Forecasting decreasing soil moisture without environmental data

Aleix Armero[1]

Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain
aleix.armero@estudiantat.upc.edu

Abstract. This work presents two models for forecasting soil moisture decreases that operate without auxiliary environmental data such as temperature or humidity. Both models rely solely on historical soil moisture measurements and their timestamps to predict future decreases over a chosen horizon. The first model combines an XGBoost classifier, which predicts when the soil moisture decrease stops, with a RANSAC linear regressor to estimate subsequent moisture values based on the observed measurements. The second model is a single XGBoost multi-output regressor that, given the most recent soil moisture measurement and its position relative to the first measurement, directly forecasts the next three soil moisture values. The first model reaches a maximum loss of 0.012 after 20 steps, while the second model achieves a slightly lower maximum loss of 0.01. These results demonstrate that soil moisture dynamics can be effectively predicted even with limited input data using classical modeling approaches.

Keywords: Forecasting · Environment prediction · Soil moisture dynamics

1 Introduction

Soil moisture forecasting plays a key role in precision agriculture, as it enables farmers to monitor crop conditions and optimize water management to maintain high productivity. Among the various environmental variables, soil moisture is particularly important because it directly reflects the amount of water available to plants. Although physical models can predict soil moisture decay with high accuracy [1, 2], they require the estimation of several site-specific parameters, which may vary significantly across locations. Moreover, these models are primarily designed to describe moisture decay following rainfall events rather than regular irrigation, leading to discrepancies when applied to controlled watering scenarios. These discrepancies are further amplified in crop fields, where physical models typically do not account for the high density of plants and their associated water uptake.

This work addresses these limitations by proposing simple, easily trainable models that rely solely on historical soil moisture measurements and their corresponding timestamps, without requiring additional environmental data. Since irrigation events are known, the models focus exclusively on predicting soil moisture decay, as the increases are already controlled. However, the challenge of site-specific variability remains, because different environments exhibit distinct soil moisture dynamics.

2 Methods

2.1 Data

The dataset used in this study was provided by SAF Sampling, a company specializing in monitoring and providing insights on agricultural data. Multiple datasets were made available, but since soil moisture decay is highly site-specific, only the largest dataset was used. This dataset contains a historical record of 45,282 soil moisture measurements with corresponding timestamps, all collected from a single sensor over a period of two years (2023–2025), with readings taken every 30 minutes, ensuring consistency throughout the data. These soil moisture measurements are given as a percentage between 0 and 1.

Data collected up to and including 2024 were used for model training, while data from 2025 were reserved exclusively for testing and performance evaluation.

2.2 Model 1 : Linear model

The linear model consists of two components: an XGBoost predictor, which determines whether a given soil moisture and time corresponds to a plateau in the decreasing trend—i.e., whether the soil moisture will stop decreasing after that point—and a RANSAC [3] linear regressor, which forecasts future soil moisture values based on previously observed measurements by fitting a line through them while discarding potential outliers arising from sensor noise.

This model was motivated by the observation that soil moisture exhibits a recurring daily pattern following irrigation: an initial decay phase followed by a plateau during the night. This behavior is illustrated in Fig. 1. The decay phase appears approximately linear, except when it transitions into a plateau or resumes decreasing. Based on this pattern, we designed a model combining a linear regressor to capture the decay and a classifier to determine the transition from decay to plateau.

To combine the two models, we first fit the linear regressor to the previously observed soil moisture values, obtaining a linear trend that serves as the baseline forecast. Then, for each predicted time step, the XGBoost classifier is applied to determine whether the current point corresponds to a plateau phase. If a plateau is detected, the predicted value at that time step is set equal to the previous prediction, thereby enforcing a constant soil moisture level. Subsequently, the remaining part of the forecasted linear trend is shifted forward by one time step to preserve continuity and maintain the original decay trajectory after the plateau.

