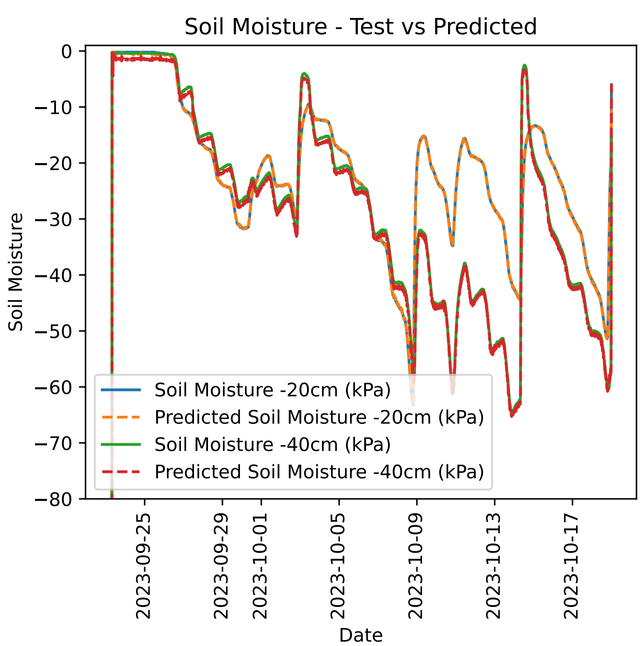

Fig. 2. Comparison of actual and 10-minute predictions obtained using the best model among those tested for the 10-minute predictions.

GRUs show better results at −40 cm. LSTMs prove to be the best option in a limited number of scenarios, chiefly at greater distances.
For the −20 cm depth, the highest R^2 scores span from 0.9947, in the case of short-term predictions, to 0.7922 for projections further out. At the −40 cm depth, the range starts from a similar R^2 of 0.9899 to 0.7047, indicating a near 10% discrepancy. Generally, the accuracy of predictions at the −40 cm depth is lower. In Fig. 2, the prediction produced by the best model at 10-minute, based on a sequence of input length 12.
Feature reduction has minimal impact on the accuracy of predictions at −20 cm using input sequences of 6. However, slight variations are observed using 7 features, yielding the best results in most cases. However, for predictions made with a 12 input sequence and an 18 output prediction, employing feature reduction results in a 4% improvement.

V. CONCLUSION

This study demonstrates the effectiveness of employing models like base RNN, GRU, and LSTM to generate precise predictions from soil potential data. Despite the advanced mechanisms of GRU and LSTM models, our findings indicate that RNN models can also be a viable option for accurate soil moisture prediction in most cases.
Even if a significant correlation exists between two variables, the simultaneous or singular presence of these variables does not substantially alter the accuracy of the results obtained.
A reliable prediction for soil potential data can be exploited in future works for water supply management and the optimization of irrigation schedules. With a forecast of the soil water potential it is possible to stop irrigation earlier reducing the effects of over-irrigation.

REFERENCES

[1] D. N. Gupta, V. Veeraiah, H. Singh, R. Anand, N. Sindhwani, and A. Gupta, "Iot-dependent intelligent irrigation system with ml-dependent soil moisture prediction," in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 2023, pp. 1296–1300.

[2] S. Chatterjee, N. Dey, and S. Sen, "Soil moisture quantity prediction using optimized neural supported model for sustainable agricultural applications," *Sustainable Computing: Informatics and Systems*, vol. 28, p. 100279, 2020.

[3] N. Wanders, D. Karssenberg, A. de Roo, S. M. de Jong, and M. F. P. Bierkens, "The suitability of remotely sensed soil moisture for improving operational flood forecasting," *Hydrology and Earth System Sciences*, vol. 18, no. 6, pp. 2343–2357, 2014.

[4] C. Massari, S. Camici, L. Ciabatta, and L. Brocca, "Exploiting satellite-based surface soil moisture for flood forecasting in the mediterranean area: State update versus rainfall correction," *Remote Sensing*, vol. 10, no. 2, 2018.

[5] D. Chaparro, M. Vall-llossera, M. Piles, A. Camps, C. Rüdiger, and R. Riera-Tatché, "Predicting the extent of wildfires using remotely sensed soil moisture and temperature trends," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 6, pp. 2818–2829, 2016.

[6] Vinodkumar and I. Dharssi, "Evaluation and calibration of a high-resolution soil moisture product for wildfire prediction and management," *Agricultural and Forest Meteorology*, vol. 264, pp. 27–39, 2019.

[7] A. Ben Abdelkader, M. Thalheimer, M. Petrillo, M. Tagliavini, and D. Zanotelli, "Automatic irrigation scheduling based on soil water potential thresholds allows water savings in a highly productive apple orchard," *Acta Horticulturae*, vol. 1346, pp. 563–570, 2022. [Online]. Available: https://doi.org/10.17660/ActaHortic.2022.1346.71

[8] M. Barezzi, U. Garlando, F. Pettiti, L. Nari, D. Gisolo, D. Canone, and D. Demarchi, "Long-range low-power soil water content monitoring system for precision agriculture," in *2022 IEEE 13th Latin America Symposium on Circuits and System (LASCAS)*, 2022, pp. 1–4.

[9] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.