Trade-offs between data-driven and process-based approaches for root-zone soil moisture retrieval in a Mediterranean vineyard

Pere Quintana-Seguí[1], Judith Cid-Giménez[1,2], Anaïs Barella-Ortiz[1], María José Escorihuela[2]

[1] Observatori de l'Ebre (URL - CSIC), Roquetes, Spain (pquintana@gencat.cat)
[2] isardSAT, Barcelona, Spain

EGU26-21629

Introduction

- Root-zone soil moisture (RZSM) data is central to drought assessment and irrigation management.
- Sensors are sinstalled, but time series are often too short for drought monitoring.
- Sensors may fail.

Goal

Evaluate the trade-offs between data-driven and process-based approaches for RZSM retrieval in a Mediterranean vineyard context, with and without SSM assimilation.

- Models allow us to:
  - Reconstruct long time series leveraging existing meteorological data.
  - Predict RZSM when sensors fail.
  - Predict future RZSM for irrigation scheduling.
  - extrapolate to non-instrumented vineyards (see Judith Cid's poster, EGU26-21620), and
  - leverage satellite surface soil moisture (SSM) products (data assimilation).
  - ...
- Until now we have used a classical approach based on a FAO56 inspired model.

Area of Study

Terra Alta vineyard region (Catalonia, NE Spain)

- Inland Mediterranean area with continental influence.
- Mean annual precipitation: ∼464 mm.
- Traditionally vineyards are rainfed, but irrigation is increasing thanks to access to surface water from the Ebro river.

Our work in the area

- We are trying to support low tech, small farmers.
- We publish a weekly bulletin with easy to understand drought indicators and support irrigation recommendations (Life eCOadapt50 project).
- Monitored vineyards are irrigated and non-irrigated.

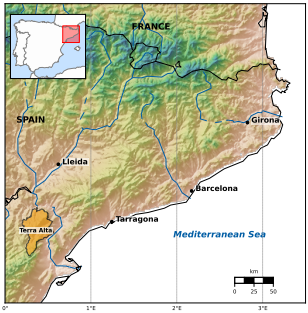
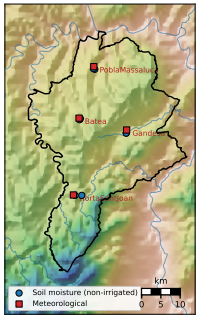


Figure: Location of the Terra Alta region within Catalonia.

Data

- **Meteorological data**: Already existing stations with historical data (SMC and AEMET).
- **Soil moisture**: METER Teros-10 probes at **5, 10, 25, 50, 70 cm**.
- **Observed SSM**: Sensors at 5 cm.
- **Observed RZSM**: Weighted mean of deep sensors (25–70 cm), representing the 15–75 cm interval.
- **Sites**: 4 non-irrigated sites (we also have irrigated sites, but they are not used in this study, for now).



Location of sites.

Instrumented field.

Methodology Overview

(1) Process based model

Based on **FAO-56 dual crop coefficient** [2] and a **multi-layer soil (DC11L)** with optional **EnKF assimilation of SSM** [4].

(2) Machine-learning (ML)

Based on **multilayer perceptron (MLP)** [6, 7] and **XGBoost** [3, 1], ingesting SSM (ML-SSM) and not ingesting SSM (ML-NOSSM).

Calibration: The models are calibrated/trained for each individual station (1D experiments)
Metric: Non-parametric KGE (KGEnp) [5].
Validation: Temporal Leave-one-out cross-validation (LOOCV).

Experimental Configurations

(3) Is precipitation data necessary?

SSM should contain enough information on actually infiltrating water.

We evaluate two scenarios to test model accuracy and robustness:

- **BASELINE**: Full meteorological forcing (P, T, ET_0) + SSM.
- **NO-P**: Precipitation is suppressed (0 for physical, removed for ML).

Approach	ID	Description
Machine Learning	**MLP / XGB**	Baseline with P, T, ET_0
	+SSM	Antecedent Surface Soil Moisture ingestion
Process based Model	**DC11L**	Dual Crop 11-Layer (Open Loop)
	ENKF	EnKF: Perturbing both P and K_{cb}

Results: Model Performance (KGEnp)

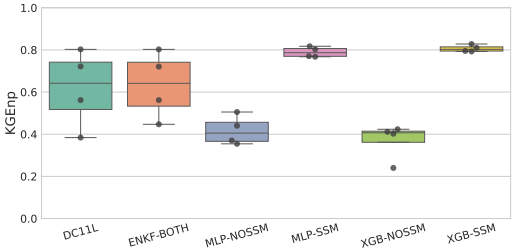


Figure: KGEnp for Baseline experiment (all stations).