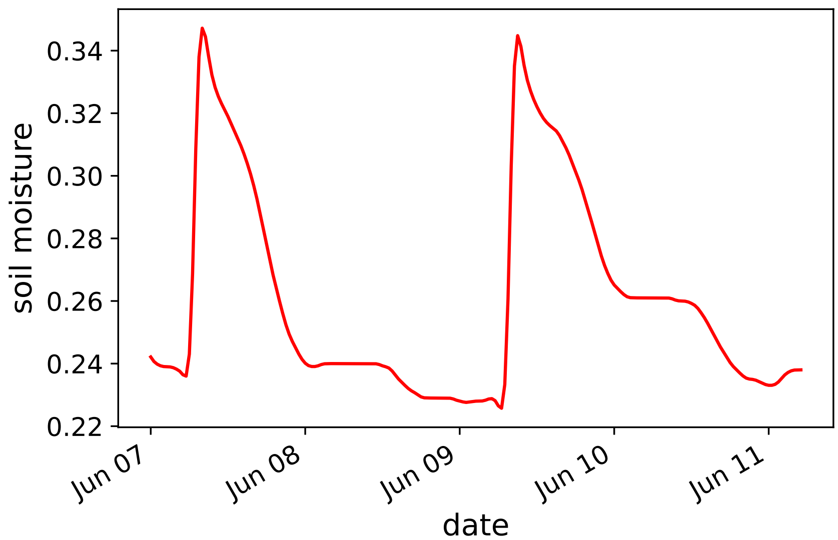


Fig. 1: Evolution of moisture decay between irrigations

2.3 Model 2 : ML-based model

The ML-based model consists of a single XGBoost multi-output regressor. Its inputs are the current soil moisture value, the time at which the measurement was taken, and the number of time steps elapsed since the first considered measurement. The model outputs forecasts of soil moisture for the subsequent three time steps. To obtain predictions over longer horizons, the model is applied recursively, using its own predictions as inputs to generate forecasts for the desired number of future time steps.

The XGBoost regressor was trained exclusively on soil moisture time series corresponding to decay phases, while periods of increasing soil moisture caused by irrigation or rainfall were excluded from the training data.

3 Results

Figure 2 shows that the linear model achieves a smaller error than the ML-based model at every forecast step. The performance gap increases as the number of forecast steps grows.

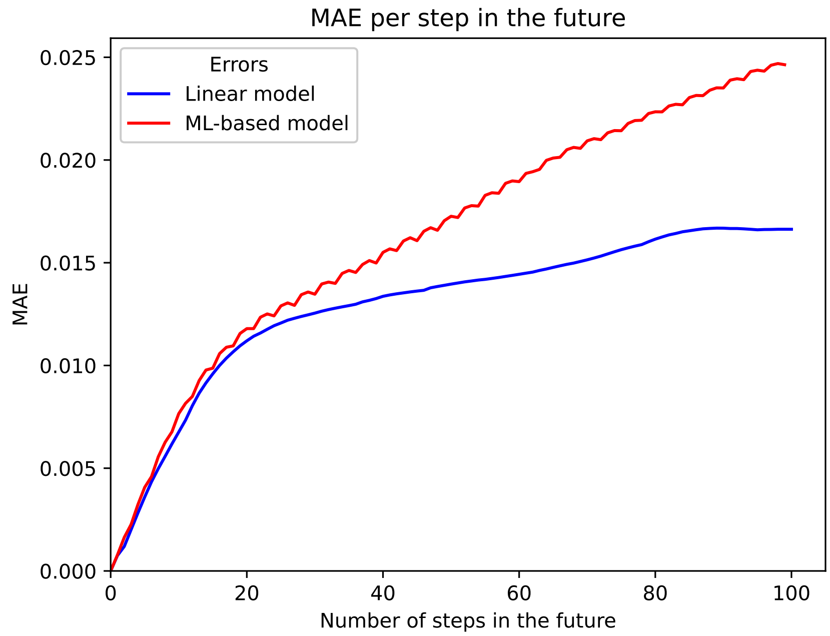


Fig. 2: Mean absolute errors of models for forecasts with different numbers of steps. Each step represents 30 minutes.

This behavior may be attributed to the winter portion of the dataset, during which irrigation events are scarce. As a result, the ML-based model lacks sufficient training examples to accurately predict soil moisture under these conditions. As shown in Fig. 3, after removing the winter data, both models exhibit comparable prediction errors. In this setting, the linear model shows higher error at shorter forecast horizons and lower error at longer horizons.