- **ML-SSM dominance**: MLP and XGBoost with SSM reach KGEnp $\sim$0.8.
- **Process Based model**: better than **ML** when no **SSM** data is available.
- **EnKF**: marginal improvement.
- **XGBoost vs MLP**: XGBoost (XGB-SSM) shows slightly higher stability.
- **Consistency**: Models show robust performance across all selected rainfed sites.

Retrieval without Precipitation

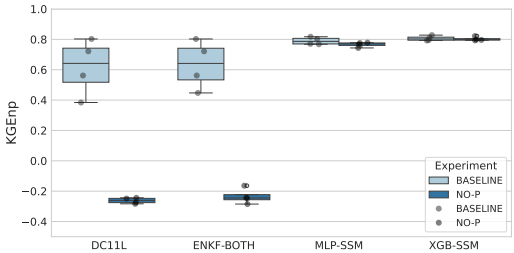


Figure: Comparison BASELINE (with P) vs NO-P (no P forcing).

- **Physical models**
  - Collapse without P forcing (KGE < 0).
  - The **EnKF** is not able to compensate that we set precipitation to zero.
- **ML models**
  - Nearly immune to missing P.
  - They are more affected by missing SSM than missing P!

Key Insight

The SSM signal itself carries enough information about infiltration events, but the physical model is not able to exploit this fact.

Conclusion

- **Data-driven superiority**: ML models with SSM ingestion reach excellent performance (KGEnp ~ 0.8), outperforming the process-based EnKF (~ 0.65).
- **ML models don't need precipitation forcing**, making it ideal for satellite-only retrieval in non instrumented plots.
- **The physical model is better without SSM** (but SSM is needed to calibrate it!)
- **XGBoost**: Shows slightly more consistency across stations than **MLP** (and it has other practictal advantages)
- **It is easier to calibrate a ML model**: Calibrating the physical model is not so easy, for example, it is difficult to get good wilting point and field capacity values from observations or soil analysis.

Main consequence

We will probably shift to an **XGBoost** model in our operational context

Perspectives

- Substitute *in-situ* SSM with downscaled satellite SSM.
- Adapt the irrigation recommendations algorithm to the ML based model.
- Include physical constraints to the loss function of the ML model.

Judith Cid PhD Thesis (EGU26-21620)

- Regionalize model parameters to evaluate performance on non-instrumented vineyards.
- Include satellite SSM on the regionalized model.
- Make irrigation recommendations for any vineyard in the region.
- Overcome the lack of training data by training on LSM data (complex physical model).

Acknowledgments & Download

This work was supported by the **DTE Hydrology Next** project (ESA) and the **LIFE21-IPC-ES- LIFE eCOadapt50** project (EU).

Download presentation:

`http://pere.quintanasegui.com/coses/quintana-EGU26-presentation.pdf`

References I

[1] A. M. Ågren, J. Larson, S. S. Paul, H. Laudon, and W. Lidberg.

Use of multiple lidar-derived digital terrain indices and machine learning for high-resolution national-scale soil moisture mapping of the swedish forest landscape.

Geoderma, 404:115280, 2021.

[2] R. G. Allen, L. S. Pereira, D. Raes, M. Smith, et al.

Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56, volume 300.

FAO Rome, 1998.

[3] 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*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[4] G. Evensen.

The ensemble kalman filter: theoretical formulation and practical implementation.

Ocean dynamics, 53(4):343–367, 2003.

References II

[5] S. Pool, M. Vis, and J. Seibert.
Evaluating model performance: towards a non-parametric variant of the kling-gupta efficiency.
Hydrological Sciences Journal, 63(13-14):1941–1953, 2018.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams.
Learning representations by back-propagating errors.
Nature, 323(6088):533–536, 1986.

[7] G. Satalino, F. Mattia, M. W. Davidson, T. Le Toan, G. Pasquariello, and M. Borgeaud.
On current limits of soil moisture retrieval from ers-sar data.
IEEE Transactions on Geoscience and Remote Sensing, 40(11):2438–2447, 2002.

Supplement: Data Availability

Station	Meteo	Soil Moisture	N_Days (SM)
GA1	2011–2025	2019–2025	1563
HA1	1998–2025	2019–2025	1939
HA2	1998–2025	2019–2025	2127
PM2	2009–2025	2019–2025	2318

Table: Data ranges and common days for comparison.

For this work we will only use data from rainfed fields.

- **Meteo**: Long-term historical data available for all sites (15+ years).
- **Soil Moisture**: Generally 6 years of data.
- **ML requirement**: Consistent training history is key for robust LOOCV performance.

Supplement: Methodology

Machine Learning

The **machine learning approach** uses **MLP** and **XGBoost**.

- **Inputs**: 14–21 day window of dynamic drivers: P, $T_{min/max/avg}$, ET_0 (+ SSM).
- **Multi-target**: Simultaneous prediction of surface (5 cm) and root-zone moisture.
- **No leakage**: Inputs are strictly antecedent ($t - W$ to $t - 1$) for predicting day t.

Process based approach

EnKF that assimilates SSM into a multilayer physical model (Dual crop FAO-56)

- **EnKF strategy**: Ensemble layer states are forced daily; SSM updates use the Kalman gain.
- **Vertical coupling**: Surface–root zone covariance propagates info downward.
- **Variants**: perturbation of P (Γ additive), K_c (Log-N mult.).

Supplementary: MLP Technical Details

Architecture

- **Type**: Feed-forward MLP (Numpy).
- **Layers**: 2 hidden layers.
- **Activation**: GeLU (hidden), Linear (out).
- **Optimizer**: Adam (Learning rate = 10^{-3}).
- **Standardization**: Z-score (per feature).
- **Targets**: Multi-target (5cm + RZSM).

Hyperparameter	Search Grid
Window Size (W)	{14, 21} days
Hidden Sizes	{(64, 32), (128, 64)}
L2 Regularization	$\{10^{-4}, 10^{-3}\}$
Max Epochs	200
Batch Size	256
Early Stopping	30 epochs (patience)

Table: MLP Tuning Search Space.

Supplementary: XGBoost Technical Details

Model Specs

- **Algorithm**: eXtreme Gradient Boosting.
- **Objective**: reg:squarederror.
- **Learning Rate**: 0.1 (η).
- **Subsample**: 0.8 (rows).
- **Colsample**: 0.8 (features).
- **Targets**: Independent model per layer.

Hyperparameter	**Search Grid**
Window Size (W)	{14, 21} days
Max Depth	{4, 6}
N Estimators	{100, 200}
Random State	123
Tree method	auto

Table: XGBoost Tuning Search Space.

Supplementary: FAO-56 & EnKF Details

Physical Model (DC11L)

- **Discretization**: 11 layers.
- **Calibrated**: θ_{FC}, CN, K_{drain}.
- **Fixed**: θ_{WP} (25th perc. min).
- **Fixed**: Z (150 cm).

EnKF Configuration

- **Ensemble size**: 32.
- **State**: Daily layer SM.
- **Update**: Daily SSM obs.

EnKF Parameter	**Search Grid**
Obs Error (R)	$\{4, 8\} \times 10^{-4}$
P Perturb. (σ_P)	$\{0.3, 0.7\}$ (Addit. Gauss)
K_{cb} Pert. (σ_K)	$\{0.03, 0.08\}$ (Mult. Log-N)
Spin-up Strategy	2 cycles (antecedent yr)

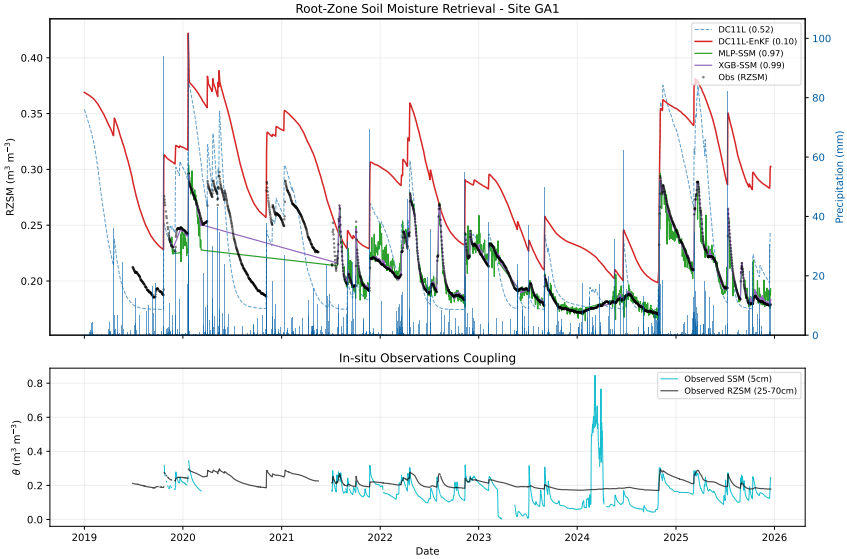
Table: EnKF Tuning Search Space.