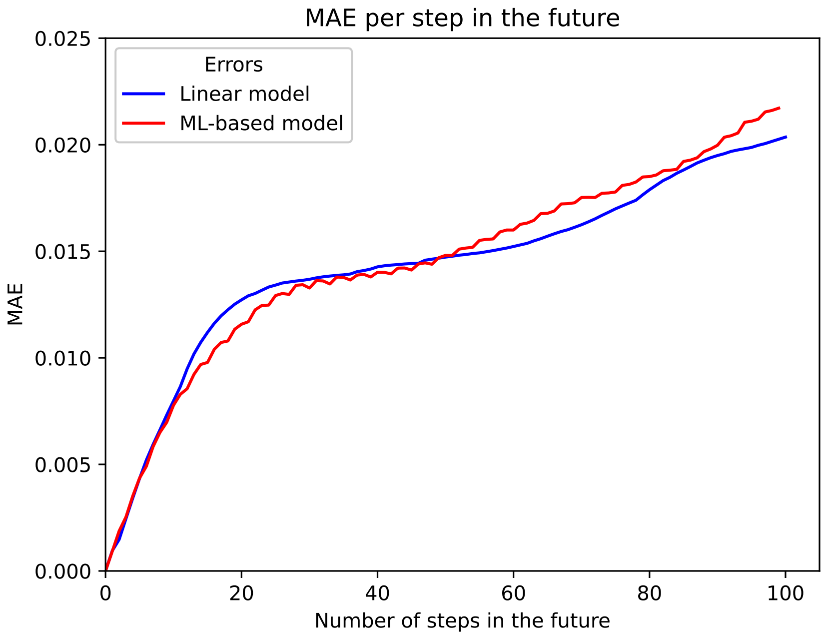
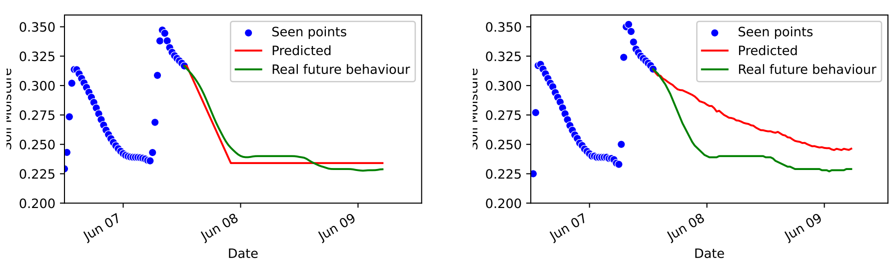


Fig. 3: Mean absolute errors of models forecasting depending on number of forecasted steps avoiding winter season

Additionally, the linear model repeatedly under- and overshoots plateau regions. In contrast, the ML-based model produces smoother forecast trajectories but fails to adequately capture the transitions between decay phases and plateau regimes. An illustrative example of this behavior is shown in Fig. 4.



(a) Linear model forecast

(b) ML based model forecast

Fig. 4: Models prediction example

4 Discussion

The results indicate that the linear model combining RANSAC regression with plateau detection consistently outperforms the ML-based XGBoost regressor in capturing the characteristic decay and plateau phases of soil moisture following irrigation at longer forecast horizons. However, this advantage does not hold for shorter forecast horizons, where both models exhibit similar performance, with the ML-based approach performing slightly better. Nonetheless, during periods with scarce irrigation, such as the winter season, the linear model appears substantially more robust than the ML-based approach, likely due to the limited availability of training examples for such conditions.

Despite its overall performance, the linear model also exhibits certain limitations. In particular, it cannot capture the smooth transition from the decay to the plateau phase, which often leads to predictions that correctly identify the timing of the plateau but misestimate the corresponding soil moisture level. This limitation could be addressed in future work by smoothing the predicted values upon detecting a plateau, which may improve the accuracy of plateau predictions and reduce the overall forecasting error. A similar approach could be applied to the transition from plateau back to decay, thereby better reflecting the continuous transitions observed in the data.

5 Conclusions

We presented and evaluated two models for soil moisture forecasting: an ML-based XGBoost regressor and a linear model combining RANSAC regression with a plateau detection classifier. The results show that both models exhibit similar performance under normal conditions, whereas the linear model substantially outperforms the ML-based approach during periods with scarce irrigation.

References

1. C. Albergel, C. Rüdiger, T. Pellarin, J.-C. Calvet, N. Fritz, F. Froissard, D. Suquia, A. Petitpa, B. Piguet, and E. Martin. From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations. *Hydrology and Earth System Sciences*, 12(6):1323–1337, December 2008. Publisher: Copernicus GmbH.
2. Mengyi Gong, Jessica Davies, Rebecca Killick, Christopher Nemeth, Shangshi Liu, and John N. Quinton. A changepoint approach to automated estimation of soil moisture drydown parameters from time series data. *Scientific Reports*, 15(1):42970, December 2025. Publisher: Nature Publishing Group.
3. Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.