Supplement: Detailed Results (KGEnp)

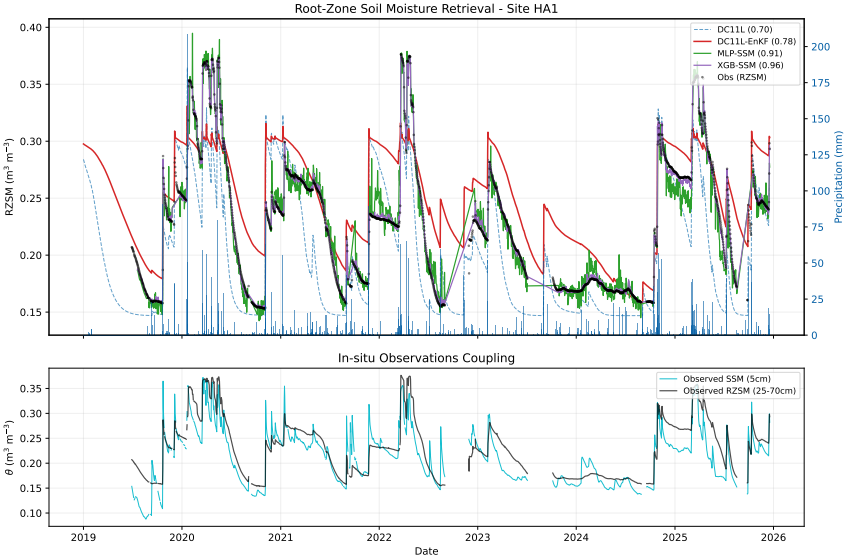
Code	Exp.	N	DC11L	ENKF	MLP-N	MLP-S	XGB-N	XGB-S
GA1	BASE	1563	0.384	0.447	0.354	0.768	0.240	0.793
GA1	NO-P	1528	-0.283	-0.285	0.230	0.743	0.250	0.796
HA1	BASE	1939	0.722	0.721	0.370	0.803	0.424	0.828
HA1	NO-P	1939	-0.249	-0.247	0.363	0.778	0.349	0.822
HA2	BASE	2099	0.562	0.562	0.505	0.770	0.411	0.795
HA2	NO-P	2127	-0.244	-0.243	0.405	0.766	0.345	0.793
PM2	BASE	2318	0.803	0.803	0.440	0.817	0.402	0.810
PM2	NO-P	2318	-0.273	-0.164	0.403	0.776	0.386	0.799

Table: Detailed KGEnp results (3 decimals). MLP-N/S: MLP-NOSSM/SSM. XGB-N/S: XGB-NOSSM/SSM. ENKF: ENKF-BOTH.

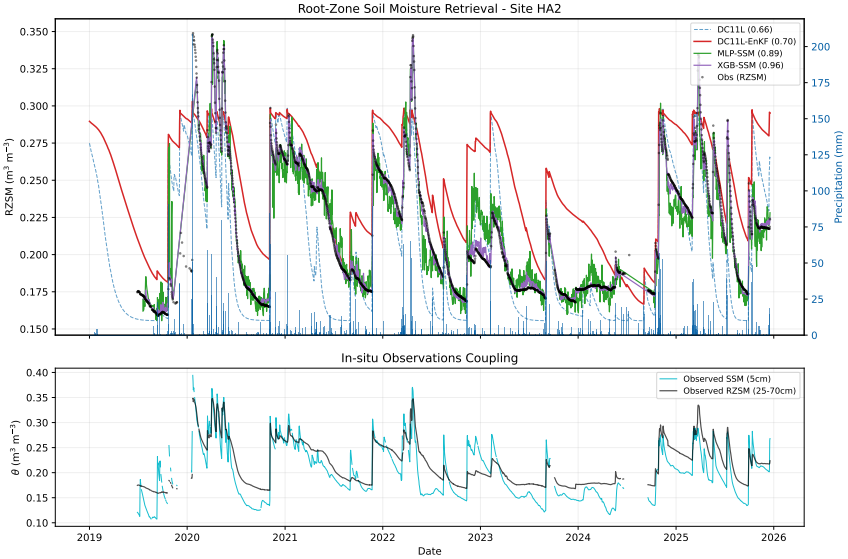
Supplement: RZSM Timeseries (GA1)



Supplement: RZSM Timeseries (HA1)



Supplement: RZSM Timeseries (HA2)



Supplement: RZSM Timeseries (PM2)